Mineração de Dados Textuais para a Classificação da Atividade Econômica Principal de Empresas

Uma proposta de aplicação em pesquisas econômicas

Ana Gabriela Faria da Silva

Dissertação apresentada ao
Instituto de Matemática e Estatística
da Universidade de São Paulo
para obtenção do título de
Mestra em Ciências

Programa: Estatística
Orientadora: Profª. Drª. Florencia Leonardi

São Paulo
13 de Maio de 2022

Mineração de Dados Textuais para a Classificação da Atividade Econômica Principal de Empresas

Uma proposta de aplicação em pesquisas econômicas

Ana Gabriela Faria da Silva

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 13 de Maio de 2022.

Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão julgadora:

Profª. Drª. Florencia Leonardi (orientadora) – IME-USP
Prof. Dr. Rafael Izbicki – UFSCar
Profª. Drª. Denise Britz do Nascimento Silva – ENCE

As opiniões expressas neste trabalho são de exclusiva responsabilidade da autora.

All models are wrong, but some are useful. (George Box)

Agradecimentos

À minha orientadora, Florencia Leonardi, por toda paciência, apoio, compreensão e generosidade durante a execução deste trabalho.

Aos professores Rafael Izbicki e Denise Britz do Nascimento Silva por toda a gentileza, contribuição e atenção despendidas ao participarem da banca de defesa.

Aos colegas e ao corpo técnico da USP com quem, apesar da distância física imposta, pude aprender e contar durante essa fase.

Ao Instituto Brasileiro de Geografia e Estatística (IBGE) pelo tempo concedido para minha participação no programa de mestrado.

À atual gerente da equipe de Métodos da COSEC, Adriana Bandeira, à sua antecessora, Maria Deolinda Cabral, e ao coordenador, Alessandro Maia Pinheiro, por todo o incentivo e auxílio durante essa empreitada.

Aos amigos Leandro Andraos, Danielle Chaves, Raquel Rabello e Amanda Soares que sempre estiveram por perto e disponíveis desde o início da minha jornada no IBGE. Agradeço ainda a ajuda da Aline Visconti, Augusto Fadel, Vinicius Mendonça, Francisco Marta, Gustavo Lima, João Carlos Rodrigues, Sonia Maria Souza e aos outros tantos companheiros do Órgão que prestaram apoio esclarecendo dúvidas, incentivando e, não menos importante, auxiliando nas questões burocráticas durante o período do mestrado.

À equipe de pesquisa Núcleo de Estudos de Foresight, da Fiocruz, e principalmente ao Grupo de Ciência de Dados, da qual tive a oportunidade de fazer parte no último ano, pela paciência, apoio e ensinamentos. Agradeço em especial à Samara Alvarez por toda confiança, incentivo e aprendizado que me proporcionou, inúmeras vezes, nesses anos de amizade.

Aos meus pais, Marly Faria e Sanderson Pereira da Silva, e ao meu irmão, Gabriel Faria da Silva, por sempre estarem ao meu lado e me apoiarem, nunca medindo esforços para me ajudar. Agradeço, também, aos demais membros da minha família por serem

compreensivos com as inúmeras ausências nesse período.

À minha prima, Paula Faria, e à sua família por terem me recebido tão bem, ainda que infelizmente por um curto período, em São Paulo.

Ao meu companheiro, Flávio Peixoto, pelo apoio desde quando o mestrado era um projeto distante. Nossas conversas, sua paciência e generosidade foram, certamente, primordiais para execução desta dissertação. Agradeço por ser meu leitor e ouvinte incansável.

Ao Danilo Peixoto e ao Heitor Peixoto por me receberem tão bem e pela paciência desproporcional as suas idades.

Aos meus sogros, cunhados e outros muitos amigos que dividiram aflições, se preocuparam comigo, torceram e se interessaram pelo andamento do estudo.

Não tenho dúvidas de que muitas pessoas contribuíram para a execução deste trabalho. Ter uma rede de apoio foi fundamental, principalmente, em um período tão nebuloso.

Muito obrigada a todos!

Resumo

Ana Gabriela Faria da Silva. **Mineração de Dados Textuais para a Classificação da Atividade Econômica Principal de Empresas: *Uma proposta de aplicação em pesquisas econômicas***. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022.

O papel das estatísticas é produzir informações que busquem retratar a realidade. Para que isso seja possível, se faz necessário o estabelecimento de padrões. As estatísticas econômicas no Brasil, seguindo diretrizes internacionais, adotam a Classificação Nacional de Atividades Econômicas (CNAE) para caracterizar as atividades desenvolvidas pelas empresas. A CNAE possui uma estrutura hierárquica onde quanto maior o número de dígitos mais específica é a atividade descrita. Este trabalho objetiva avaliar o uso do aprendizado supervisionado, no âmbito da mineração de dados textuais, para a obtenção da CNAE que corresponde à atividade econômica principal das empresas. Para tanto, são utilizados textos como variáveis preditoras, obtidos via *web scraping*, de páginas da *web* e o oriundo da própria *URL* da companhia. Tanto a *URL* quanto a variável resposta, a CNAE, têm como origem as Pesquisas Estruturais por Empresa, do Instituto Brasileiro de Geografia e Estatística (IBGE). Por conta da estrutura hierárquica da classificação são testadas duas abordagens para o ajuste dos modelos. A primeira, denominada classificação plana, tem por objetivo obter diretamente a classe mais específica. Já a segunda, enquadrada na categoria de classificação hierárquica, consiste na construção de diversos classificadores locais independentes para cada nível da hierarquia de classes. Nos dois casos, dentre os algoritmos testados, a Regressão Logística apresentou o melhor desempenho, se mostrando apta para extrair padrões capazes de identificar a classificação. As duas abordagens forneceram resultados diferentes por classe, tendo o classificador plano exibido um comportamento mais adequado em categorias que tendiam a ser mais difíceis de caracterizar nos níveis superiores, ou seja, naqueles que representam atividades menos específicas. Apesar disso, nas duas abordagens o resultado ao se considerar todas as classes foi próximo.

Palavras-chave: classificação de atividades econômicas. web scraping. mineração de dados textuais. aprendizado automático. classificação hierárquica.

Abstract

Ana Gabriela Faria da Silva. **Text Mining for Classifying Main Economic Activity of Companies: *A proposal for application in business surveys*.** Thesis (Master's). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2022.

The role of statistics is to produce information that aims to portray reality. To make this possible, it is necessary to establish standards. Economic statistics in Brazil, following international guidelines, adopts the National Classification of Economic Activities (CNAE). The CNAE has a hierarchical structure where the greater the number of digits more specific the activity described. The purpose of the present study is to evaluate the use of supervised learning, in the context of text mining, to achieve the CNAE which corresponds to the main economic activity of the companies. Therefore, it is used texts as predictors variables, obtained via web scraping, from business websites and URLs. Both URLs and the response variable, the CNAE, derive from the Annual Business Surveys, from the Brazilian Institute of Geography and Statistics (IBGE). Due to the hierarchical structure of the classification, two approaches are tested to fit the models. The first one, called flat classification, aims to directly obtain the most specific class. The second approach, which is framed in the category of hierarchical classification, consists of training several independent local classifiers for each level of the class hierarchy. In both cases, among the tested algorithms, the Logistic Regression classifier presented the best performance, being able to extract patterns fit to identify the classification. The two approaches provided different results by class, having the flat classifier exhibited a more adequate behavior in categories that tended to be more difficult to characterize in the higher levels, that is, in those that represent less specific activities. Despite this, the result was similar in both approaches when considering all classes.

Keywords: classification of economic activities. web scraping. text mining. machine learning. hierarchical classification.

Lista de Abreviaturas

BoW	*Bag of Words*
CAGED	Cadastro Geral de Empregados e Desempregados
CATI	*Computer Assisted Telephone Interview*
CBS	Cadastro Básico de Seleção
CEMPRE	Cadastro Central de Empresas
CGI.br	Comitê Gestor da *Internet* no Brasil
CIIU	*Clasificación Industrial Internacional Uniforme*
CNAE	Classificação Nacional de Atividades Econômicas
CNPJ	Cadastro Nacional da Pessoa Jurídica
CONCLA	Comissão Nacional de Classificação
CPF	Cadastro de Pessoas Físicas
DAG	*Direct Acyclic Graph*
DV	Dígito Verificador
ESS	*European Statistical System*
HTML	*HyperText Markup Language*
IA	Inteligência Artificial
ISIC	*International Standard Industrial Classification of All Economic Activities*
IBGE	Instituto Brasileiro de Geografia e Estatística
OCDE	Organização para Cooperação e Desenvolvimento Econômico
KDD	*Knowledge Discovery in Databases*
LGPD	Lei Geral de Proteção de Dados Pessoais
MSV	Métodos de Suporte Vetorial
NB	Naive Bayes
NIC.br	Núcleo de Informação e Coordenação do Ponto BR
PAC	Pesquisa Anual de Comércio
PAIC	Pesquisa Anual da Indústria da Construção

PAS	Pesquisa Anual de Serviços
PIA - Empresa	Pesquisa Industrial Anual - Empresa
PIA - Produto	Pesquisa Industrial Anual - Produto
PIB	Produto Interno Bruto
PPEmp	Pesquisa Pulso Empresa
RAIS	Relação Anual de Informações Sociais
RL	Regressão Logística
SIMCAD	Sistema de Manutenção Cadastral
SMOTE	*Synthetic Minority Over-sampling Technique*
TF	*Term Frequency*
TF-IDF	*Term Frequency - Inverse Document Frequency*
TI	Tecnologia da Informação
UF	Unidade da Federação
UL	Unidade Local
UNECE	*United Nations Economic Commission for Europe*
UNSD	*United Nations Statistical Division*
URL	*Uniform Resource Locator*

Lista de Figuras

1.1 Esquema ilustrativo de ações executadas e bases criadas no decorrer do desenvolvimento do estudo de caso proposto 6

3.1 Empresas (%) que mudaram de CNAE, segundo os níveis da classificação econômica, em um triênio 29
3.2 Empresas (%), sem verificação da CNAE por técnicos, com divergência entre a CNAE do cadastro de seleção e a informada na pesquisa, por pesquisa, segundo os níveis da classificação econômica - 2014-2019 30
3.3 Empresas (%) que lançaram ou passaram a comercializar novos produtos ou serviços, por pesquisa, em cada quinzena 34
3.4 Empresas (%) após a extração e o pré-processamento dos dados, por pesquisa, segundo as Grandes Regiões 41

4.1 Exemplo de parte da estrutura de árvore, oriunda da organização da CNAE, englobando a Seção G 51
4.2 Ilustração dos classificadores locais criados 53

5.1 Melhores resultados obtidos, através dos classificadores locais, considerando a métrica *Macro* F_1 *score*, a partir da base de treino, por nó, segundo os algoritmos testados 69
5.2 Diagrama de Venn representando as empresas classificadas em classes finais incorretas de acordo com a abordagem plana e local 93

Lista de Quadros

2.1 Organização hierárquica da CNAE 2.0/CNAE-Subclasses 2.3 18
2.2 Categorias da CNAE 2.0 - PIA-Empresa 22
2.3 Categorias da CNAE 2.0 - PAIC 23
2.4 Categorias da CNAE 2.0 - PAC 24
2.5 Categorias da CNAE 2.0 - PAS 24
5.1 Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por pesquisa - Nó: Pesquisa 70
5.2 Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PAC - Nó: PAC-CNAE2D 71
5.3 Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 3 dígitos da Divisão 45 da PAC - Nó: PAC45-CNAE3D 72
5.4 Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 3 dígitos da Divisão 46 da PAC - Nó: PAC46-CNAE3D 73
5.5 Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 3 dígitos da Divisão 47 da PAC - Nó: PAC47-CNAE3D 74
5.6 Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PIA-Empresa - Nó: PIA-CNAE2D 75
5.7 Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PAS - Nó: PAS-CNAE2D 78
5.8 Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por classe final - Classificação Plana 81
A.1 Hiperparâmetros e representações testadas ao se considerar cada classificador 103

A.1 Estrutura detalhada de parte da CNAE 2.0: Códigos e denominações . . . 117

Lista de Tabelas

3.1 Empresas no CBS, selecionadas pelas pesquisas e utilizadas no estudo, por pesquisa . . . 36
3.2 Empresas no âmbito do estudo, após a extração e o pré-processamento dos dados, por pesquisa, segundo o estrato de seleção . . . 41
3.3 Empresas após a extração e o pré-processamento dos dados na PIA-Empresa, PAC e PAS, segundo a Seção e a Divisão da CNAE, e na PAC, segundo o Grupo da CNAE . . . 42

4.1 Os cinco *tokens* com maior cobertura e os cinco com maior frequência absoluta presentes na base de treino . . . 46
4.2 Os cinco *tokens* com maior cobertura e os cinco com maior frequência absoluta presentes na base de treino após o uso do *stemmer* . . . 48

5.1 Melhores resultados obtidos, através dos classificadores planos, considerando a métrica *Macro* F_1 *score*, a partir da base de treino, segundo os algoritmos testados . . . 80
5.2 Resultados obtidos na base de teste com os melhores algoritmos e configurações encontradas na base de treino, por tipo de métrica, segundo as abordagens testadas . . . 88
5.3 Resultados obtidos na base de teste considerando as métricas planas, por tipo de classificador, segundo as classes finais . . . 88
5.4 Empresas, de classes que evidenciaram a dificuldade encontrada pelos classificadores locais, classificadas incorretamente . . . 91
5.5 Classes que tiveram mais de 40% das empresas preditas inadequadamente em uma classe específica, considerando apenas as classes que apresentaram 5 ou mais companhias classificadas erroneamente . . . 92

5.6 Comparação das CNAEs do histórico e da classificação fornecida pelo informante em 2020 com o resultado obtido através de cada um dos classificadores - apenas para as empresas que foram classificadas de forma inadequada nas abordagens plana e local e tiveram uma CNAE informada em 2020 diferente da considerada em 2019 95

B.1 Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: Pesquisa . 105
B.2 Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC-CNAE2D 106
B.3 Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC45-CNAE3D 107
B.4 Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC46-CNAE3D 109
B.5 Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC47-CNAE3D 110
B.6 Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PIA-CNAE2D 111
B.7 Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAS-CNAE2D 113
B.8 Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem plana 114

Sumário

1 Introdução **1**
1.1 Motivação . . . 3
1.2 Objetivos . . . 4
1.3 Estrutura do Trabalho . . . 4

2 Contexto **7**
2.1 Ascensão da Ciência de Dados . . . 7
2.2 Uso de Megadados e de Ferramentas da Ciência de Dados em um Órgão Oficial de Estatística . . . 11
2.3 Instituto Brasileiro de Geografia e Estatística – IBGE . . . 14
2.4 O Sistema Estatístico Nacional e as Convenções . . . 15
2.4.1 Classificação Nacional de Atividades Econômicas - CNAE . . . 17
2.5 Pesquisas Estruturais por Empresa e seus Cadastros . . . 19
2.5.1 Cadastro Central de Empresas - CEMPRE . . . 20
2.5.2 Cadastro Básico de Seleção - CBS . . . 21
2.5.3 Pesquisa Industrial Anual - PIA–Empresa . . . 22
2.5.4 Pesquisa Anual da Indústria da Construção - PAIC . . . 22
2.5.5 Pesquisa Anual de Comércio - PAC . . . 23
2.5.6 Pesquisa Anual de Serviços - PAS . . . 24

3 Avaliação Preliminar e Construção das Bases de Dados **27**
3.1 Dados das Pesquisas Estruturais por Empresa . . . 28
3.1.1 Avaliação Preliminar . . . 28
3.1.2 Estudo Principal . . . 35
3.2 Dados da *Web* . . . 37
3.2.1 Extração dos Dados da *Web* . . . 38
3.2.2 Pré-Processamento dos Dados da *Web* . . . 39
3.3 Base Final . . . 40

3.4 Divisão da Base Final - Treino e Teste 44

4 Variáveis Preditoras Finais e Modelagem 45
4.1 Descrição dos *Tokens* e Redução de Dimensionalidade da Base de Treino 46
4.1.1 Redução de Dimensionalidade a partir da Cobertura dos *Tokens* . 47
4.1.2 Redução de Dimensionalidade através de Técnicas de *Stemming* 47
4.2 Representações do tipo *Bag-of-Words* (Saco de Palavras) 48
4.3 Dados Desbalanceados 49
4.4 Classificação Hierárquica 50
4.5 Algoritmos de Aprendizado Supervisionado 54
4.5.1 Naive Bayes - NB 54
4.5.2 Regressão Logística - RL 56
4.5.3 *Support Vector Machines* (Métodos de Suporte Vetorial - MSV) . . 58
4.5.4 Boosting 60
4.6 Validação Cruzada 62
4.7 *Random Search* (Busca Aleatória) 62
4.8 Métricas para Avaliação de Desempenho 63

5 Apresentação do Modelo Final e Análise dos Resultados 67
5.1 Resultados Obtidos na Base de Treino 68
5.1.1 Resultados Obtidos com os Classificadores Locais 68
5.1.2 Resultados Obtidos com os Classificadores Planos 80
5.2 Resultados Obtidos na Base de Teste 87
5.2.1 Investigação do Histórico das Empresas Classificadas Incorretamente 94
5.2.2 Empresas Classificadas Incorretamente e o seu Comportamento em 2020 94

6 Conclusões 97
6.1 Limitações e Trabalhos Futuros 98

Apêndices

A Funções Utilizadas para o Ajuste dos Algoritmos e Valores de Hiperparâmetros Aplicados na Busca Aleatória 103

B Melhores Resultados Obtidos com cada Algoritmo 105

Anexos

A Códigos e Denominações de parte da Classificação Nacional de Atividades Econômicas - CNAE **117**

Referências **123**

Capítulo 1

Introdução

O atual patamar do desenvolvimento tecnológico permite a criação de vultosos bancos de dados, muitos deles a partir da *internet*. Esses, por sua vez, impulsionam o avanço de uma estrutura tecnológica que permite a sua análise e/ou uso pelos mais diversos grupos de pessoas. Nesse cenário, surge o atual conceito de ciência de dados. Apesar de sua relativa popularização, sua difusão ainda se encontra longe do que se poderia tomar por democratizada. Entretanto, é essa pequena pluralidade que molda essa ciência e a insere em discussões sobre o mau uso dos dados, que vão desde a falta de conhecimentos estatísticos apropriados para a construção das bases e para o uso das ferramentas até a percepção de mecanismos de perpetuação de preconceitos e da necessidade da transparência e reprodutibilidade. Essas não eram discussões em voga no passado, quando os instrumentos de coleta de dados eram mais restritos, bem como as capacidades de análise e utilização. Ciente dos desafios, é preciso também valorizar o extenso número de possibilidades que surgem ao se conseguir transformar esses dados em informações úteis.

A legislação relacionada à *internet* é recente no Brasil. A Lei que estabelece princípios, garantias, direitos e deveres para o seu uso no País foi assinada em 2014 e ficou conhecida como Marco Civil da *Internet*. Poucos anos antes outras leis associadas a delitos informáticos e à contratação no comércio eletrônico também foram sancionadas. Já a Lei Geral de Proteção de Dados Pessoais (LGPD) que, de acordo com a mesma, "[...] dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural."(Brasil, 2018), é ainda mais recente, tendo sido aprovada em agosto de 2018 com vigência a partir de agosto de 2020. Leis similares a essa são recentes também em outros países. Esse é o caso do *General Data Protection Regulation* (Regulamento Geral sobre a Proteção de Dados), aplicável desde 25 de maio de 2018, que objetivou harmonizar as diretrizes na União Europeia.

O estudo aqui apresentado focou no uso de megadados públicos em conjunto com ferramentas da ciência de dados, como o aprendizado automático supervisionado, em um órgão oficial de estatística. No trabalho proposto visou-se testar se as informações oriundas da *web*, coletadas via *web scraping*, serviriam para complementar, atualizar e validar os dados relativos à Classificação Nacional de Atividades Econômicas (CNAE) de empresas já

pertencentes aos bancos de dados do Instituto Brasileiro de Geografia e Estatística (IBGE). Esse tipo de coleta também poderia abrir portas para obtenção de outras informações passíveis de dar origem a novas pesquisas. Muitos institutos pelo mundo estão realizando a aquisição de dados, de forma rotineira e intensificada nos anos pandêmicos, dessa maneira. Alguns deles já possuem uma política relacionada ao *web scraping*. Esse é o caso do Reino Unido (ONS, 2020) e do Canadá (StatCan, 2021), que inclusive a atualizaram durante a pandemia. Já a União Europeia (Eurostat, 2022a) possui apenas um rascunho, ainda não se tratando de uma política acordada. Essa é uma etapa ainda a ser pensada no Brasil.

O mesmo desenvolvimento tecnológico que possibilita este estudo faz com que seja cada vez mais difícil classificar a atividade econômica principal de uma companhia. Segundo Tigre (2019b), as novas tecnologias associadas às infraestruturas e modelos organizacionais facilitaram, por exemplo, a transformação de bens materiais em serviços e a agregação de serviços aos produtos, sendo esse tipo de oportunidade explorada, principalmente, por novas empresas, mais enxutas, flexíveis e com maior foco na demanda potencial dos clientes.

Apesar da dinamicidade relacionada às atividades econômicas, a CNAE precisa ser estática, uma codificação previamente acordada, para servir de pilar para as estatísticas e fazer com que estas sejam informativas. O uso do aprendizado supervisionado, nesse cenário, tem o intuito de colaborar para que o processo de identificação da classificação econômica de uma empresa seja mais ágil e menos custoso do que o executado manualmente. Embora os modelos tenham natureza estacionária, ou seja, sejam qualificados para capturar padrões e falhem quando a forma como um conceito é expresso pela população muda ou quando o próprio conceito se transforma, eles podem ser capazes de cumprir o objetivo.

Zeelenberg (2016), do Instituto de Estatística da Holanda, destaca a necessidade de se aprender a criar e trabalhar com classificações flexíveis. De acordo com ele, a classificação padrão não deve mais ocupar uma posição central, uma vez que para muitos fenômenos complexos, outras classificações são muito mais importantes, como, por exemplo, pequenas/grandes empresas, empresas exportadoras/domésticas e empresas que usam/produzem Tecnologia da Informação (TI). Ele ainda acrescenta que a classificação padrão atual também está conceitualmente ultrapassada, pois as empresas são classificadas de acordo com a sua atividade principal e, hoje em dia, as companhias mudam com muito mais frequência suas atividades e, muitas vezes, suas atividades subsidiárias podem ser quase tão importantes quanto a principal. Além disso, o autor pontua que, do ponto de vista da sociedade, é mais importante saber quais produtos e serviços estão sendo fornecidos e demandados do que onde eles estão sendo produzidos. Ele destaca ainda que esse movimento levará a mais uso de métodos estatísticos, como análise de agrupamentos e outros métodos de classificação, bem como de análise. Apesar de já estar clara a necessidade de novas formas, nenhum método novo despontou, sendo a CNAE ainda o principal meio utilizado para o fornecimento de estatísticas econômicas.

É nesse contexto, portanto, que este trabalho buscou integrar ferramentas existentes na ciência de dados e a CNAE através da proposta de um estudo de caso. Em geral, uma atividade envolvendo ciência de dados é um processo iterativo: os resultados são analisados; modificam-se os dados; a forma como é realizada a limpeza e o pré-processamento; e,

novamente, os resultados são examinados. As iterações seguem até que seja obtido um resultado satisfatório. No estudo aqui executado, uma gama de variações foi testada e outras tantas foram citadas quando se tratou dos trabalhos futuros. Assim, vale reconhecer que esta é uma primeira experiência e que muito ainda pode ser testado e melhorado.

1.1 Motivação

O IBGE é o principal provedor de informações sociais, demográficas e econômicas do País. Essas informações contribuem tanto para a formulação e avaliação de políticas públicas como para viabilizar a existência de uma democracia, permitindo que os indivíduos possam se tornar conscientes da realidade que estão inseridos. A aceitação das estatísticas oficiais depende de sua pertinência, da credibilidade e da boa reputação dos órgãos responsáveis, o que se torna uma missão muito mais delicada em um cenário de propagação da ideia de Estado corrupto e ineficiente, no qual é possível observar o ataque dos próprios governantes às instituições públicas.

Nesse contexto, possibilitado pelo surgimento de novas tecnologias, meios e métodos baseados muitas vezes na estatística, nem sempre novos, mas com novas perspectivas de usos e aplicações, estão emergindo. Os mecanismos da ciência de dados vêm ganhando notoriedade em um cenário onde a abundância de dados prontamente acessíveis faz com que informações defasadas sejam cada vez menos utilizadas, apesar de já ser sabido que a qualidade da informação é (ou deveria ser) um fator de alto peso na balança. Nessa conjuntura, cabe a exploração do tema ao se olhar para os órgãos oficiais de estatística.

Os motivos para a escolha do assunto deste trabalho estão relacionados ao entendimento da autora da importância do papel do IBGE perante a sociedade, onde se pode mencionar tanto a necessidade de informações mais ágeis[1], quanto a importância de desonerar o respondente[2], além de amenizar o efeito da redução de trabalhadores do quadro do Órgão. Soma-se também a esse entendimento a vontade de compreender a respeito de questões relacionadas ao uso de novas bases de dados e ao reconhecimento de padrões, que nem sempre são isentos de preconceitos, e sobre a importância de humanos se responsabilizarem pelas decisões e conclusões geradas pelos algoritmos. Tende-se a presumir que qualquer sistema construído com base em matemática deve, de alguma forma, ser objetivo, isento dos vieses que permeiam a tomada de decisão humana, o que nem sempre se verifica na realidade. Somado a isso, muitas vezes, quando se entende o contrário, a responsabilidade recai sobre o próprio modelo, numa tentativa de desonerar quem o construiu.

Para a autora, o encanto da estatística, aflorado com o advento da ciência de dados, tem relação com o equilíbrio que ela requer, onde não se deve confiar irrefletidamente em resultados gerados através do uso da matemática, como já muito difundido no caso das correlações espúrias; como não se deve "mentir" para que os resultados reflitam algo que o pesquisador gostaria de concluir; como nos casos também já corriqueiros onde se informa a média, mas não se diz qual foi utilizada; ou como quando se apresentam gráficos fora da

[1] Como ficou ainda mais evidente nos anos pandêmicos, onde informações prontamente acessíveis são capazes de salvar vidas.

[2] Visando diminuir a quantidade de não-respostas e melhorar a qualidade da informação obtida.

escala. Os exemplos citados já são velhos conhecidos dos estatísticos e, nos últimos anos, o uso massivo de dados vem dando forma a vários outros.

1.2 Objetivos

O objetivo geral deste trabalho é investigar o uso do ferramental oriundo da ciência de dados, através da mineração de dados textuais, em conjunto com as novas bases disponíveis, para obtenção da classificação econômica de uma empresa. O trabalho tem o intuito de auxiliar e agilizar o processo de produção de estatísticas oficiais no Brasil e também pode atuar como um mecanismo de fornecimento de novas ideias.

Assim, definiu-se como objetivo específico, e principal, avaliar a utilização do aprendizado automático supervisionado para determinar a Classificação Nacional de Atividades Econômicas (CNAE) de empresas. Para essa tarefa, as variáveis preditoras foram os textos extraídos das páginas da *web*, em 2021, da respectiva organização e os textos das próprias *URLs*. Esse último e a variável resposta, no caso a CNAE, foram obtidos através das informações prestadas pelas companhias às Pesquisas Estruturais por Empresa do IBGE referentes ao ano de 2019.

O trabalho tem ainda como objetivos secundários tentar avaliar a estabilidade temporal da classificação econômica principal de uma empresa e, com o intuito de buscar possíveis explicações, os erros de classificação gerados pelo método desenvolvido.

1.3 Estrutura do Trabalho

Este trabalho está organizado em 6 capítulos, 2 apêndices e 1 anexo. A organização dos capítulos procura apresentar uma estrutura lógica partindo do contexto no qual as ferramentas utilizadas e o problema tratado estão inseridos, passando por uma investigação preliminar que visa pautar a factibilidade do estudo de caso proposto, seguindo para a sua apresentação, onde é descrita desde a coleta dos dados até a avaliação dos resultados encontrados.

O Capítulo 1, no qual o leitor se encontra, apresenta uma ideia geral do trabalho, explicitando a sua motivação, os objetivos e a estrutura a ser encontrada no decorrer da leitura.

O Capítulo 2 tem no seu início a exposição do contexto no qual a ciência de dados se fundamenta e desenvolve. Ele fornece um panorama histórico dessa e definições dos principais termos empregados por ela. Na sequência, discorre sobre o uso de megadados e ferramentas da ciência de dados em um órgão oficial de estatística. Assim, chega ao IBGE, culminando nos trabalhos que o Instituto desenvolve e focando nas Pesquisas Estruturais por Empresa e na Classificação Nacional de Atividades Econômicas, que podem ser beneficiados pela utilização das fontes alternativas de dados, das novas tecnologias e técnicas estatísticas.

Ao Capítulo 3 cabe averiguar a viabilidade do estudo de caso pretendido e definir os dados a serem empregados na sua execução. Ele detalha as variáveis das Pesquisas

Estruturais por Empresas, apresentadas na Capítulo 2, a serem utilizadas e também define o material a ser coletado da *web* e das *URLs*. A necessidade da avaliação preliminar está relacionada à defasagem temporal entre as diferentes fontes de dados. Após essa análise, os moldes dos dados usados no estudo principal são delineados. O Capítulo se encerra com a construção de duas bases de dados, uma a ser operada na etapa de treinamento e outra na de teste.

Já no Capítulo 4 são apresentados todos os passos realizados para execução da modelagem. A ele compete descrever os procedimentos executados para tratar a alta dimensionalidade dos dados, o seu desbalanceamento e o fato da CNAE apresentar uma estrutura hierárquica. Além disso, ele exibe os algoritmos de aprendizado supervisionado utilizados e pormenoriza como o ajuste é executado e como os resultados são avaliados.

Na sequência, a função do Capítulo 5 é expor os resultados obtidos após a realização das etapas detalhadas no Capítulo 4, sendo utilizadas as bases descritas no Capítulo 3 e se valendo das definições e do contexto delineado no Capítulo 2. Acrescenta-se, ainda, no final desse Capítulo, um exame das empresas para as quais os resultados foram diferentes do esperado.

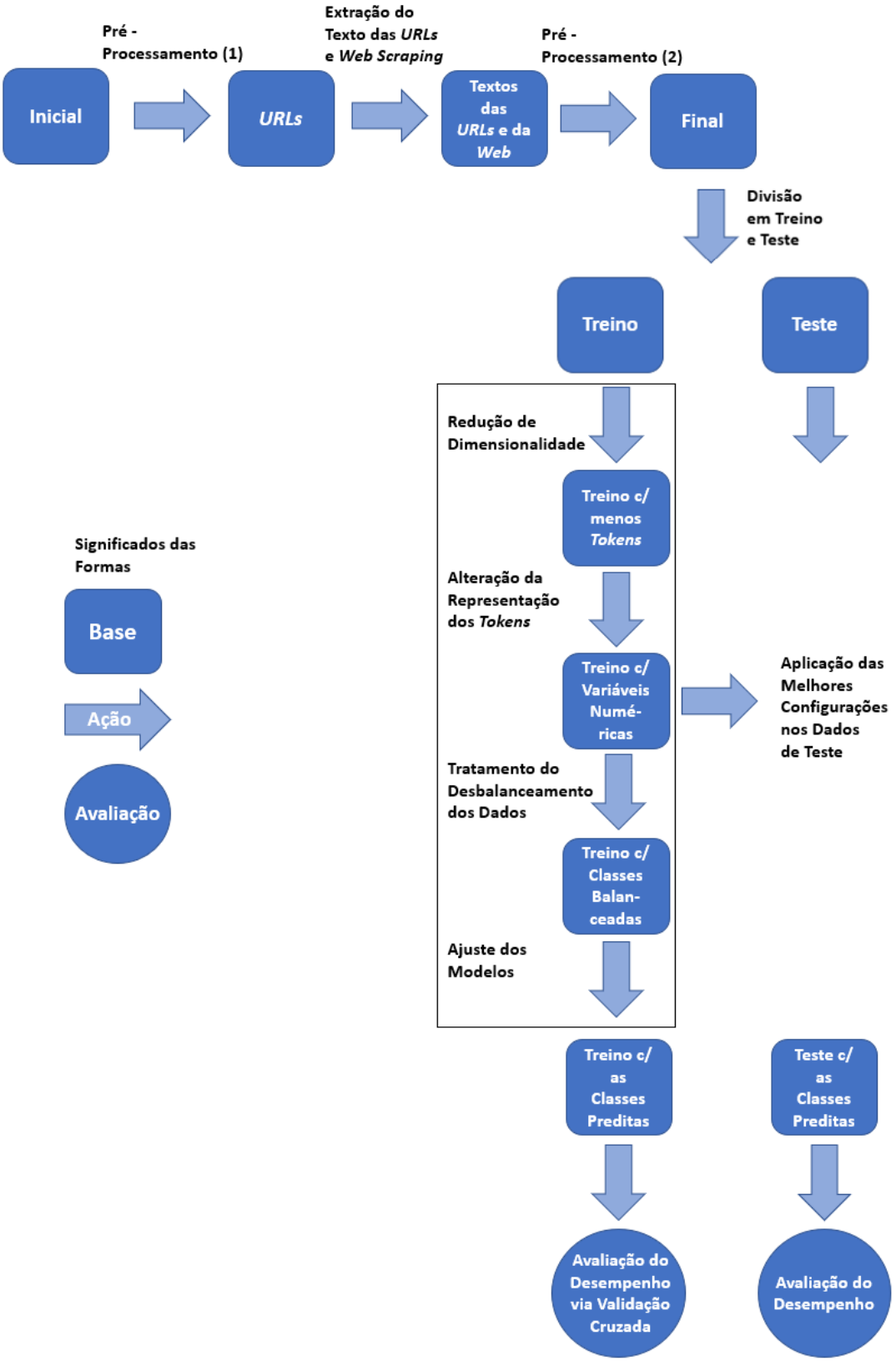
Por último, no Capítulo 6, são feitas as considerações finais acerca do tema e, mais precisamente, da análise proposta. Além disso, não menos importante, são pontuadas as limitações do trabalho realizado e são propostos trabalhos futuros.

Vale ainda mencionar que compete ao Apêndice A exibir as funções utilizadas em *Python* para o ajuste dos modelos e os valores de hiperparâmetros aplicados na busca aleatória e ao Apêndice B listar os melhores resultados obtidos na base de treino com cada algoritmo e suas configurações ao se considerar a abordagem hierárquica e a plana. E, para finalizar, cabe ao Anexo A apresentar os códigos e denominações da CNAE utilizados no estudo.

A Figura 1.1 traz ainda uma representação esquemática de alguns dos principais pontos desenvolvidos no decorrer do trabalho, a partir do Capítulo 3 até o Capítulo 5. Nela, as setas representam ações, os quadrados indicam as diferentes bases de dados criadas no decorrer do processo e os círculos apontam as etapas de avaliação dos resultados.

Seguindo o esquema ilustrativo, pode-se mencionar que todo o processo, desde a criação da base até a sua divisão em treino e teste, consta do Capítulo 3. O *"Pré-Processamento (1)"*, a *"Extração do Texto das URLs e Web Scraping"* e o *"Pré-Processamento (2)"* são descritos na Subseção 3.1.2, na Seção 3.2 e na Seção 3.3 e, para finalizar, a *"Divisão em Treino e Teste"* é retratada na Seção 3.4. Já os passos que seguem, podem ser encontrados no Capítulo 4. A *"Redução de Dimensionalidade"* está na Seção 4.1, a *"Alteração da Representação dos Tokens"* na Seção 4.2, o *"Tratamento do Desbalanceamento dos Dados"* na Seção 4.3 e o *"Ajuste dos Modelos"* engloba as Seções 4.4, 4.5, 4.6 e 4.7. No Capítulo 4, na Seção 4.8, encontra-se ainda a descrição teórica das métricas utilizadas para a avaliação do desempenho dos modelos. Assim, fica reservado para o Capítulo 5 a exposição do melhor modelo ajustado, dos números obtidos quando da avaliação de desempenho e da análise das empresas classificadas incorretamente.

Figura 1.1: Esquema ilustrativo de ações executadas e bases criadas no decorrer do desenvolvimento do estudo de caso proposto



Fonte: Autoria própria.

Capítulo 2

Contexto

Vive-se hoje em um ambiente de larga produção de dados e de rápida propagação de novas tecnologias, fatores esses que permitem cada vez mais a exploração de dados inéditos, das mais variadas formas e em grandes volumes. Apesar dessa nova gama de possibilidades, já está posto que elas vêm cercadas de muitos desafios, como o que fazer para conter robôs propagando notícias falsas, como evitar avaliações injustas passíveis de serem realizadas com base em algoritmos, como identificar a disseminação de vieses através do uso de bases cuja representação é de difícil avaliação, entre outros. Consciente das adversidades e questionando constantemente se existem outras, diferentes das já mapeadas, é possível obter informações capazes de agregar novos conhecimentos ou aprimorar os já existentes.

Inserido nesse contexto, encontram-se os órgãos oficiais de estatística. Muitos desses institutos já vêm utilizando dados de diversas fontes, bem como se beneficiando das tecnologias atuais para desenvolver novas pesquisas, detectar informações problemáticas, realizar classificações, entre outras tarefas, mas muito ainda resta a ser explorado. Nesse cenário, no Brasil, se encontra o Instituto Brasileiro de Geografia e Estatística e os seus diversos trabalhos.

Este capítulo busca apresentar um panorama do advento, do avanço e da difusão da ciência de dados e adiciona nesse ambiente o IBGE, de modo a tentar fornecer argumentos a respeito de como, do porquê e de onde essa ciência se encaixa na realidade do Instituto.

2.1 Ascensão da Ciência de Dados

Nos últimos anos a ciência de dados vem ganhando cada vez mais notoriedade em todo o mundo e em várias esferas, tais quais acadêmica, empresarial e governamental. Universidades estão disponibilizando cursos de graduação e programas de pós-graduação voltados para o assunto, assim como já existem equipes de trabalho, tanto no setor privado como no público, com essa denominação e são conhecidos os casos de uso de algoritmos no dia a dia, como nos sistemas de recomendação, que já se tornaram velhos conhecidos dos usuários da *internet*. Esse movimento de ascensão está ligado à quantidade cada vez

maior de dados disponíveis, à acessibilidade a computadores mais potentes e à necessidade de transformar esses dados em informações úteis a respeito do mundo. Em consonância com essa mutabilidade, modificam-se também os significados atribuídos ao termo "ciência de dados".

Essa expressão não é recente. Press (2013), em seu artigo *A Very Short History Of Data Science*, ao traçar uma linha do tempo da evolução do termo e de outros a ele relacionados, menciona que Naur (1974) definiu ciência de dados como "A ciência de lidar com dados, uma vez que eles estejam estabelecidos, enquanto a relação dos dados com o que eles representam é delegada a outros campos e ciências."(tradução da autora) [1].

Já segundo o Instituto de Tecnologia da Georgia (2021), o termo foi cunhado em uma aula inaugural ministrada, em 1997, por C. F. Jeff Wu, chefe de Estatística aplicada à Engenharia (*Engineering Statistics*) na Coca-Cola e professor do Instituto, que sugeriu que a estatística fosse renomeada para ciência de dados e os estatísticos fossem chamados de cientistas de dados. Outras fontes dizem ainda que o termo já vinha sendo usado desde a década de 1980 por C. F. Jeff Wu em suas palestras.

Blei e Smyth (2017) definem a ciência de dados como sendo filha da estatística e da ciência da computação e fornecem uma visão holística a respeito dela. Segundo os pesquisadores, ela é mais do que uma combinação das disciplinas citadas anteriormente. Para eles "a ciência de dados holística requer que nós entendamos o contexto dos dados, que sejam avaliadas as responsabilidades envolvidas no uso de dados privados e públicos e que seja claramente comunicado o que um conjunto de dados pode ou não dizer sobre o mundo." (Blei e Smyth, 2017, p.8691, tradução da autora)[2].

Em um contexto em que a disponibilidade de dados é cada vez maior, surge a ideia de *big data* (megadados)[3]. O termo está relacionado com o fenômeno de rápido crescimento da capacidade de sistemas de computação para reunir e transportar grandes quantidades de dados. A maioria dos autores concorda que as características essenciais dos megadados podem ser descritas por três V's, nomeadamente Volume (quantidade de dados), Variedade (formas diferentes de dados – imagens, texto, entre outros) e Velocidade (fluxo contínuo de dados). Alguns autores usam ainda um quarto V para defini-los, a Veracidade (incerteza sobre os dados). Já outros adicionam ainda mais V's que variam entre Valor (utilidade dos dados), Vulnerabilidade (proteção dos dados dos indivíduos), Variabilidade (fluxos de dados imprevisíveis), entre outros. Não existe hoje um consenso sobre o assunto.

Nesse cenário, métodos estatísticos e computacionais deixam de ser suficientes para explorar a abundância de dados para previsão, investigação, compreensão e intervenção, assim nasce a ciência de dados tal qual é conhecida atualmente. Alguns dos novos problemas que surgem são computacionais, como conjuntos enormes de dados e metadados complexos, outros são estatísticos, como as vastas interações entre variáveis correlacionadas e, não

[1] *"The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."*

[2] *"Holistic data science requires that we understand the context of data, appreciate the responsibilities involved in using private and public data, and clearly communicate what a dataset can and cannot tell us about the world."*

[3] Para traduções de termos estatísticos optou-se pelas oriundas do Glossário Inglês - Português produzido pela Associação Brasileira de Estatística e pela Sociedade Portuguesa de Estatística.

menos importante, algumas questões são mais indistintas e filosóficas, como manter um equilíbrio entre suposições adequadas e métodos computacionalmente eficientes para lidar com a multidisciplinaridade em torno da exploração e compreensão dos dados (Blei e Smyth, 2017).

Oriunda da necessidade de transformar grandes quantidades de dados em informações úteis a serem utilizadas para previsões e decisões, enfatiza-se também a relevância de estudos de técnicas de *machine learning* (aprendizado automático)[4]. Jordan e Mitchell (2015, p.255, tradução da autora) definem aprendizado automático como:

> [...] uma disciplina focada em duas questões inter-relacionadas: Como construir sistemas de computador que melhorem automaticamente através da experiência? E Quais são as leis estatísticas-computacionais-informacionais-teóricas fundamentais que governam todos os sistemas de aprendizagem, incluindo computadores, humanos e organizações?[5].

De acordo com os autores, o aprendizado automático despontou como método preferido para desenvolvimento de *softwares* relacionados a reconhecimento de fala, processamento de linguagem natural, controle de robôs e outras aplicações. Reconheceu-se, no caso do aprendizado supervisionado por exemplo, que pode ser mais fácil treinar um sistema lhe mostrando exemplos do comportamento desejado (dados de entrada e saída) do que programá-lo manualmente antecipando a resposta esperada para todas as possíveis entradas.

Os autores pontuam que um problema de aprendizagem pode ser definido como um problema de como melhorar uma medida de desempenho, ao executar uma determinada tarefa, por meio de algum tipo de experiência de treinamento. Assim, algoritmos de aprendizado automático, conceitualmente, podem ver vistos como a busca por um programa computacional em um amplo conjunto de candidatos, guiados pela experiência de treinamento, que otimize a métrica de desempenho considerada.

Jordan e Mitchell (2015) destacam que as tentativas de caracterizar algoritmos de aprendizado automático levaram a misturas de teoria estatística e computacional, pois o objetivo é qualificar, simultaneamente, a complexidade da amostra e a computacional (quanta computação é necessária) e especificar como elas dependem das características do algoritmo de aprendizagem, como a representação que ele usa para o que aprende. Os autores enfatizam ainda que o aprendizado automático é um campo jovem e que tem muito a ser explorado, podendo algumas dessas oportunidades serem vistas contrastando as abordagens atuais de aprendizado automático com os tipos de aprendizado observados em sistemas que ocorrem naturalmente, como com os humanos e outros animais, organizações, economias e na evolução biológica.

A maioria dos problemas de aprendizado automático pode ser enquadrado como supervisionado ou não. No aprendizado automático supervisionado está disponível uma

---

[4] Neste trabalho *machine learning* (aprendizado automático) e *statistical learning* (aprendizado com estatística) são considerados sinônimos.

[5] *"[...] a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? and What are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans, and organizations?"*

variável resposta e a partir de padrões verificados nas variáveis preditoras deve-se prevê-la. Já no não supervisionado a variável resposta não está presente na base utilizada para a modelagem e o foco é descrever associações e padrões entre as variáveis preditoras de forma a fornecer, por exemplo, agrupamentos dos dados.

Aprendizado automático e ciência de dados são campos correlatos. O primeiro se concentra no desenvolvimento e na avaliação de algoritmos para extrair padrões dos dados enquanto o segundo possui um escopo mais amplo e, além de englobar o primeiro, envolve também outros desafios como a coleta, a limpeza e a estruturação dos dados (Kelleher e Tierney, 2018).

A Organização para Cooperação e Desenvolvimento Econômico - OCDE (2019), ao que tudo indica, difere de alguns autores que caracterizam aprendizado automático como sinônimo de inteligência artificial (IA). A OCDE recomenda que sistema de inteligência artificial seja entendido como "[...] um sistema baseado em máquina que pode, para um determinado conjunto de objetivos definidos pelo homem, fazer previsões, recomendações ou tomar decisões que influenciam ambientes reais ou virtuais"(tradução da autora)[6]. A Organização ressalta ainda que a extração das percepções a respeito de ambientes reais e/ou virtuais, a partir dos dados fornecidos a esse sistema de IA, pode ocorrer de maneira automatizada (por exemplo, com aprendizado automático) ou manualmente.

A visão da OCDE, ao que parece, vai de encontro a fornecida por Chollet e Allaire (2017), onde IA também engloba abordagens que não envolvem aprendizado. Um exemplo de uma delas, fornecido pelos autores, está nos primeiros programas de xadrez, que se baseavam apenas em regras codificadas elaboradas por programadores.

Data mining (mineração de dados) é outro termo intimamente relacionado à ciência de dados. Ele se refere ao processo de descoberta de padrões interessantes e extração de conhecimento de grandes quantidades de dados (desde a limpeza dos dados até a apresentação do conhecimento gerado por eles), sendo essa, segundo Han *et al.* (2011), uma visão ampla da funcionalidade. Mineração de dados acabou substituindo, em um contexto mais comercial, a expressão *knowledge discovery in databases – KDD* (descoberta de conhecimentos através de bases de dados), mais prevalente no meio acadêmico. Mais uma vez, não existe concordância a respeito do tema. O termo é usado como sinônimo de *KDD* e também como um subconjunto deste, geralmente relacionado à análise de dados estruturados. Sendo assim, na visão mais abrangente, mineração de dados diz respeito a todo o processo de descoberta de padrões, desde a limpeza dos dados até a apresentação do conhecimento extraído, por isso sinônimo de *KDD*, enquanto na visão mais restrita ela se limita a uma etapa do processo, executada após a estruturação dos dados e antes das avaliações dos padrões encontrados.

Considerando o conceito mais amplo de mineração de dados, pode-se definir *text mining* (mineração de dados textuais) como seu subconjunto. Mineração de dados textuais se refere ao processo de descoberta de padrões interessantes e extração de conhecimento de textos.

Os dados textuais são considerados não estruturados, ou seja, eles não se ajustam aos

[6] *"AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments[...]"*

moldes tradicionais de bases de dados, facilmente enquadrados em matrizes, exigindo assim uma série de transformações antes de serem aplicadas técnicas como as relacionadas ao aprendizado automático.

Somado à falta de estrutura desse tipo de dado, tem-se ainda o fato de os aspectos linguísticos influenciarem no desempenho e na quantidade de ferramentas hoje disponíveis para se trabalhar com mineração de texto. A utilização de determinadas línguas pode tornar as tarefas mais árduas ou mais tangíveis.

Existem atualmente cerca de 6 000 línguas em uso no mundo. Algumas possuem um percentual de difusão maior do que outras, bem como estruturas diferenciadas. As línguas são dinâmicas, isto é, os idiomas nascem e morrem. De acordo com Castilho (2017), não se sabe como nasceram todas as línguas do mundo, mas conhece-se como elas morrem. A morte de uma língua ocorre através da sua substituição ou por desaparecimento da comunidade que a fala. A substituição linguística pode estar relacionada com uma invasão cultural e material suficientemente forte para desorganizar a cultura do povo invadido. Já o desaparecimento pode ter como causa o extermínio da população ou razões econômicas, como quando a emigração suplanta os nascimentos.

Além dos idiomas nascerem e morrerem, as palavras também surgem, se modificam (tanto no significado quanto na escrita) e desaparecem. Segundo Bagno (1999, p.117), "A Língua é viva, dinâmica, está em constante movimento – toda língua viva é uma língua em decomposição e em recomposição, em permanente transformação".[7] Um exemplo do processo de transformação do significado de uma palavra pôde ser observado no exercício de tentar definir "ciência de dados" e outras expressões relacionadas ao assunto. A dificuldade na determinação de um sentido exato para cada termo está possivelmente relacionada ao momento de ascensão da área, o que caminha junto com a descoberta de novas questões associadas a ela e, por vezes, à necessidade de reformulação de ideias mais antigas.

Nesse contexto de popularização e aprimoramento das tecnologias, megadados em conjunto com técnicas de ciência de dados passaram a ser aplicados em diferentes campos, sendo um deles a produção de estatísticas oficiais. A sua utilização vem acompanhada de debates, já em voga nos últimos anos, a respeito das oportunidades e desafios.

2.2 Uso de Megadados e de Ferramentas da Ciência de Dados em um Órgão Oficial de Estatística

A principal atribuição dos órgãos oficiais de estatística é produzir estatísticas oficiais de alta qualidade e imparciais para pautar tomada de decisões e permitir que os indivíduos, de maneira geral, tenham acesso a informações a respeito do País, viabilizando a existência de uma democracia. Por conta disso, as estatísticas oficiais são consideradas um bem público. Rolland (2017) destaca que bem público possui dois significados, um coloquial

[7] Vale destacar que no texto de origem, diferente do aqui escrito, o trecho se insere no contexto do preconceito linguístico, que, de acordo com o autor, está ligado à confusão entre língua e gramática normativa. Segundo ele, a gramática tenta caracterizar a língua como algo fechado e acabado, mas não é assim. A língua é viva, são pessoas vivas que a falam, não sendo a gramática a língua.

que o caracteriza como um bem produzido pelo governo e, geralmente, disponível para o benefício de seus cidadãos, e um científico, que o define como um bem não excludente e não rival no consumo[8]. De acordo com o glossário da OCDE, "estatísticas oficiais são estatísticas divulgadas pelo sistema estatístico nacional, exceto aquelas que são explicitamente declaradas não oficiais."(tradução da autora)[9].

De acordo com o IBGE (2021c), dada a necessidade de se produzir dados adequados e confiáveis, a Comissão de Estatística das Nações Unidas adotou, em 1994, os Princípios Fundamentais das Estatísticas Oficiais das Nações Unidas (UNSD, 2022c). Existem, ao todo, dez princípios fundamentais: relevância, imparcialidade e igualdade de acesso; padrões profissionais e ética; responsabilidade e transparência; prevenção do mau uso dos dados; eficiência; confidencialidade; legislação; coordenação nacional; uso de padrões internacionais; e cooperação internacional.

A aceitação das estatísticas oficiais depende da sua pertinência e da credibilidade e boa reputação dos órgãos responsáveis. Assim, em um cenário onde um novo conjunto de meios e métodos, baseados em megadados e em mecanismos da ciência de dados, vem ganhando notoriedade é necessário que esses institutos sejam prudentes ao avaliar, pautados em seus princípios, os benefícios e riscos do uso dessas tecnologias na produção de informações.

De acordo com Yung, Karkimaa *et al.* (2018), apesar dos dados digitais disponíveis hoje serem oriundos de processos não estatísticos, eles contêm sinais de atividade humana que podem ser explorados para a produção de estatísticas oficiais. O uso dos megadados para produzir informações honestas, precisas e não viesadas, que podem ser usadas para a formulação de políticas baseadas em evidências, é um grande desafio para os produtores de estatísticas. O aprendizado automático é uma ferramenta indispensável para tentar resolver esse problema. Além disso, o fato de o aprendizado automático habilitar computadores para aprender tarefas estatísticas instiga uma revisão do processo de produção de estatísticas atual para identificar em que etapas o ferramental por ser útil.

A revisão do processo atual de produção já começou a ser feita. Os algoritmos de aprendizado automático têm se mostrado úteis na realização de tarefas de classificação, imputação, detecção de informações problemáticas (como problemas de preenchimento de questionário), entre outras. As ferramentas vêm cada vez mais sendo testadas, em diferentes etapas da produção das pesquisas, e alguns institutos já as aderiram. Além disso, se vislumbra, ainda que em fase embrionária, que os megadados possam substituir parcialmente[10] ou totalmente fontes de dados existentes hoje, como pesquisas amostrais, como também fornecer dados complementares a essas e, até mesmo, permitir a criação de novas estatísticas. Todos esses recursos tecnológicos são capazes de contribuir para

---

[8] O significado coloquial fornecido por Rolland (2017) tem origem em Holcombe (1997) e o científico em Samuelson (1954). Uma discussão mais ampla sobre estatísticas oficiais e bens públicos pode ser encontrada em Rolland (2017).

[9] *"Official statistics are statistics disseminated by the national statistical system, excepting those that are explicitly stated not to be official."*

[10] Atualmente o IBGE já faz uso de técnicas de *web scraping* para coletar preços de passagens aéreas e de transportes por aplicativo para a produção de Índices de Preços ao Consumidor. Os detalhes podem ser vistos em IBGE (2020b).

a redução do tempo de produção de uma pesquisa, pois com a proliferação de dados prontamente acessíveis informações defasadas se fazem cada vez menos proveitosas.

Todavia, não se pode deixar de destacar os desafios associados às oportunidades. Uma tarefa árdua é garantir a representatividade e a harmonização dos megadados. Para que esses dados sirvam de fonte para estatísticas é preciso saber quais populações são representadas por eles, pois sem essa informação não é possível garantir a sua imparcialidade, transparência e relevância, por exemplo. A geração de estatísticas exige a correspondência entre o fenômeno que se objetiva averiguar e a base de dados a ser utilizada. Além disso, outra questão sensível está relacionada à incerteza a respeito da continuidade dos dados. Existe o risco de determinados megadados deixarem de existir ou passarem a retratar um fenômeno diferente.

Outro tópico que merece menção é o acesso aos dados. Os megadados não são necessariamente de domínio público. Organizações como Google, Amazon e provedores de serviços de telefonia celular estão coletando grandes quantidades de dados e os analisando de acordo com os seus interesses. Constituindo, muitas vezes, essas análises o centro do modelo de negócios das companhias, pois com elas se busca entender os comportamentos de seus usuários ou clientes para obter vantagens sobre seus concorrentes. O uso desse tipo de dado, em muitos países, por um órgão oficial de estatística vai ao encontro de questões legais. No entanto, embora existam leis semelhantes que abrangem organizações do setor privado, a implementação e o cumprimento, provavelmente, variam de acordo com a empresa (Yung, 2021). Vale mencionar também que ainda que os megadados sejam passíveis de consulta pelo público geral, por uma questão de privacidade existe a possibilidade de não estarem disponíveis.

No âmbito do uso do aprendizado automático, pode-se também apontar alguns desafios. Um deles está relacionado a uma questão conhecida como *concept drift* (mudança de conceito). De acordo com Puts e Daas (2021), modelos têm natureza estacionária e, por conta disso, são capazes de detectar mudanças no número de ocorrências de um determinado conceito em uma população. Entretanto, podem ocorrer mudanças na forma como o conceito é expresso pela população, sendo os resultados obtidos através da modelagem também afetados. Assim, não há representatividade ao longo do tempo. Além disso, pode-se ainda acrescentar que muitas vezes é difícil justificar como um algoritmo gerou uma solução específica e, portanto, garantir transparência e reprodutibilidade. Ademais, tem-se o fato do aprendizado automático ser, por vezes, associado à ideia de que as decisões tomadas pelas máquinas são mais consistentes que as tomadas por humanos, porém, se não for claro para os humanos em que tipo de base de dados e como um algoritmo atua, ele pode ser um mecanismo de perpetuação de valores já postos pela sociedade como inadequados ou de continuidade de injustiças já detectadas.[11].

Ainda no contexto de responsabilidade associada à utilização de dados, pode-se citar o termo *mathwashing* (lavagem matemática) cunhado por Fred Benenson, o segundo contratado da Kickstarter e ex-vice-presidente de dados dessa empresa. Em uma entrevista a Woods (2016) ele esclarece que lavagem matemática deveria ser um aviso para que os cientistas não negligenciem a subjetividade inerente de se utilizar dados só porque

[11] Mais detalhes podem ser vistos em O'Neil (2020) e em Zhou *et al.* (2021).

estão usando a matemática. Ele enfatiza que algoritmos e produtos baseados em dados sempre refletirão as escolhas dos humanos que o construíram, e é irresponsável assumir o contrário. Destaca também que embora a matemática possa ajudar a descrever os dados subjacentes de maneira objetiva, a ciência de dados é muito mais do que isso, ela trata do desenvolvimento de produtos usando as observações dos cientistas sobre o mundo, medindo fatos complexos como o comportamento humano e fenômenos naturais. Ele conclui dizendo que não existem dados perfeitos e que se forem coletados incorretamente, nenhuma matemática pode impedir a construção de um produto ruim ou perigoso. Assim, para se ater aos números, existe a necessidade de ser intelectualmente honesto sobre a compreensão de como esses números foram registrados.

Nessa mesma linha, Yung, Karkimaa *et al.* (2018) ressaltam que uma recomendação importante, para garantir qualidade, reprodutibilidade e transparência, é que se desenvolva uma estrutura de qualidade específica para o uso dessas novas tecnologias. A justificativa para tanto está no fato das estruturas de qualidade para as estatísticas tradicionais, diferentes das novas que estão sendo pesquisadas, pressuporem que o processo de geração de dados e outras etapas de processamento são explicitamente conhecidas. Somado a isso, destacam também que parece útil projetar diretrizes para reprodutibilidade/ transparência/ inferência causal de estatísticas baseadas em aprendizado automático. E, não menos importante, relatam ainda que deve-se estar ciente do possível viés algorítmico e da falta de justiça ao aplicar as técnicas de aprendizado automático, se fazendo muito importante estar comprometido com análises posteriores para diminuir os riscos.[12]

Inserido nesse cenário, tem-se, no Brasil, o Instituto Brasileiro de Geografia e Estatística que é o órgão oficial de estatística do País. O Instituto tem o seu trabalho pautado nos Princípios Fundamentais das Estatísticas Oficiais das Nações Unidas e também em outros códigos de ética e boas práticas.

2.3 Instituto Brasileiro de Geografia e Estatística – IBGE

O Instituto Brasileiro de Geografia e Estatística, entidade da administração pública federal, é o principal provedor de dados e informações oficiais do Brasil. A sua missão institucional é "Retratar o Brasil com informações necessárias ao conhecimento de sua realidade e ao exercício da cidadania."(IBGE, 2021d).

[12] Entendendo a necessidade do desenvolvimento de uma estrutura de qualidade específica para o uso dessas novas tecnologias a Equipe de Machine Learning da UNECE realizou uma primeira tentativa voltada para algoritmos. A proposta, que recebeu o nome de *Quality Framework for Statistical Algorithms*, é baseada em cinco dimensões: *explainability*, *accuracy*, *reproducibility*, *timeliness* e *cost effectiveness*. Mais detalhes podem ser vistos em Yung, Tam *et al.* (2022).

As informações geradas pelo IBGE atendem às necessidades dos mais diversos segmentos da sociedade civil e órgãos das várias esferas governamentais. O IBGE (2021e) aponta que dentre os usuários das suas informações estão:

> **Setor público** – instituições governamentais crescentemente utilizam as pesquisas do IBGE como subsídio para formular e avaliar políticas públicas, bem como para planejar ações de impacto sobre a sociedade. São informações que podem auxiliar o processo de tomada de decisão e gerar benefícios a toda a coletividade.
> **Setor privado** – os dados do IBGE são referência para empresas e estabelecimentos na definição de suas ações de posicionamento no mercado, no conhecimento de características estruturais e conjunturais da economia brasileira, na realização de estudos para prospecção de potenciais clientes e consumidores, e na definição de sua política de preços e de investimentos.
> **Sociedade civil** – organizações não-governamentais, associações de classe e profissionais, líderes comunitários, outras entidades e grupos sociais organizados utilizam os dados do IBGE para conhecer características da sociedade brasileira, bem como do público-alvo de suas ações, tornando seu planejamento mais consistente e eficaz.
> **Comunidade científica e estudantes** – instituições de pesquisa, universidades, profissionais do meio acadêmico e estudantes de todos os níveis encontram nas informações do IBGE os alicerces para a produção e multiplicação do conhecimento nas mais diversas áreas do saber.
> **Indivíduos** – qualquer cidadão pode ter acesso às informações do IBGE e conhecer em maiores detalhes características do País, de seu estado ou município, tornando-se mais consciente de sua realidade.

Dentre as principais funções do Instituto, pode-se mencionar a produção e análise de informações estatísticas e geográficas, a coordenação e consolidação das informações estatísticas e geográficas, a estruturação e implantação de um sistema de informações ambientais, a documentação e disseminação de informações e a coordenação dos sistemas estatístico e cartográfico nacionais. O IBGE assume, além do papel de produzir estatísticas, a posição de coordenador do Sistema Estatístico Nacional Brasileiro.

2.4 O Sistema Estatístico Nacional e as Convenções

O Sistema Estatístico Nacional Brasileiro compreende órgãos e entidades que exercem atividades estatísticas com o objetivo de possibilitar o conhecimento da realidade física, econômica e social do País. As estatísticas produzidas por eles podem ser primárias[13], derivadas[14] ou ainda uma sistematização de dados sobre meio ambiente e recursos naturais com referência a sua ocorrência, distribuição e frequência (Brasil, 1974).

Os dados que representam os fatos a serem estudados devem ser ordenados, sistematizados e padronizados para se transformarem em informação, em outras palavras,

[13] Contínuas e censitárias

[14] Indicadores econômico e sociais, sistemas de contabilidade social e outros sistemas de estatísticas derivadas.

ganharem significância (Feijó e Valente, 2006). As estatísticas produzidas pelo Sistema Estatístico Nacional precisam ser baseadas em uma codificação prévia (explícita ou implícita), estabelecida por produtores e usuários, para alcançarem o que propõem, ou seja, serem informativas e, possivelmente, também comparáveis.

Segundo Feijó e Valente (2006), estatísticas são um mecanismo institucionalizado para gerar informações sobre a realidade, isto é, são convenções. Elas refletem uma maneira de se perceber a realidade, não a própria. Ao mesmo tempo que as estatísticas precisam ser algo convencionado para serem aceitas, dificilmente o escopo estabelecido acompanha as transformações que se processam na esfera econômica e social. Por vezes fronteiras bem definidas em um determinado momento histórico dão lugar a situações intermediárias, podendo até mesmo haver uma ruptura. Nesses momentos, estatísticas tais como concebidas atualmente são postas em xeque. Tornar as estatísticas algo menos defasado com relação à realidade é um grande desafio.

São inúmeros os esforços dos Sistemas Estatísticos para harmonizarem nacionalmente e internacionalmente seus métodos, questionários, nomenclaturas e as definições que utilizam. Um exemplo atual pode ser verificado nas pesquisas de Inovação. Tal pesquisa pretende, dentre outros objetivos, estudar o uso de novas tecnologias, como as relacionadas à ciência de dados. Entretanto, como mencionado na Seção 2.1, esse é um campo em ascensão onde poucas definições são consensuais entre os usuários, tornando a tarefa de estabelecer uma convenção (ou um padrão) por parte dos órgãos de estatística muito mais complexa.

O reconhecimento da necessidade de se estabelecerem convenções relacionadas às estatísticas e também aos registros administrativos, muitas vezes utilizados como base para a construção das estatísticas, não é recente. Inclusive, em diversas ocasiões, se faz necessária uma sistematização maior do que a instituída pelas convenções, é preciso definir padrões. Padronizar significa garantir previsibilidade em vários aspectos, tanto no que diz respeito à qualidade quanto ao que se refere ao formato das informações a serem disponibilizadas. Um padrão pode ser refletido através de uma classificação ou, ainda, na concepção de uma publicação.

United Nations (2021, p.379, tradução da autora) caracteriza as classificações da seguinte forma:

> Classificações agrupam e organizam as informações de forma significativa e sistemática em um formato padrão que é útil para determinar a semelhança de ideias, eventos, objetos ou pessoas. Preparar uma classificação significa criar um conjunto exaustivo e estruturado de categorias mutuamente excludentes e bem descritas, muitas vezes apresentadas como uma hierarquia refletida pelos códigos numéricos ou alfabéticos atribuídos a elas[15].

No início da década de 1990, visando o uso pelos produtores de informações econômicas do Brasil, o IBGE em conjunto com outros órgãos elaboraram uma classificação de atividades econômicas. Para tanto, objetivando a comparabilidade internacional, foi

[15] *"Classifications group and organize information meaningfully and systematically into a standard format that is useful for determining the similarity of ideas, events, objects, or persons. The preparation of a classification means creating an exhaustive and structured set of mutually exclusive and well-described categories, often presented as a hierarchy reflected by the numeric or alphabetical codes assigned to them."*

tomada como referência a classificação padrão desenvolvida pela Divisão de Estatísticas das Nações Unidas, a *International Standard Industrial Classification of All Economic Activities - ISIC* (*Clasificación Industrial Internacional Uniforme - CIIU*), 3ª revisão. Antes dessa especificação não havia uma padronização nacional e uma harmonização com as classificações internacionais. Assim, para a realização de comparações eram utilizadas tabelas de conversão. (IBGE e CONCLA, 2021)

2.4.1 Classificação Nacional de Atividades Econômicas - CNAE

Segundo o IBGE (2021b), o objetivo da utilização da CNAE é:

> Unificar os códigos de atividades pelos órgãos gestores de cadastros e de registros administrativos. A desagregação das classes CNAE visa atender às necessidades de maior especificação das atividades para identificação de segmentos produtivos sujeitos à regulamentação ou a tratamento tributário específicos ou, ainda, cuja visibilidade era necessária ao gestor público.

O IBGE é órgão gestor da CNAE e presidente da Comissão Nacional de Classificação (Concla). Sendo essa última responsável pelo monitoramento, padronização e definição das normas de utilização das classificações estatísticas usadas no Sistema Estatístico e nos cadastros e registros da Administração Pública.

De acordo com IBGE e CONCLA (2020), a CNAE foi construída pautada na ideia de cobrir todo o universo representado, possuir categorias mutuamente excludentes e uma base conceitual e de princípios metodológicos que permita a alocação consistente dos elementos nas várias categorias da classificação.

Além disso, a CNAE possui uma estrutura hierárquica cujo objetivo é possibilitar o uso para diferentes propósitos estatísticos. Nos dois primeiros níveis, Seções e Divisões, a CNAE adota a estrutura da *CIIU*/*ISIC*[16], inclusive na definição dos códigos. Nos dois níveis seguintes, a CNAE introduz, sempre que necessário, uma maior especificidade para refletir a estrutura da economia brasileira, possibilitando a reconstituição das categorias da classificação internacional. O código de quatro dígitos das Classes CNAE é acompanhado de um Dígito Verificador (DV). Os dígitos seguintes representam as Subclasses CNAE. Essas Subclasses são um detalhamento das Classes e têm como finalidade atender necessidades da organização de cadastros de Pessoa Jurídica e Física no âmbito da Administração Pública, de modo especial da área tributária.

[16] Mais informações podem ser obtidas em UNSD (2022b).

Toda a estrutura e descrições da CNAE constam em IBGE e CONCLA (2020). Um exemplo de código, extraído de IBGE e CONCLA (2020, p.16), pode ser visto a seguir.

Seção	A	Agricultura, pecuária, produção florestal, pesca e aquicultura
Divisão	01	Agricultura, pecuária e serviços relacionados
Grupo	01.1	Produção de lavouras temporárias
Classe	01.11-3	Cultivo de cereais
Subclasse	0111-3/01	Cultivo de arroz

Já o número de grupamentos existentes pode ser verificado no Quadro 2.1.

Quadro 2.1: Organização hierárquica da CNAE 2.0/CNAE-Subclasses 2.3

Nome	Nível	Número de Grupamentos	Identificação
Seção	Primeiro	21	Alfabético de 1 dígito
Divisão	Segundo	87	Numérico de 2 dígitos
Grupo	Terceiro	285	Numérico de 3 dígitos
Classe	Quarto	673	Numérico de 4 dígitos + DV
Subclasse	Quinto	1 332	Numérico de 7 dígitos (incluindo DV)

Fonte: IBGE e CONCLA (2020, p.16).

Apesar do extenso número de Subclasses da CNAE, por ser estática, ela não é capaz de acompanhar a dinâmica da economia. Para evitar uma grande defasagem, a classificação econômica sofre revisões periódicas. As atualizações ao quinto nível, que não dependem de uma organização em nível internacional, ocorrem com frequência superior às demais, que ficam em função da necessidade de obediência ao calendário internacional de revisões.

Quando surge uma atividade nova, não contemplada pelos códigos existentes, ela acaba se misturando em diferentes classificações antes de ser identificada. A partir do momento que ela é reconhecida, convenciona-se a sua alocação em uma classificação específica, ainda que essa não a retrate em sua plenitude. Já quando a CNAE é revista, essa nova atividade recebe o seu próprio código, refletindo e estabelecendo um padrão. De acordo com IBGE e CONCLA (2020), os benefícios da inserção de novos códigos devem ser altos para justificar os custos de retrabalhar as bases de dados do Sistema Estatístico Nacional, os cadastros da Administração Pública e revisar séries históricas. Faz parte dos princípios de construção da CNAE a estabilidade durante um determinado período.

Um dos desafios associados à utilização da CNAE está no fato de ser grande a quantidade de empresas que desempenham diversas atividades econômicas, principalmente no contexto atual de intensificação tecnológica. Quando isso acontece, busca-se classificar a empresa de acordo com a sua atividade principal. Segundo IBGE e CONCLA (2020), a

atividade econômica está baseada na criação de valor adicionado[17]a partir da produção de bens e serviços, com a utilização de trabalho, de capital e de insumos (matérias-primas). Em consequência, a atividade principal de uma unidade de produção é definida de acordo com a que mais contribui para a geração do valor adicionado.

A tarefa de classificação não é fácil nem mesmo para a própria empresa. Visando a correta atribuição do código, o IBGE fornece instrumentos e mecanismos de apoio. Elencando-os têm-se: o banco de descritores, que lista as atividades que compõem cada classe e subclasse CNAE e alimenta o mecanismo de busca do aplicativo Pesquisa CNAE; o manual de orientação da codificação em subclasses CNAE; o roteiro de procedimentos-padrão; a central de dúvidas; entre outros.

O interesse do Instituto na correta classificação da empresa é grande. Várias estatísticas econômicas divulgadas por ele, como as Pesquisas Estruturais por Empresa, têm como pilar a CNAE existente nos seus cadastros.

2.5 Pesquisas Estruturais por Empresa e seus Cadastros

As Pesquisas Estruturais por Empresa são compostas pela Pesquisa Industrial Anual - Empresa (PIA–Empresa), pela Pesquisa Anual da Indústria da Construção (PAIC), pela Pesquisa Anual de Comércio (PAC) e pela Pesquisa Anual de Serviços (PAS)[18]. Essas pesquisas amostrais têm como objetivo fornecer informações sobre as características estruturais básicas de um determinado segmento empresarial no país. Sua unidade de investigação é a empresa[19] formalmente constituída, sediada em território nacional[20], cuja principal fonte de receita pertença a uma determinada atividade. Sendo a receita utilizada como aproximação do valor adicionado, mencionado na Subseção 2.4.1, para determinar a atividade principal da companhia.

O desenho amostral utilizado nessas pesquisas é baseado em amostragem aleatória estratificada simples com dois níveis. Assim, inicialmente, são selecionados os estratos naturais e, dentro desses, os estratos finais.

Os estratos naturais são construídos a partir da união de empresas com a mesma combinação de Unidade da Federação (UF) e agrupamentos de CNAE, exceto para as

[17] Valor Adicionado é o valor bruto da produção menos o custo das matérias-primas, bem como de outros consumos intermediários.

[18] A Pesquisa Industrial Anual - Produto (PIA-Produto) também faz parte das Pesquisas Estruturais por Empresa mas, por suas características muito diversas das demais pesquisas, não será objeto deste trabalho. A PIA-Produto é desenhada como uma subamostra intencional da PIA-Empresa.

[19] A PIA-Empresa também tem como unidade de investigação a Unidade Local (UL). A UL se refere ao espaço físico no qual uma ou mais atividades econômicas são desenvolvidas, correspondendo a um endereço de atuação da empresa ou a um sufixo do seu CNPJ. (IBGE, 2021i)

[20] Em particular para as Unidades da Federação da Região Norte (Rondônia, Acre, Amazonas, Roraima, Pará, Amapá e Tocantins), na PAC e na PAS, são consideradas apenas aquelas que estão sediadas nos Municípios das Capitais, com exceção do Pará, onde são consideradas aquelas que estão sediadas nos municípios da Região Metropolitana de Belém.

empresas com 1 a 4 pessoas ocupadas da PIA-Empresa para as quais não se considera a UF. A CNAE atribuída à empresa corresponde à sua atividade principal, ou seja, a que gerou a maior receita no ano da pesquisa.

Posteriormente, os estratos finais são obtidos pela subdivisão de cada estrato natural em outros dois estratos (certo ou amostrado) para a PIA-Empresa e para a PAIC e em outros três estratos (certo, gerencial ou amostrado) para a PAC e a PAS. As empresas do estrato certo e do gerencial são selecionadas em sua totalidade, já para as pertencentes ao estrato amostrado é executada uma seleção probabilística. A alocação das empresas em cada um desses estratos é dada pelo total de pessoal ocupado, de acordo com o cadastro, e varia conforme cada pesquisa. Para o estrato certo considera-se também, quando disponível, a receita obtida pela empresa no ano anterior e para o gerencial a quantidade de UFs de atuação. O estrato certo da PIA-Empresa e da PAIC é ainda dividido em três grupos (C1, C2 e C3) a depender do pessoal ocupado e da receita auferida no ano anterior. Já o estrato amostrado é composto por quatro (A1, A2, A3 e A4) conjuntos na PIA-Empresa e na PAIC e por três (A1, A2 e A3) na PAC e na PAS estipulados de acordo com o número de pessoas ocupadas pela empresa.

Sendo assim, a classificação econômica e a UF da sede da empresa são peças-chave do desenho amostral. Esses elementos são determinantes para a seleção da amostra e para o cálculo do peso que é atribuído às empresas no momento da expansão da amostra para representação do universo. Consequentemente, e por esse ser o objetivo, é a partir deles que está pautada toda a divulgação de informações.

A periodicidade das Pesquisas Estruturais por Empresa é anual e seus resultados são divulgados para Brasil, Grandes Regiões e Unidades da Federação. Essas pesquisas são importantes fontes de dados para o planejamento público e privado e para o cálculo do Produto Interno Bruto (PIB) pelo sistema de Contas Nacionais. Além disso, é a partir delas que se movem as demais pesquisas por empresa, tanto as de acompanhamento conjuntural, com periodicidade inferior a um ano, como as de aprofundamento temático.

Para a seleção de suas amostras, as Pesquisas Estruturais por Empresa possuem um Cadastro Básico de Seleção (CBS), de acordo com o âmbito de cada uma delas, que é obtido a partir do Cadastro Central de Empresas (Cempre).

2.5.1 Cadastro Central de Empresas - CEMPRE

O Cempre é formado por empresas e outras organizações, e suas respectivas unidades locais formalmente constituídas, registradas no Cadastro Nacional de Pessoa Jurídica (CNPJ). As principais fontes de dados que atualizam anualmente o Cempre são as Pesquisas por Empresas do IBGE, o Sistema de Manutenção Cadastral (Simcad)[21] e os registros administrativos do Ministério do Trabalho, em particular, a Relação Anual de Informações Sociais (RAIS) e o Cadastro Geral de Empregados e Desempregados (Caged) (IBGE, 2021a).

[21] Simcad é um sistema do IBGE de 'Entrevistas por Telefone Assistidas por Computador', denominado *Computer Assisted Telephone Interview — CATI*. Sua finalidade é atualizar os dados existentes no Cempre, principalmente a classificação econômica. São trabalhadas pelo Simcad algumas das empresas que não foram selecionadas pelas Pesquisas Econômicas naquele ano.

O Cempre traz informações de Unidade da Federação da sede da empresa, Classificação Nacional de Atividades Econômicas, total de pessoal ocupado, salários, entre outras. O Cadastro é uma importante fonte de dados sobre a atividade econômica do Brasil. As estatísticas oriundas do Cadastro Central de Empresas são divulgadas anualmente pelo IBGE.

De acordo com IBGE (2021a, p.19), a metodologia utilizada para a atribuição da classificação de atividade principal no Cempre segue a seguinte ordem:

- Para as organizações especiais, como as prefeituras municipais, órgãos da administração pública e algumas empresas públicas, por meio do acompanhamento da classificação ano a ano, a classificação econômica é atribuída pela Coordenação de Cadastro e Classificações;
- Para as empresas e unidades locais selecionadas nas pesquisas anuais nas áreas de Indústria, Construção, Comércio e Serviços do IBGE, a classificação econômica é atribuída pela pesquisa;
- Para as empresas e outras organizações que foram selecionadas para compor o painel de pesquisa do Simcad devido a suspeitas de erro de preenchimento do registro administrativo, a classificação econômica é atribuída pelo Simcad;
- Caso a empresa ou organização não tenha sido investigada pelas pesquisas do IBGE no ano de referência, é mantida a classificação mais recente atribuída pelas Pesquisas Anuais por Empresas ou pelo Simcad nos últimos três anos, independentemente da classificação existente no registro administrativo;
- No caso de empresas e outras organizações que possuam mais de um registro no ano de referência, as informações das Pesquisas Anuais por Empresas do IBGE ou do Simcad têm precedência sobre as informações do registro administrativo; e
- No caso de não existirem informações nas Pesquisas Anuais por Empresas do IBGE ou do Simcad, permanece a classificação econômica proveniente do registro administrativo do ano de referência.

Dado o exposto, fica nítida a importância e também a maior precisão da classificação considerada nas Pesquisas Estruturais por Empresa, bem como a relevância do Cempre para a construção dessas pesquisas.

Todo ano, a partir das informações referentes à CNAE e também de outras, é extraído do Cempre o Cadastro Básico de Seleção das Pesquisas Estruturais por Empresa.

2.5.2 Cadastro Básico de Seleção - CBS

Cada uma das pesquisas possui um Cadastro Básico de Seleção de acordo com o seu âmbito. O CBS nada mais é do que uma triagem do Cempre.

Vale destacar que o CBS que dará origem a uma Pesquisa Estrutural por Empresa de determinado ano tem como base o Cempre que, por sua vez, é composto pelas informações da RAIS do ano anterior, do Caged dos meses de janeiro a dezembro do mesmo ano das pesquisas e das Pesquisas Estruturais por Empresas do ano anterior.

O CBS possui as variáveis necessárias para a construção de cada uma das amostras detalhadas a seguir.

2.5.3 Pesquisa Industrial Anual - PIA–Empresa

A PIA-Empresa, de acordo com IBGE (2021i), fornece informações sobre o segmento empresarial da atividade industrial no país. Fazem parte do estrato certo da pesquisa as empresas com mais de 30 pessoas ocupadas e/ou que auferiram receita bruta, oriunda das vendas de produtos e serviços industriais, superior a um determinado valor no ano anterior ao de referência da pesquisa. Em 2019, utilizou-se o corte de R$15,8 milhões. Para as empresas do estrato certo também são investigadas informações a respeito das suas ULs. Já as empresas que ocuparam entre 1 e 29 pessoas compõem o estrato amostrado.

O âmbito da PIA-Empresa inclui empresas com atividade principal, de acordo com o Cempre, compreendida nas Seções B ou C (Indústrias extrativas ou Indústrias de transformação, respectivamente) da CNAE 2.0. A estrutura da atividade industrial na CNAE 2.0, na PIA-Empresa, pode ser vista no Quadro 2.2.

Quadro 2.2: Categorias da CNAE 2.0 - PIA-Empresa

Nível	Código	Número de categorias da indústria
Seção	Alfabético de 1 dígito	2
Divisão	Numérico de 2 dígitos	29
Grupo	Numérico de 3 dígitos	111
Classe	Numérico de 4 dígitos + DV	274

Fonte: IBGE (2021i, p.8).

A pesquisa divulga informações em níveis geográficos diferentes a depender do número de pessoas ocupadas na empresa. Para empresas com 1 ou mais pessoas ocupadas, os resultados são disponibilizados para Brasil, por Divisões da CNAE. Para empresas com 5 ou mais pessoas ocupadas são divulgados dados do Brasil por Divisões e Grupos da CNAE e, além disso, são publicados dados por Divisões e Grupos CNAE para Minas Gerais, Rio de Janeiro, São Paulo, Paraná, Santa Catarina e Rio Grande do Sul e dados por Divisões da CNAE para as demais UFs. Já para as empresas do estrato certo é possível obter informações mais detalhadas, por Classes da CNAE e por municípios.

2.5.4 Pesquisa Anual da Indústria da Construção - PAIC

A PAIC, como pode ser visto em IBGE (2021f), fornece informações sobre o segmento empresarial da atividade de construção no país. Fazem parte do estrato certo da pesquisa as empresas com mais de 30 pessoas ocupadas e/ou que auferiram receita bruta oriunda da construção superior a um determinado valor no ano anterior ao de referência da pesquisa. Em 2019, utilizou-se o corte de R$14,9 milhões. Já as empresas que ocuparam entre 1 e 29 pessoas compõem o estrato amostrado.

O âmbito da PAIC inclui empresas com atividade principal, de acordo com o Cempre, compreendida na Seção F (Construção) da CNAE 2.0. A atividade da construção na CNAE 2.0, na PAIC, estrutura-se da seguinte forma:

Quadro 2.3: Categorias da CNAE 2.0 - PAIC

Nível	Código	Número de categorias da construção civil
Seção	Alfabético de 1 dígito	1
Divisão	Numérico de 2 dígitos	3
Grupo	Numérico de 3 dígitos	9
Classe	Numérico de 4 dígitos + DV	21

Fonte: Compilação da autora a partir de IBGE (2021f).

O IBGE (2021f, p.20) descreve como ocorre a disseminação dos resultados da pesquisa.

> Para Brasil, as informações do conjunto de empresas que ocupam 1 a 4 pessoas são apresentadas por Divisão da CNAE 2.0 (dois dígitos da classificação). Para as empresas cujo total de pessoal ocupado varia de 5 a 29 pessoas, a abertura se dá no nível de grupo (três dígitos). Por fim, para as empresas com 30 ou mais pessoas ocupadas, as informações são apresentadas por Classe (quatro dígitos, nível mais desagregado da classificação). Apresentam-se, também, as informações segundo o grupo e a faixa de pessoal ocupado.

2.5.5 Pesquisa Anual de Comércio - PAC

A PAC, de acordo com IBGE (2021g), fornece informações sobre o segmento empresarial do comércio atacadista e varejista no país. Fazem parte do estrato certo da pesquisa as empresas com mais de 20 pessoas ocupadas e/ou que individualmente apresentam receita total no mesmo patamar das empresas do estrato certo da pesquisa do ano anterior. Já as empresas com menos de 20 pessoas ocupadas e que atuam em mais de uma UF fazem parte do estrato gerencial. Por fim, as empresas que ocuparam entre 0 e 19 pessoas compõem o estrato amostrado.

O âmbito da PAC inclui empresas com atividade principal, de acordo com o Cempre, compreendida na Seção G (Comércio; reparação de veículos) da CNAE 2.0. Entretanto, por mais que façam parte da Seção G, são excluídos serviços de manutenção e reparação de veículos automotores e motocicletas e comércio ambulante e outros tipos de comércio varejista. A atividade de comércio na CNAE 2.0, na PAC, estrutura-se da seguinte forma:

Quadro 2.4: Categorias da CNAE 2.0 - PAC

Nível	Código	Número de categorias do comércio
Seção	Alfabético de 1 dígito	1
Divisão	Numérico de 2 dígitos	3
Grupo	Numérico de 3 dígitos	20
Classe	Numérico de 4 dígitos + DV	92

Fonte: Compilação da autora a partir de IBGE (2021g).

A pesquisa divulga resultados do total de empresas comerciais, em nível Brasil, segundo Divisões, Grupos e Classes de atividades. Além disso, apresenta resultados de acordo com as Grandes Regiões e as UFs de atuação das empresas por Divisões e Grupos de atividades. Já exclusivamente para as empresas com 20 ou mais pessoas ocupadas é possível obter informações segundo Divisões, Grupos e Classes de CNAE. O desenho amostral da PAC e, consequentemente, também a sua disseminação sofrem alterações de acordo com a UF da sede da empresa e com a sua classificação econômica, podendo todos os detalhes serem vistos em IBGE (2021g).

2.5.6 Pesquisa Anual de Serviços - PAS

A PAS, como pode ser visto em IBGE (2021h), fornece informações sobre o segmento empresarial de serviços não financeiros no país. Fazem parte do estrato certo da pesquisa as empresas com mais de 20 pessoas ocupadas e/ou que individualmente apresentam receita total no mesmo patamar das empresas do estrato certo da pesquisa do ano anterior. Já as empresas com menos de 20 pessoas ocupadas e que atuam em mais de uma UF fazem parte do estrato gerencial. Por fim, as empresas que ocuparam entre 0 e 19 pessoas compõem o estrato amostrado.

Diferente das outras Pesquisas Estruturais por Empresa, a PAS abrange um conjunto de atividades com características econômicas diversificadas e genericamente referidas como setor produtor de serviços, englobando assim várias Seções da CNAE. As atividades que constam na PAS, de acordo com a CNAE 2.0, estruturam-se da seguinte forma[22]:

Quadro 2.5: Categorias da CNAE 2.0 - PAS

Nível	Código	Número de categorias do serviço
Seção	Alfabético de 1 dígito	13
Divisão	Numérico de 2 dígitos	38
Grupo	Numérico de 3 dígitos	88
Classe	Numérico de 4 dígitos + DV	166

Fonte: Compilação da autora a partir de IBGE (2021h).

[22] Vale destacar que, assim como detalhado para a PAC, dizer que a PAS engloba 13 Seções não implica informar que todas as Divisões oriundas dessas Seções integrem a Pesquisa. Todos os detalhes podem ser vistos em IBGE (2021h).

O plano tabular completo da PAS compreende 81 tabelas que podem ser divididas em três partes. A primeira engloba uma tabela-resumo que confronta os dados do ano de referência com os do ano anterior, contemplando as principais variáveis referentes aos serviços empresariais não financeiros, e mais duas tabelas com informações sobre a origem da receita operacional líquida e outras variáveis pesquisadas, por atividades. Já do segundo grupo fazem parte 77 tabelas que apresentam resultados em nível nacional para cada um dos 7 grupamentos estudados na PAS – Serviços prestados principalmente às famílias; Serviços de informação e comunicação; Serviços profissionais, administrativos e complementares; Transportes, serviços auxiliares aos transportes e correio; Atividades imobiliárias; Serviços de manutenção e reparação; e Outras atividades de serviços. Por último, no terceiro grupo, estão as tabelas com dados regionalizados, segundo as Grandes Regiões, as Unidades da Federação e as atividades. IBGE (2021h) traz todos os pormenores sobre o assunto.

Capítulo 3

Avaliação Preliminar e Construção das Bases de Dados

O IBGE utiliza em suas pesquisas a Classificação Nacional de Atividades Econômicas para categorizar empresas de acordo com a sua atividade principal. Estabelecer essa classificação não é uma tarefa simples, ela exige a elaboração de diversos materiais pelo Instituto e demais órgãos responsáveis por orientar as empresas, além de mobilizar diversos técnicos para realizar atividades de conferência da CNAE. Dito isso, fica a indagação se seria possível utilizar as diferentes bases de dados hoje disponíveis em conjunto com as novas tecnologias e as técnicas estatísticas para contribuir com a empreitada.

Este trabalho combinou dados da PIA-Empresa, PAC e PAS, ano de referência 2019, com dados textuais, coletados em maio de 2021 via *web scraping*, dos *sites* de algumas das empresas que compõem essas pesquisas. Ele teve como objetivo central avaliar a possibilidade de se classificar a atividade econômica principal de uma empresa através de técnicas de mineração de dados textuais, com base no texto da sua *URL*, no texto extraído do seu *site* e na CNAE utilizada pelo IBGE para classificar a companhia.

Devido à defasagem temporal de, aproximadamente, dois anos, foram utilizadas também informações referentes aos anos de 2014 a 2019 da PAIC, PIA-Empresa, PAC e PAS para realização de uma avaliação preliminar com o intuito de determinar qual percentual de empresas mudou de atividade principal, de acordo com as pesquisas, em um triênio. Optou-se pelos anos mais recentes pois acreditou-se que eles seriam mais aptos a refletir o período entre 2019 e 2021, apesar da pandemia de Coronavírus (SARS-CoV-2, chamado de Covid-19).

Nota-se que para o estudo inicial foram empregados dados da PAIC, mas não para o estudo principal. Essa escolha se deu pelo fato da PAIC possuir 22 961 empresas na sua amostra e cerca de 2 000 *sites* disponíveis, quantidades essas muito inferiores às das demais pesquisas utilizadas.

Assim, este capítulo apresenta inicialmente um estudo preliminar que buscou avaliar se o descompasso temporal entre a coleta dos dados da *web* e a CNAE utilizada pelo IBGE inviabilizaria a tarefa de se obter a classificação através dos textos dos *sites*. Na sequência, estabelece quais empresas pertencentes às pesquisas estruturais foram objeto do estudo

principal, bem como quais variáveis foram utilizadas. A seguir, fornece uma descrição de como os dados disponíveis na *internet* foram coletados e quais procedimentos empregados para a sua limpeza. Ao final, apresenta uma análise descritiva da base que alimentou a modelagem.

3.1 Dados das Pesquisas Estruturais por Empresa

Os dados das Pesquisas Estruturais por Empresa, descritas na Seção 2.5, foram obtidos através de questionários, com formatos distintos para cada uma delas, disponibilizados para o preenchimento via aplicativo *web*.

O trabalho utilizou dados relativos à CNAE da respondente, ao seu *site*, à UF da sua sede e ao estrato de seleção[1]. As informações são solicitadas para as empresas no primeiro semestre de cada ano e são referentes ao ano anterior, sendo este denominado ano de referência da pesquisa. Os dados são publicados pelo Instituto no ano seguinte ao do seu recebimento.

As informações permitiram a realização de uma avaliação preliminar, para embasar a realização do estudo principal, e também foram fundamentais para construção da base utilizada na execução deste último.

3.1.1 Avaliação Preliminar

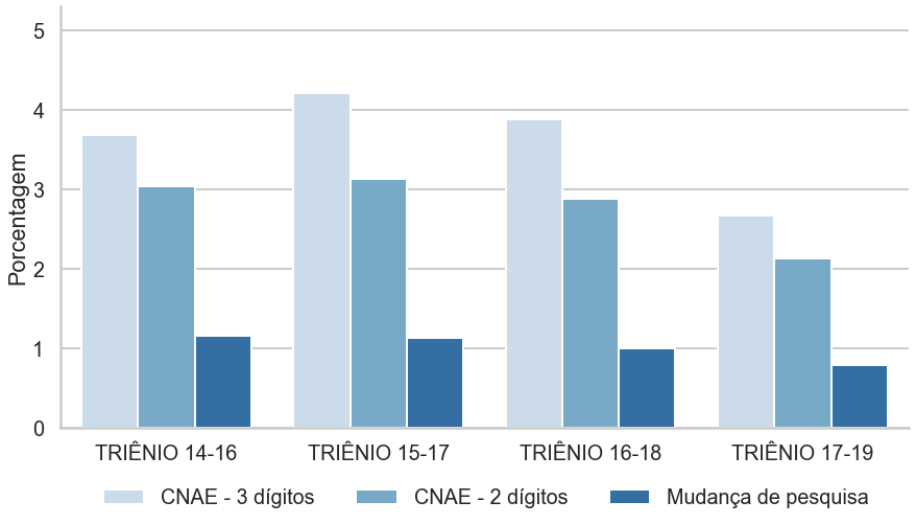
O estudo preliminar visou avaliar a factibilidade da utilização da CNAE principal da empresa em 2019 em conjunto com os dados dos *sites* recolhidos no primeiro semestre de 2021.

Nesse passo foram consideradas as informações das pesquisas - PAIC, PIA-Empresa, PAC e PAS - de 2014 a 2019. As variáveis utilizadas foram um identificador único para cada respondente (CNPJ) e a sua classificação econômica principal. Destaca-se ainda que foram consideradas apenas as empresas que preencheram o questionário de alguma das pesquisas em, pelo menos, dois anos do triênio a ser avaliado. Assim, o total de empresas em cada triênio variou entre, aproximadamente, 204 000 e 212 000.

Na Figura 3.1 é possível verificar o percentual de respondentes que sofreram alteração da classificação econômica, levando em consideração Divisão e Grupo CNAE, nos triênios de 2014 a 2016, 2015 a 2017, 2016 a 2018 e 2017 a 2019. Pode-se também observar o percentual de empresas que mudaram de pesquisa, considerando a PAIC, a PIA-Empresa, a PAC e a PAS, nos períodos.

[1] Destaca-se que o uso de dados restritos ao Instituto se deu pelo fato de a autora ser servidora do IBGE, lotada na Coordenação de Serviços e Comércio, e ter desenvolvido o trabalho em gozo de afastamento para realização de programa de pós-graduação, mediante submissão de projeto de pesquisa, consoante com as regras do Instituto.

Figura 3.1: Empresas (%) que mudaram de CNAE, segundo os níveis da classificação econômica, em um triênio



Fonte: Autoria própria.

Os valores apresentados não exibem transformações profundas quando os triênios são comparados, exceto pelo mais recente, de 2017 a 2019, no qual são verificados percentuais mais baixos, principalmente, para níveis de maior detalhamento da CNAE. Esse acontecimento possivelmente está relacionado com a política de revisão de dados das pesquisas, na qual ao se divulgar um ano são sempre revistos os resultados dos dois anos anteriores. Assim, quando as informações referentes a 2020 forem divulgadas, serão analisados novamente os dados de 2018 e 2019. Além disso, também é importante destacar que uma empresa que mudou de pesquisa no triênio, automaticamente, teve também a sua CNAE a 2 e a 3 dígitos[2] alterada. Sendo assim, todas as empresas que mudaram de pesquisa compõem o total de empresas que mudou de CNAE ao se considerar os 2 dígitos e, da mesma forma, todas as que mudaram ao se considerar os 2 dígitos também constituem o total das que apresentaram mudanças a 3 dígitos.

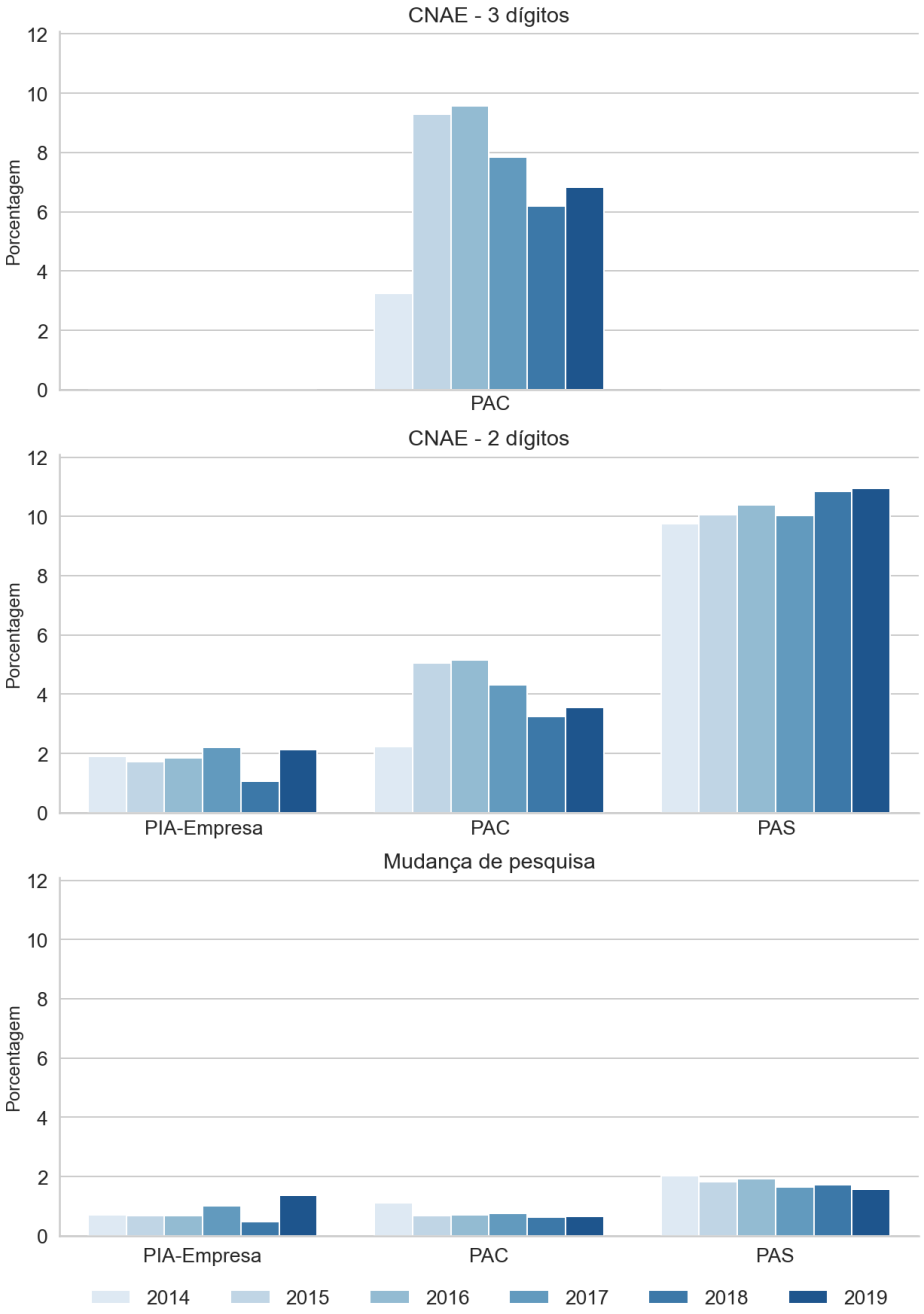
Os valores apontados em todos os períodos são baixos, em especial, para menores níveis de detalhamento da CNAE. Tal situação fornece indícios de que é factível confrontar dados textuais coletados em 2021 com a informação fornecida pela empresa em 2019.

Entretanto, um outro ponto que merece atenção é o fato de que para uma companhia ter a sua classificação econômica alterada no triênio é necessário que um técnico das Pesquisas Estruturais por Empresa a averigue e efetue a mudança manualmente ou que a classificação mude por outros meios, como o Simcad, no Cempre. Assim, fica explícito que os resultados apresentados na Figura 3.1 sofrem influência da capacidade de absorção de trabalho das equipes envolvidas na tarefa de classificação. E, mais do que isso, que o fato de a empresa ter mudado de CNAE pode significar que sua atividade principal realmente mudou ou que a usada anteriormente não era a mais adequada.

[2] Com exceção da Divisão 45 (Comércio e reparação de veículos automotores e motocicletas) que consta na PAC e na PAS. Empresas com essa classificação podem mudar de pesquisa sem sofrer alteração na CNAE a 2 dígitos. O mesmo se aplica ao Grupo 454 (Comércio, manutenção e reparação de motocicletas, peças e acessórios), que também pertence às duas pesquisas.

De forma a tentar ter uma dimensão do quanto a capacidade de absorção de trabalho das equipes interfere na análise, a Figura 3.2 exibe o percentual de CNAEs, não verificadas pelos técnicos do Instituto, informadas pelas empresas diferentes das CNAEs registradas no CBS por pesquisa, ano e detalhamento da classificação. Vale destacar que para essa análise, diferente da anterior, foram consideradas todas as empresas que preencheram um questionário por pesquisa e ano, sendo a quantidade média dessas empresas, considerando os anos de 2014 a 2019, de 46 000 na PIA-Empresa, 70 000 na PAC e 95 000 na PAS.

Figura 3.2: Empresas (%), sem verificação da CNAE por técnicos, com divergência entre a CNAE do cadastro de seleção e a informada na pesquisa, por pesquisa, segundo os níveis da classificação econômica - 2014-2019



Fonte: Autoria própria.

Antes da análise da Figura 3.2 é importante frisar que o fato de a CNAE do informante ser diferente da CNAE de seleção não implica que a classificação usada na seleção esteja incorreta, mas traz indícios de que ela merece atenção. Além disso, vale também pontuar que alterar a atividade econômica de uma empresa após a seleção da amostra não é uma tarefa trivial, visto que seu peso amostral tem como função representar a população-alvo de empresas da atividade na qual foi selecionada. Da mesma forma, manter uma empresa com a classificação que não lhe é devida não soluciona o problema, pois pode ser que ela apresente dados muito destoantes das outras companhias que efetivamente pertencem àquela atividade, gerando resultados distorcidos.

Dificilmente todas as empresas computadas na Figura 3.2 se enquadram no perfil que se busca quantificar, de companhias que tiveram a sua atividade econômica alterada em um triênio. Assim, embora a PAS possua, em alguns anos, um percentual próximo dos 11% de empresas com CNAEs de seleção diferentes das CNAEs fornecidas pelos informantes e a PAC um percentual que se aproxima dos 10% ao se considerar os 3 dígitos da classificação, o estudo principal proposto parece viável.

Apesar da conclusão ao que diz respeito à utilização da classificação aplicável em 2019 em conjunto com os dados coletados em 2021, o percentual de CNAEs de seleção diferentes das CNAEs fornecidas pelos informantes representa, em valores absolutos, uma grande quantidade de empresas para serem avaliadas manualmente por um corpo técnico cada vez menor, o que serve de motivação para o estudo principal.

Ainda com relação à Figura 3.2, convém mencionar que tais discrepâncias podem também ser justificadas ao se verificar as distintas naturezas das atividades que compõem essas categorias.

A despeito do fato da classificação econômica referente às empresas industriais ser composta por 2 Seções e 29 Divisões, indicando que as atividades nelas desenvolvidas possuem importantes particularidades, essas empresas possuem características estruturais mais parecidas entre si do que aquelas de outras categorias, como as de Comércio e Serviços. Sendo o principal objetivo de suas atividades a transformação de matérias-primas (insumos) em bens tangíveis (produtos), elas se valem da instalação de uma estrutura produtiva concreta, associada a custos fixos (custos que independem do nível de produção) e variáveis (que dependem da quantidade produzida) que se traduzem no uso de mão-de-obra, energia, transporte e ativos imobilizados em geral. Tais custos se referem àqueles relacionados às economias de escala[3] e escopo[4], resultado de investimentos em ativos que não podem ser transacionados sem perda total ou parcial de seu valor, os chamados custos irrecuperáveis. A necessidade de adaptações e mudanças nesta estrutura produtiva, evidentemente mais rígida, dependerá dos custos de oportunidade a eles associados, ou seja, aqueles relacionados às avaliações das melhores alternativas de alocação de recursos empregados para os objetivos daquelas atividades. Nesse sentido, as mudanças associadas a tal estrutura são tanto menos flexíveis e dinâmicas, quanto mais custosas e, portanto, menos sujeitas às variações de curto e médio prazo, do que aquelas verificadas nas empresas

[3] "Reduções de custos médios derivados do aumento da produção, por meio da diluição de custos fixos e uso de máquinas e equipamentos com maior capacidade."(Tigre e Pinheiro, 2019, Glossário, p.277).

[4] "Reduções de custos derivados da produção conjunta de pelo menos dois produtos/serviços."(Tigre e Pinheiro, 2019, Glossário, p.277).

de outras atividades.

As empresas de Serviços, por sua vez, abarcam um conjunto um tanto mais amplo de atividades econômicas do que o das empresas industriais. Composta por 13 Seções e 38 Divisões, as empresas definidas como de Serviços no âmbito da PAS se dividem em distintos e heterogêneos grupos de empresas cujas atividades variam tanto em sua natureza quanto finalidade.

Essas atividades incluem desde serviços pessoais simples até serviços complexos e intensivos em conhecimentos. Em geral, esse conjunto de atividades possui entre si algumas características em comum quando comparados aos produtos físicos resultados da produção industrial. São eles: a intangibilidade, heterogeneidade, a simultaneidade entre produção e consumo e a ausência de transferência de propriedade (Tigre, 2019a).

Apesar das similitudes dessas características, seus padrões produtivos e inovativos, em geral relacionados à intensidade em trabalho e capital, são bastante distintos, fazendo com que o nível de complexidade tecnológica e a intensidade do uso de conhecimento e de pessoas variem significativamente nas suas distintas atividades, o que torna seu agrupamento enquanto um conjunto análogo de análise um tanto desafiador.

Nesse cenário, e sobretudo pelo seu caráter intangível e a inseparabilidade de sua produção e consumo, as mudanças ocorridas nos seus processos produtivos são muito mais dinâmicas e flexíveis, e portanto, mais rápidas, do que aquelas das empresas industriais. Essa mesma dinamicidade e flexibilidade, por sua vez, faz com que as possíveis mudanças de atividades, ora dentro do próprio âmbito dos Serviços ora, menos frequente, para fora dele, sejam mais comuns do que as que ocorrem nas empresas industriais e de comércio. Essas características tornam a classificação das empresas em CNAEs uma tarefa difícil até mesmo para humanos.

No caso das empresas de comércio, ainda que pertençam a uma única seção e apenas três divisões da estrutura de classificação, suas atividades também possuem o potencial de rápidas mudanças em um curto período, algumas vezes, inclusive, se confundindo com algumas atividades de Serviços, em particular aquelas relacionadas ao "Comércio e reparação de veículos automotores e motocicletas", atividade cujos grupos estão divididos entre as pesquisas de Comércio e Serviços.

Vale destacar ainda que embora o uso de dados de um período anterior possa funcionar bem em períodos de estabilidade econômica, o mesmo pode não se verificar em períodos de excepcionalidade, como os anos de 2020 e 2021, quando se viveu uma pandemia.

Por conta da Covid-19, nesses dois anos, muitos países implementaram medidas restritivas, de maneiras distintas, com relação à circulação de pessoas e à forma de funcionamento das empresas. Nesse contexto, foram divulgados, e são atualizados constantemente, pelo Eurostat, organização estatística da União Europeia, diversos guias metodológicos e, dentre eles, um que visa fornecer orientações sobre o impacto da crise da Covid-19 nas estatísticas econômicas anuais.

No que diz respeito às atividades desenvolvidas pelas empresas, o Eurostat (2021) destaca que algumas delas podem ter assumido temporariamente atividades alternativas, como a produção de desinfetantes ou máscaras e serviços de entrega em domicílio, não

sendo recomendado reclassificá-las para o ano de referência em causa, já que elas podem não afetar a estrutura do negócio no longo prazo. Salienta ainda que, embora a informação possa ser interessante, estatísticas experimentais *ad-hoc* podem ser preferíveis para esse fim.

Assim, fica claro que usar dados de 2019 em conjunto com informações de 2021 pode ser uma tarefa delicada, sendo difícil avaliar a sua viabilidade, pois não é trivial mensurar o quanto atividades temporárias impactariam em dados exibidos no *site* de uma empresa. Pode-se especular que, possivelmente, as maiores alterações nos *sites* seriam nas páginas dedicadas aos produtos ou serviços oferecidos pelas companhias e não em páginas que objetivam contar a sua história e explicitar "quem ela é".

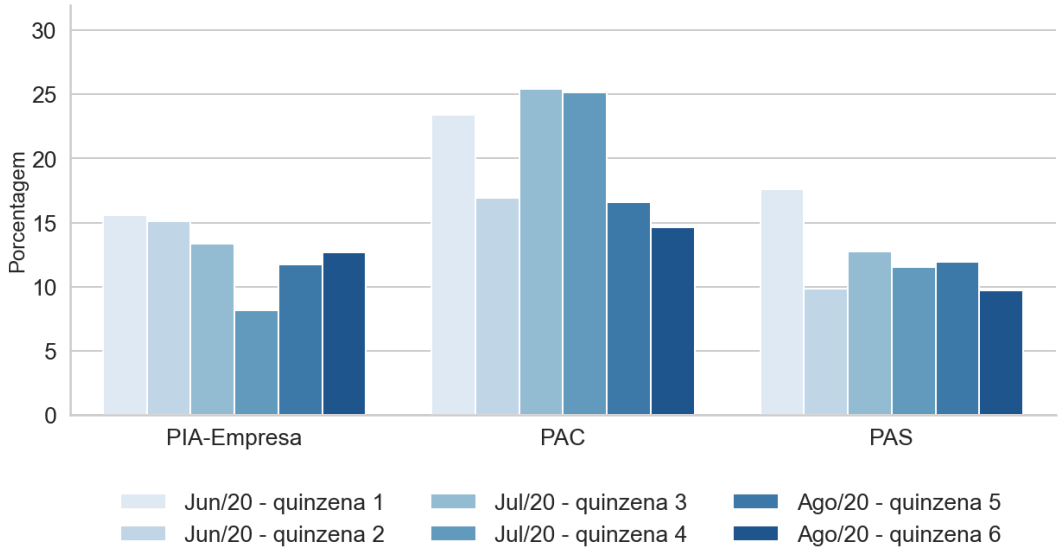
Buscando mais evidências da exequibilidade do estudo principal, ainda tratando do contexto pandêmico, é possível citar a Pesquisa Pulso Empresa (PPEmp). Ela se enquadra na categoria das estatísticas experimentais[5] divulgadas pelo IBGE e tem como objetivo estimar os impactos da Covid-19 na economia brasileira, tendo como unidade de investigação as empresas não financeiras representativas das atividades de Indústria, Construção, Comércio e Serviços.

A PPEmp (IBGE, 2020a) foi realizada quinzenalmente, com início em 15 de junho de 2020 e enceramento em 01 de outubro de 2020, por telefone, e seu âmbito seguiu o adotado pelas Pesquisas Estruturais por Empresa (PAIC, PIA-Empresa, PAC e PAS). Pode-se ainda mencionar que foi adotado um plano amostral estratificado que considerou dez estratos de atividades econômicas (Indústria, Construção, Comércio - varejista; por atacado; e de veículos, peças e motocicletas - e Serviços - prestados às famílias; de informação e comunicação; profissionais; transportes, serviços auxiliares e correios; e outros serviços), cinco estratos de recortes territoriais formados pelas Grandes Regiões e três estratos de tamanho das empresas (empresas com menos de 50 pessoas ocupadas; com 50 ou mais e menos de 500 pessoas ocupadas; e com 500 ou mais pessoas ocupadas).

Um resultado fornecido pela pesquisa que pode dar indícios a respeito das atividades econômicas, apesar de talvez ter mais ligação com o conceito de inovação, é a quantidade de empresas que lançaram ou passaram a comercializar novos produtos ou serviços. Os percentuais de empresas por quinzenas consideradas para PIA-Empresa, PAC e PAS podem ser vistos na Figura 3.3. Vale destacar que a primeira divulgação, referida como *"Jun/20 - quinzena 1"* na legenda do Gráfico, trouxe comparações entre a primeira quinzena de junho e o período anterior ao início da pandemia e as demais trouxeram comparações com a quinzena imediatamente anterior.

[5] Segundo o IBGE o uso dessas estatísticas requer cuidado pois são novas e ainda estão em fase de teste e sob avaliação.

Figura 3.3: Empresas (%) que lançaram ou passaram a comercializar novos produtos ou serviços, por pesquisa, em cada quinzena



Fonte: Compilação da autora a partir de IBGE (2020a).

Como pode ser observado, o percentual de empresas que lançaram ou passaram a comercializar novos produtos ou serviços não é alto para a PIA-Empresa e para a PAS. Entretanto, para a PAC, nas quinzenas 1, 3 e 4, ele ultrapassa os 20%, alcançando as duas últimas quinzenas mencionadas um valor de, aproximadamente, 25%. Vale frisar que o lançamento ou o início da comercialização de novos produtos ou serviços não implicam necessariamente na alteração da atividade principal da empresa, especialmente em um nível mais agregado da classificação. Além disso, como será descrito na Subseção 3.2.1, buscou-se textos de páginas da *web*, para classificar a atividade econômica principal da empresa, que descrevessem quem era a companhia, além de outros mais relacionados aos seus serviços e produtos, presumindo que o seu conteúdo seria mais estável perante a mudanças momentâneas relacionadas as atividades econômicas exercidas.

Logo, ainda que a PPemp tenha trazido alguns percentuais altos de empresas que lançaram ou passaram a comercializar novos produtos ou serviços, entende-se que esses valores refletem cenários pouco prováveis[6] quando relacionados a mudança da CNAE principal da empresa. Sendo assim, apesar dos números exibidos na Figura 3.3 alertarem quanto a necessidade de análises posteriores, sobretudo relacionadas a PAC, não trazem fortes indícios de que o estudo principal não deva ser realizado.

[6] Conjectura a autora que a ocorrência da alteração da atividade econômica principal de todas as empresas que lançaram ou iniciaram a comercialização de novos produtos ou serviços é um evento pouco provável.

3.1.2 Estudo Principal

Na execução do estudo principal, do banco de empresas selecionadas na PIA-Empresa, PAC e PAS, ano de referência 2019, inicialmente foram consideradas as empresas em operação, com questionário preenchido e que não declararam ter sofrido mudanças estruturais ou outras formas de ligação entre empresas. Além disso, foram retiradas as empresas integralmente imputadas[7], as que precisaram ser enviadas de uma pesquisa para outra após o recebimento do questionário e as que preencheram o questionário mesmo sem terem sido selecionadas.

Em um segundo momento foram excluídas as empresas que não informaram o seu *site*. Ou seja, empresas que deixaram o campo *site* do questionário em branco ou que escreveram *n/t*. Foram retiradas também as empresas que responderam com os dados do seu Instagram ou Facebook. Além disso, foram desconsideradas as empresas cujo campo *site*, do questionário, continha um dos seguintes textos: *s/d*, *nao*, *não*, *google*, *ibge*, *xxx*, *www.com.br*, *www.site.com* e *www.site.com.br* ou que continham as duas palavras *site* e *sem*. Ressalta-se ainda que foram excluídos espaços em branco presentes no meio do texto e pontos (sinal de pontuação) seguidos foram substituídos por um único ponto.

Na sequência, foram retiradas as empresas cuja CNAE determinada pelo informante não era a mesma utilizada pelo IBGE no momento de seleção da amostra e que, além disso, não tiveram a verificação executada manualmente por um técnico do Instituto.

Notou-se ainda que, por vezes, diferentes empresas, considerando o CNPJ a oito dígitos, de uma mesma pesquisa ou de pesquisas diferentes informavam o mesmo *site*. Esse fenômeno se mostrou relativamente raro quando foram buscadas empresas com diferentes CNPJs, mesmo *site* e CNAEs principais distintas. Na maior parte das vezes se tratavam de companhias grandes que utilizam um CNPJ para cada atividade econômica. Porém, também foi possível identificar casos em que o *site* informado era referente à consultoria de contabilidade contratada pela empresa para responder ao questionário e não à empresa selecionada. Nos casos em que se verificou mais de uma CNAE, todas as empresas foram retiradas, nos outros, uma das observações foi mantida.

Exclusivamente para o estrato certo da PIA-Empresa[8], também foi possível avaliar o comportamento das unidades locais. A pesquisa solicita, para esse grupo de empresas, que informem se possuem mais de uma UL produtiva[9] e, mais do que isso, qual é a CNAE de cada uma delas, podendo tais ULs possuírem diferentes classificações.

---

[7] "Imputação é um processo usado para determinar e atribuir valores de substituição para dados faltantes, inválidos ou inconsistentes.[...]"(United Nations, 2021, p.212, tradução da autora)

[8] O estrato certo da PIA-Empresa é composto por empresas com mais de 30 pessoas ocupadas e/ou que auferiram receita bruta, oriunda das vendas de produtos e serviços industriais, superior a um determinado valor no ano anterior ao de referência da pesquisa. Mais detalhes, referentes à amostra da PIA-Empresa, podem ser vistos na Seção 2.5.

[9] De acordo com IBGE (2004, p.13): "Unidade local produtiva é o local/endereço onde a atividade produtiva, principal ou secundária, é exercida."

De acordo com IBGE (2004, p.27) a determinação da classificação da unidade local produtiva e da empresa ocorre da seguinte forma:

> [...] A empresa com unidade local única recebe a classificação desta unidade.
>
> No caso de empresas com mais de uma unidade local produtiva, sua atividade principal, e portanto, sua Classe CNAE, será determinada a partir da Classe CNAE de suas unidades locais e respectivos valores das expedições (vendas + transferências) realizadas, pelo método descendente (*top down*), ou seja, obedecendo à sequência descrita a seguir: primeiro determina-se a Seção sobre a qual recai o montante maior de expedições; dentro desta Seção, determina-se a Divisão que detém o maior volume desta variável; dentro desta Divisão, determina-se o Grupo com maior valor da variável e, dentro deste Grupo, determina-se a Classe com maior participação no valor das expedições. Esta Classe identifica a atividade principal da empresa. [...]

Decidiu-se também retirar da base as empresas que relataram possuir mais de uma UL com CNAEs distintas na PIA-Empresa.

Informações sobre a quantidade de empresas presentes em cada base de dados podem ser vistas a seguir na Tabela 3.1:

Tabela 3.1: Empresas no CBS, selecionadas pelas pesquisas e utilizadas no estudo, por pesquisa

Empresas	PIA-Empresa	PAC	PAS
Cadastro de seleção (1)	390 869	1 707 406	1 721 026
Selecionadas para a pesquisa (1)	48 113	79 266	116 788
Utilizadas no estudo (2)	12 743	10 201	15 932

(1) Fonte: Compilação da autora a partir de IBGE (2021i), IBGE (2021g) e IBGE (2021h).

(2) Fonte: Autoria própria.

Assim, a base final destinada ao estudo contou com 38 876 empresas e com a variável *site*, isto é, a *URL* informada pelas companhias, e com a sua CNAE principal. A partir do texto da *URL* criou-se o primeiro conjunto de variáveis preditoras.

URL (*Uniform Resource Locator*) como Variável Preditora

As *URLs* são os endereços virtuais, no caso aqui tratado, das empresas. Elas são compostas por várias partes, dentre elas o nome do domínio e a sua categoria. Para ilustrar, tem-se na *URL https://www.ibge.gov.br/* o nome *ibge* como nome do domínio e *gov* como categoria. As categorias possuem significados, o uso de *gov* revela que se trata de uma instituição do governo federal.

No Brasil, o Núcleo de Informação e Coordenação do Ponto BR (NIC.br) tem entre suas funções o registro e manutenção dos nomes de domínios que usam o <.br>, a promoção de estudos e a recomendação de procedimentos, normas e padrões técnicos e operacionais para a segurança das redes e serviços de *internet*, visando assim a sua crescente e adequada utilização pela sociedade, além de diversas outras relacionadas ao seu papel de implementar as decisões e os projetos do Comitê Gestor da *Internet* no Brasil (CGI.br). Sendo o CGI.br o responsável por coordenar e integrar as iniciativas e serviços da *internet* no País.

O Registro.br é o departamento do NIC.br responsável pelas atividades de registro e manutenção dos nomes de domínios que usam o <.br>. Ele oferece diversas categorias de domínios (NIC.br, 2022a) aos usuários. Os domínios de pessoa física e profissionais liberais só podem ser registrados por um titular com CPF. Os domínios de pessoa jurídica devem ser associados a um CNPJ. Já os domínios genéricos e de cidades podem ser registrados por CPF ou CNPJ. Algumas categorias possuem ainda restrições adicionais por serem direcionadas a empresas de setores específicos, sendo necessária comprovação por meio de envio de documentos.

Apesar das diversas opções existentes, as estatísticas (NIC.br, 2022b) fornecidas pelo Registro.br evidenciam que a imensa maioria das *URLs* pertence à categoria *com.br*. O mesmo fenômeno pôde também ser verificado nas bases oriundas das Pesquisas Estruturais por Empresa. Nessas bases também foram observados diversos *sites* fora do domínio *.br*. Alguns deles eram referentes a domínios de outros países e outros entravam na categoria de domínios internacionais, ou seja, não possuíam uma extensão relacionada a nenhum país. Nesse último conjunto estão as *URLs* com final *.com*, *.net*, *.org*, entre outros.

Ainda que não exista nenhum padrão rígido para a formação das *URLs*, o nome do domínio e a sua categoria podem fornecer indícios da atividade econômica desempenhada pela companhia. Berardi *et al.* (2015) obtiveram resultados muito satisfatórios com o uso desses campos, ao avaliarem o desempenho alcançado por diferentes conjuntos de variáveis preditoras, na tarefa de classificação de uma empresa de acordo com sua atividade econômica. Para tanto, os autores utilizaram *sites* em língua inglesa.

Por conta disso, este trabalho utilizou a parte central das *URLs* extraídas das Pesquisas Estruturais por Empresa para formar agrupamentos de caracteres de tamanhos 3, 4, 5 e 6. Para tanto, a princípio, retirou-se a parte inicial (*www.*, *http://* e *https://*) e a parte final (qualquer informação a partir de *.com* ou *.br*) dos endereços das páginas da *web*, sendo o restante repartido se houvessem pontos finais no seu interior. A partir de cada texto extraído, com tamanho superior a 3, formou-se conjuntos de caracteres. No caso de textos de tamanho menor ou igual a 3, o próprio foi utilizado. Por exemplo, se o *site* fosse *https://www.ibge.gov.br/*, seriam utilizadas as palavras *ibge* e *gov*, como a última tem tamanho 3 nada seria executado, já no caso da primeira seriam obtidos *ibg*, *bge* e *ibge*. Logo, as variáveis preditoras geradas seriam *ibg*, *bge*, *ibge* e *gov*.

Após a montagem das bases de dados utilizando informações das Pesquisas Estruturais por Empresa foram recolhidos, através dos *sites*, e pré-processados os textos disponíveis na *web*.

3.2 Dados da *Web*

A extração dos dados e o seu pré-processamento foram realizados em *Python* utilizando a ferramenta *Notebook Jupyter*. A primeira tarefa foi executada com o auxílio das bibliotecas *JusText* (Pomikálek, 2011), *Scrapy* e *NLTK* (Steven Bird e Loper, 2009). Já a segunda utilizou os pacotes *NLTK* e *Langdetect*.

3.2.1 Extração dos Dados da *Web*

A estrutura do conteúdo dos *sites*, *HyperText Markup Language - HTML* (Linguagem de Marcação de HiperTexto), foi acessada através da biblioteca *Scrapy* e também foi ela que permitiu que outras páginas da mesma companhia fossem alcançadas.

Através do pacote foram rastreadas, além da página presente no questionário (ou o redirecionamento desta), todas as outras que pudessem possuir textos que estabelecessem quem era a empresa. Buscou-se *links* na página inicial que possuíam no seu corpus ou no texto associado a ele as palavras *sobre*, *somos*, *about*, *empresa*, *institucional* ou *historia*. Além dessas, rastreou-se outras páginas que pudessem possuir textos que estabelecessem quais produtos a empresa comercializa e/ou quais serviços presta. Nesse passo, foram recolhidos *links* na página inicial que possuíam no seu corpus ou no texto associado a ele as palavras *produto*, *serviço* e *servico*. Desta vez a extração foi limitada aos dois *links* com menor número de caracteres encontrados. Optou-se por essa restrição após observar que algumas empresas possuíam uma página para cada produto/serviço, o que gerava uma extensa quantidade de textos coletados, podendo demandar uma capacidade computacional muito grande nos outros passos do estudo.

Não foi possível obter dados de todos os *sites* mapeados. Alguns deles bloquearam a extração automática de informações. Somado a isso, ainda haviam *URLs* incorretas e páginas fora do ar.

Em conjunto com a biblioteca *Scrapy*, para realizar a extração dos textos, foi utilizada a biblioteca *JusText*. A sua principal função é remover *boilerplate*, partes supostamente não informativas, de páginas em *HTML* e extrair textos. A biblioteca foi projetada, principalmente, para identificar textos que contêm frases completas.

JusText, inicialmente, segmenta o conteúdo da página analisada com base em marcações *HTML* com o objetivo de formar blocos de texto suficientemente homogêneos em termos de conteúdo para que possam ser classificados como bons ou ruins.

Após a etapa de segmentação é realizado o pré-processamento. Nesse passo são removidos conteúdos com determinadas marcações, como <header>, e outros são marcados como ruins, como blocos que contém o símbolo de *copyright*.

Na fase seguinte, os conjuntos ainda não rotulados são classificados em uma das quatro classes: blocos *boilerplate* – ruim; blocos de conteúdo principal – bom; blocos curtos para uma decisão confiável – pequeno; e blocos em uma categoria entre os pequenos e os bons – quase bons. A atribuição das classes tem como fatores determinantes a quantidade de *tokens*[10], de *links* e de *stopwords*[11]. A ferramenta possui listas de *stopwords* em várias línguas, incluindo o português, e também permite que sejam utilizadas outras definidas pelo usuário. Este trabalho utilizou a lista de *stopwords* da biblioteca *NLTK*.

Por último, os segmentos classificados como pequenos ou quase bons são realocados nas categorias bom ou ruim. Essa classificação é feita de acordo com a classe dos blocos

[10] *Token* é uma sequência contígua de caracteres entre dois espaços, entre um espaço e sinais de pontuação ou até mesmo um sinal de pontuação sozinho.

[11] *Stopwords* são palavras não relevantes para análises de textos pois possuem grande probabilidade de estarem presentes em todos os documentos.

vizinhos e também considerando a posição do bloco na página.

Planejou-se, inicialmente, buscar apenas textos bem estruturados que retratassem uma sequência lógica e estivessem presentes nas páginas referentes a quem era a empresa, entretanto esse tipo de extração gerou um conjunto muito pequeno de observações a serem analisadas. Por conta disso, optou-se por captar textos pertencentes a qualquer classe quando se tratava da página informada pela empresa ou das páginas referentes a produtos e serviços. Já das páginas relacionadas a quem era a empresa, esperava-se encontrar textos mais bem estruturados e decidiu-se por extrair todos os segmentos não classificados como *boilerplate* (ruim).

A lógica final adotada para a extração foi baseada em relatos de técnicos do IBGE que, eventualmente, acessam páginas da *web* das empresas buscando sanar dúvidas relacionadas à classificação econômica. Além do elencado a respeito do que foi extraído, os técnicos também mencionaram procurar nas páginas quais eram os clientes da companhia. Neste trabalho, preferiu-se não buscar essa informação, pois caso os clientes fossem outras empresas poderia ser que os dados textuais relacionados a essas dificultassem a modelagem. Após a extração dos textos foi realizado o seu pré-processamento. Essa etapa englobou a limpeza e a padronização do texto bruto extraído.

3.2.2 Pré-Processamento dos Dados da *Web*

A princípio todo o texto foi reescrito em letra minúscula, foi realizada a *tokenização*[12], as *stopwords* foram retiradas e também os acentos e cedilhas. Somado a isso, foram mantidos apenas os *tokens* nos quais todos os caracteres eram parte do alfabeto e os seus tamanhos estavam contidos no intervalo de [2, 46]. O corte superior, de 46, foi definido de acordo com a quantidade de letras da maior palavra dicionarizada da língua portuguesa. Alguns *tokens* possuíam total de caracteres superior a esse por terem sido recolhidas também partes consideradas *boilerplate*, que nem sempre são textos bem estruturados. Na sequência, foram excluídas as empresas cujos textos possuíam menos de 20 *tokens*, pois notou-se que se tratavam em sua imensa maioria de *sites* ainda em desenvolvimento ou fora do ar.

Por último, através da biblioteca *Langdetect*, avaliaram-se os idiomas dos textos extraídos. Grande parte deles estava em português, já em segundo e terceiro lugares detectou-se a presença do inglês e do espanhol, respectivamente. Aproximadamente 9% das empresas da PIA-Empresa apresentaram textos em inglês e 5% da PAC e da PAS. Já em espanhol foram observados menos de 50 ocorrências em cada uma delas. A quarta linguagem mais encontrada, não superando os 25 documentos, na PIA-Empresa foi o alemão e na PAC e na PAS foi o italiano. Sendo assim, foram retiradas cerca de 10% das empresas da PIA-Empresa e 6% das empresas da PAC e da PAS para as quais foi detectada uma língua diferente do português.

Após essa fase passou a existir dois conjuntos de variáveis preditoras, as obtidas das páginas da *web* após as etapas de extração e pré-processamento e as oriundas da *URL*, descritas na Subseção 3.1.2. Esses dados deram origem à base final.

---

[12] *Tokenizar* significa dividir strings em substrings (*tokens*) de acordo com determinadas regras. Este trabalho utilizou a função *word_tokenize* do *NLTK* para realizar a *tokenização*.

3.3 Base Final

A base final foi constituída pelas empresas para as quais foi possível coletar textos da *web* em português e que continuaram possuindo um número mínimo de *tokens* após o pré-processamento descrito na Subseção 3.2.2. Nessa base, os *tokens* obtidos através do texto da *URL* de cada empresa foram adicionados aos extraídos das páginas da *web*. Optou-se por essa abordagem devido ao fato do conjunto de palavras oriundas da *web* já não estar organizado de acordo com uma sequência lógica por conta dos critérios adotados para a sua coleta. Após essa etapa, uma última limpeza foi realizada.

A limpeza teve como objetivo encontrar empresas cujo *site* informado era referente ao da consultoria de contabilidade contratada pela companhia para responder ao questionário e não ao da mesma[13]. Para tanto, buscou-se nos textos os *tokens contabil*, *contab* e *contabilidade*. Foram removidas as observações que apresentaram cinco vezes ou mais os termos e não pertenciam a Divisão 69 (Atividades jurídicas, de contabilidade e de auditoria) da CNAE, totalizando 227 empresas. Compreende-se que nesta etapa os dados de teste foram bisbilhotados (ocorreu *data snooping*), pois as bases ainda não haviam sido separadas quando as empresas foram retiradas e foi necessário identificar a Divisão 69. Todavia, entendeu-se que a existência da inconsistência no preenchimento do questionário é um problema já identificado a ser sanado durante a etapa de coleta dos mesmos, conscientizando o informante ou, ainda, implementando outras formas de validação ou até mesmo de obtenção das *URLs*[14]. Além disso, seria possível estimar o erro gerado na base de teste caso a exclusão não tivesse sido efetuada[15].

Vale ainda ressaltar que de forma alguma as empresas que compõem a base final fazem parte de uma amostra aleatória ou retratam um universo. Elas foram usadas unicamente para a execução de um estudo de caso, tendo sido evitadas complexidades e inconsistências detectadas nas bases iniciais oriundas das Pesquisas Estruturais por Empresa[16]. As variáveis *estrato* ao qual a empresa pertence e *região* onde se localiza a sua sede não foram utilizadas na etapa de modelagem, elas são apresentadas a seguir apenas para fornecer informações sobre quais empresas estão representadas no trabalho.

---

[13] Essas companhias não apareceram quando a base das Pesquisas Estruturais por Empresa foi avaliada, conforme descrito na Subseção 3.1.2, pois esses não eram casos em que os *sites* se repetiam para CNAEs distintas.

[14] Informações mais detalhadas sobre o tema foram fornecidas na Seção 6.1, onde se menciona os trabalhos desenvolvidos pelo ESSnet Big Data II.

[15] Cerca de 20% das 227 empresas, ou seja, aproximadamente 45 companhias se uniriam as 6 elencadas na Seção 5.2.

[16] O total de empresas presentes nas amostras de cada pesquisa e o total de empresas para as quais buscou-se coletar os textos são exibidos na Tabela 3.1. Já o total de empresas para as quais o texto das páginas da *web* foi obtido é apresentado na Tabela 3.2.

O total de empresas, de acordo com o seu estrato e por pesquisa, após a execução de todos os passos descritos neste capítulo pode ser visto na Tabela 3.2.

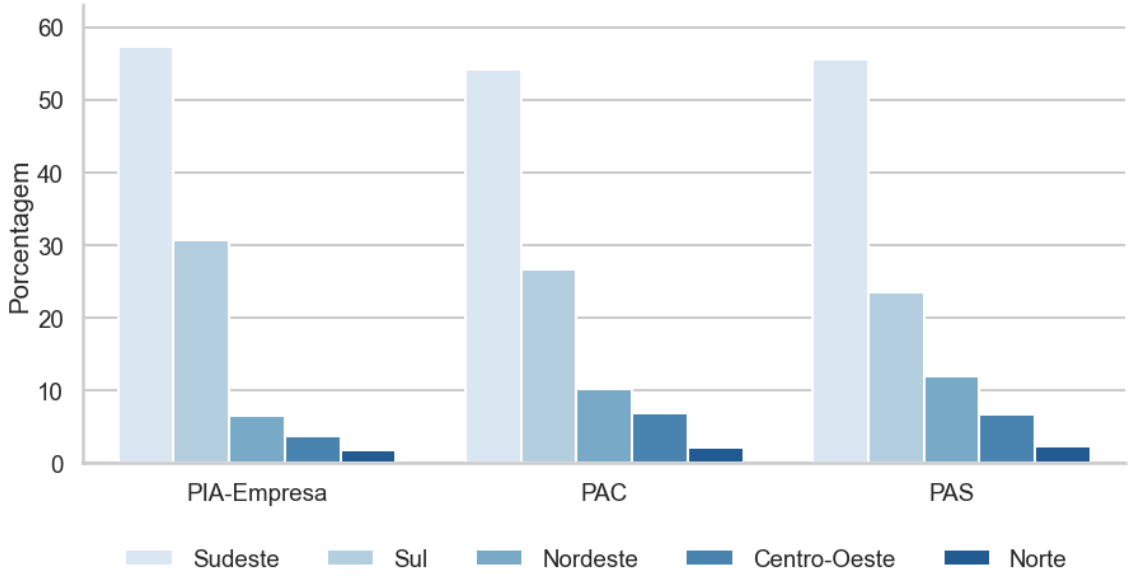
Tabela 3.2: Empresas no âmbito do estudo, após a extração e o pré-processamento dos dados, por pesquisa, segundo o estrato de seleção

Empresas	PIA-Empresa	PAC	PAS
Estrato certo ou gerencial	7 815 (86, 99%)	6 192 (93, 97%)	10 082 (89, 66%)
Estrato amostrado	1 169 (13, 01%)	397 (6, 03%)	1 163 (10, 34%)
Total	8 984	6 589	11 245

Fonte: Autoria própria.

Já o percentual das empresas existentes na base final, por pesquisa e segundo as Grandes Regiões do Brasil, é exibido na Figura 3.4.

Figura 3.4: Empresas (%) após a extração e o pré-processamento dos dados, por pesquisa, segundo as Grandes Regiões



Fonte: Autoria própria.

Na sequência, a Tabela 3.3 apresenta a quantidade de empresas por classificação econômica a 2 dígitos na PIA-Empresa, PAC e PAS e por classificação econômica a 3 dígitos na PAC. A única CNAE a 2 dígitos, pertencente a uma das três pesquisas, sem nenhuma empresa coletada foi a 92 (Atividades de exploração de jogos de azar e apostas), na PAS. A escolha do detalhamento da CNAE a ser trabalhado em cada pesquisa foi pautada na base que se conseguiu construir durante o processo de coleta e pré-processamento, ou seja, na quantidade de observações obtidas.

Tabela 3.3: Empresas após a extração e o pré-processamento dos dados na PIA-Empresa, PAC e PAS, segundo a Seção e a Divisão da CNAE, e na PAC, segundo o Grupo da CNAE

PIA-Empresa		
Seção	Divisão	Empresas
B	05 (1)	4
	06 (1)	11
	07 (1)	23
	08	126
	09 (1)	18
C	10	1 077
	11	153
	12 (1)	11
	13	340
	14	372
	15	194
	16	143
	17	257
	18	189
	19	108
	20	653
	21	142
	22	689
	23	411
	24	225
	25	792
	26	302
	27	391
	28	888
	29	332
	30	108
	31	444
	32	380
	33	201

(a) Empresas da PIA-Empresa, segundo a Seção e a Divisão da CNAE

PAC			
Seção	Divisão	Grupo	Empresas
G	**45**	45.1	692
		45.3	420
		45.4	159
	46	46.1	44
		46.2	153
		46.3	356
		46.4	656
		46.5	103
		46.6	420
		46.7	273
		46.8	537
		46.9	128
	47	47.1	286
		47.2	155
		47.3	194
		47.4	643
		47.5	415
		47.6	122
		47.7	335
		47.8	498

(b) Empresas da PAC, segundo a Seção, a Divisão e o Grupo da CNAE

Tabela 3.3: Empresas após a extração e o pré-processamento dos dados na PIA-Empresa, PAC e PAS, segundo a Seção e a Divisão da CNAE, e na PAC, segundo o Grupo da CNAE

PAS		
Seção	Divisão	Empresas
A	01	44
	02 (1)	18
E	37 (1)	34
	38	125
	39 (1)	13
G	45	173
H	49	1 301
	50	65
	51	39
	52	451
	53	80
I	55	687
	56	815
J	58	153
	59	137
	60	157
	61	295
	62	775
	63	186
K	66	341
L	68	335
M	69	1 407
	70	132
	71	451
	73	195
	74	136
N	77	384
	78	178
	79	146
	80	225
	81	434
	82	473
P	85	285
R	90	37
	92 (1)	0
	93	175
S	95	188
	96	175

(c) Empresas da PAS, segundo a Seção e a Divisão da CNAE

Fonte: Autoria própria.
(1) Classes excluídas.

Como mostrou a Tabela 3.3, alguns níveis da CNAE passaram a contar com poucas observações. Por causa disso, foram excluídas da base final as empresas pertencentes a PIA-Empresa com CNAE a 2 dígitos 05 (Extração de carvão mineral), 06 (Extração de petróleo e gás natural), 07 (Extração de minerais metálicos), 09 (Atividades de apoio à extração de minerais) e 12 (Fabricação de produtos do fumo) e as da PAS com CNAE a 2 dígitos 02 (Produção Florestal), 37 (Esgoto e atividades relacionadas), 39 (Descontaminação e outros serviços de gestão de resíduos) e 92 (Atividades de exploração de jogos de azar e apostas), em outras palavras, as classes que possuíam menos de 35 observações. Assim, a base final compreendeu 26 686 empresas.

A seguir, a base final foi dividida em base de treino e de teste para dar início à etapa de modelagem.

3.4 Divisão da Base Final - Treino e Teste

Do conjunto de empresas pertencentes à base final, 20% foram alocadas na base de teste e o restante na base de treino. Assim, a base de teste contou com 5 338 empresas e a base de treino com 21 348. A divisão considerou uma estratificação da variável *pesquisa*[17] e da CNAE a 2 dígitos para a PIA-Empresa e para a PAS e da variável *pesquisa* e da CNAE a 3 dígitos para a PAC, totalizando 78 categorias.

[17] A variável pesquisa é composta por três categorias: PIA-Empresa, PAC e PAS.

Capítulo 4

Variáveis Preditoras Finais e Modelagem

A determinação das variáveis preditoras finais e de como a modelagem é desenvolvida está intimamente relacionada com a base de dados utilizada, as características do problema tratado e a capacidade computacional disponível.

Este capítulo se inicia com a descrição dos *tokens* coletados no Capítulo 3, seguindo com o tratamento da alta dimensão dos dados. Neste trabalho optou-se por uma redução de dimensionalidade de acordo com a cobertura de cada *token*, bem como por testar o desempenho dos modelos também após o uso de um *stemmer*[1].

Na continuação são exibidas representações do tipo *Bag-of-Words* (saco de palavras). Escolha essa baseada no fato da tentativa de coleta apenas de páginas da *web* não classificadas como *boilerplate*, pelo *JusText*, ter gerado um número muito pequeno de observações por classe final do problema, sendo então preferida a coleta de segmentos pertencentes a todas as categorias e seguindo a lógica utilizada pelos técnicos do IBGE. Em geral, é melhor treinar um modelo simples com mais dados do que um modelo mais complexo com menos dados. Ademais, como as classes finais do problema apresentaram quantidades muito diversas de empresas, discorre-se sobre o seu desbalanceamento.

Na sequência, devido à hierarquia de classes da Classificação Nacional de Atividades Econômicas e a sua utilização nas Pesquisas Estruturais por Empresa, é possível testar classificadores locais e planos ao se considerar o aprendizado supervisionado. Assim, para o primeiro são implementados os algoritmos Naive Bayes, Regressão Logística e os Métodos de Suporte Vetorial. Já para o segundo acrescenta-se a essa lista o Boosting.

A seguir, para medir o desempenho e selecionar os hiperparâmetros adequados para o método de aprendizagem utilizado, é apresentada a validação cruzada, sendo a busca aleatória o método escolhido para a obtenção dos hiperparâmetros a serem testados. Por fim, são descritas as métricas para verificação de desempenho.

Nesse contexto de modelagem, vale ainda mencionar que os dados fornecidos pela *web*

[1] Ver Subseção 4.1.2.

aqui operados estão indiretamente relacionados com o fenômeno de interesse, a CNAE principal da empresa, o que torna a tarefa de predição mais desafiadora. Esse tipo de aplicação é identificada como "uso derivado" por Eurostat (2020). Eles descrevem o uso como um processo em duas etapas, onde a primeira consiste na geração dos dados e a segunda no uso do modelo para obter o resultado derivado. Além disso, chamam atenção para o fato de que qualquer coisa que afete a relação entre os dados gerados e o fenômeno (derivado) estudado influenciará na qualidade das estatísticas com base em tal fonte. Por conta disso, enfatizam que se faz necessária a verificação do modelo em uma base regular. Assim, caso se observe uma mudança, o modelo deve ser ajustado para manter a detecção de padrões úteis. Essa observação está pautada na ideia de mudança de conceito, mencionada na Seção 2.2.

4.1 Descrição dos *Tokens* e Redução de Dimensionalidade da Base de Treino

Trabalhar com dados textuais, em geral, significa tratar de dados que têm alta dimensão, ou seja, aqueles para os quais o número de variáveis é muito superior à quantidade de unidades amostrais. Assim, objetivando a análise da quantidade e do perfil dos *tokens* que estavam presentes na base de treino para, posteriormente, realizar a sua redução, foi calculada a frequência absoluta para cada um deles e a sua cobertura. O termo cobertura se refere ao percentual de documentos em que um determinado *token* apareceu.

Para tanto, os *tokens* obtidos através do texto da *URL* de cada empresa foram adicionados aos retirados das páginas da *web*, como descrito na Seção 3.3. Nesse conjunto foi possível verificar a existência de 414 237 *tokens* únicos. A Tabela 4.1 mostra os cinco com maior cobertura e os cinco com maior frequência. Vale destacar que os *tokens* com menor cobertura e contagem foram omitidos pois alguns se tratavam dos nomes das empresas.

Tabela 4.1: Os cinco *tokens* com maior cobertura e os cinco com maior frequência absoluta presentes na base de treino

Cobertura	Frequência absoluta
(*contato*, 81,56%)	(*produtos*, 123 687)
(*empresa*, 72,72%)	(*empresa*, 104 736)
(*todos*, 72,57%)	(*servicos*, 91 902)
(*qualidade*, 70,20%)	(*contato*, 76 911)
(*clientes*, 68,57%)	(*qualidade*, 67 617)

Fonte: Autoria própria.

De acordo com a Tabela 4.1, a palavra *contato* estava presente em 81,56% dos documentos da base de treino, isto é, no texto coletado de 81,56% das empresas. Já a palavra *produtos* apareceu 123 687 vezes considerando todas as ocorrências em todos os documentos. A mesma lógica se aplica aos demais *tokens*. A partir dessas informações foi possível reduzir a dimensionalidade da base de dados.

4.1.1 Redução de Dimensionalidade a partir da Cobertura dos *Tokens*

Visando diminuir o total de *tokens* únicos presentes no conjunto de documentos, fez-se uso da cobertura calculada. Essa decisão foi tomada com base no fato de que tanto os recursos muito frequentes como os pouco frequentes agregarem menos informação a um modelo, pois estão presentes em quase todos os documentos ou em apenas alguns, não servindo para identificar padrões.

Optou-se por retirar todos os *tokens* que estivessem aparecido em 20 documentos ou menos, ou seja, com cobertura inferior a 0,093%. A maioria deles se tratavam de nomes próprios, palavras que perderam o sentido (por exemplo, duas palavras sem espaço entre elas) durante a coleta ou, ainda, *tokens* gerados a partir do texto da *URL*. Nenhum corte superior foi definido para a exclusão. Após a remoção restaram 28 335 *tokens* únicos, 6,84% do número inicial. Essa decisão aparentou ser razoável devido ao fato de a menor classe na base de treino possuir 30 observações, da inviabilidade computacional de se trabalhar com a totalidade original de *tokens* e da possível inclusão de ruídos que não gerariam um melhor desempenho.

Assim, após a redução, o número mínimo de *tokens* para uma empresa foi de 18, o primeiro quartil de 243, a mediana de 432 e o terceiro quartil de 794. Constaram ainda 9 documentos com mais de 30 000 *tokens*. O valor máximo para uma empresa foi de 128 222. Contudo, devido ao fato de ainda terem restado mais variáveis preditoras, 28 335 *tokens* únicos, do que amostras também foi testada uma normalização através de *stemming*.

4.1.2 Redução de Dimensionalidade através de Técnicas de *Stemming*

A abordagem consiste na aplicação de uma série de regras para se obter uma *substring* a partir da *string* original. O objetivo é remover afixos, isto é, trabalhar com a raiz das palavras e, assim, reduzir a quantidade de *tokens* (Bengfort *et al.*, 2018).

A execução foi realizada através do *stemmer* *RSLPStemmer* do pacote *NLTK*. Vale destacar que essa não é uma tarefa fácil de ser realizada na língua portuguesa e o *stemmer* considerado, segundo a sua documentação, foi desenvolvido com finalidades unicamente de demonstração e didáticas.

A Tabela 4.2 mostra os cinco *tokens* com maior cobertura e os cinco com maior frequência considerando o *stemming* das palavras.

Tabela 4.2: Os cinco *tokens* com maior cobertura e os cinco com maior frequência absoluta presentes na base de treino após o uso do *stemmer*

Cobertura	Frequência absoluta
(*tod*, 85,74%)	(*produt*, 168 215)
(*contat*, 83,84%)	(*empr*, 137 081)
(*empr*, 76,48%)	(*serv*, 127 541)
(*client*, 74,75%)	(*tod*, 116 742)
(*atend*, 74,12%)	(*client*, 95 549)

Fonte: Autoria própria.

A maioria dos *tokens* são as raízes dos *tokens* apresentados na Tabela 4.1, tendo as suas quantidades aumentadas substancialmente. Após o uso do *stemmer* a quantidade de *tokens* únicos passou de 28 335 para 13 361.

Os *tokens*, sequências de símbolos, não podem alimentar diretamente os algoritmos pois a maioria desses espera vetores numéricos de tamanho fixo. Para contornar a questão, utilizou-se representações do tipo *Bag-of-Words* (saco de palavras).

4.2 Representações do tipo *Bag-of-Words* (Saco de Palavras)

Devido à lógica adotada na extração dos textos da *web*, descrita na Subseção 3.2.1, e à obtenção de *tokens* a partir dos textos das próprias *URLs*, formados por conjuntos de caracteres, conforme detalhado na Subseção 3.1.2, foram utilizadas representações do tipo *Bag-of-Words (BoW)* e *unigrams*[2]. Nessa abordagem cada *token* é tratado individualmente e o texto de cada documento, ou seja, o texto referente a cada empresa, é representado como um saco de seus *tokens*, desconsiderando a sua ordem no texto e aspectos gramaticais, mas mantendo informações sobre as suas quantidades. Assim, cada documento é representado por um vetor de seus *tokens* e o corpus é retratado por uma matriz onde cada linha é referente a um documento e cada coluna a um *token*. Vale ainda observar que quanto maior a quantidade de variáveis preditoras, ou seja, quanto maior a dimensionalidade, mais altas as chances destas matrizes serem esparsas.

Através do pacote *Scikit-Learn* (Pedregosa *et al.*, 2011) foram implementadas duas abordagens diferentes para o preenchimento da matriz, uma utilizando a frequência do termo e a outra a frequência do termo levando em consideração a frequência inversa do documento.

[2] Trabalhar com *unigrams*, n-grams onde n=1, significa tratar cada *token* individualmente. Já quando n=2, tem-se os *bigrams* e cada dupla de *tokens* vizinhos são considerados como um único (como uma espécie de palavra composta). A mesma lógica se aplica quando n=3, onde tem-se os *trigrams*, e para outros valores de n. Neste trabalho chegou-se a testar o uso de *unigrams* e *bigrams* em conjunto quando do ajuste da Regressão Logística nos dados de treino, entretanto essa abordagem não forneceu melhores resultados e aumentou consideravelmente o tempo de execução.

Term Frequency - TF

Nesse tipo de tratamento a matriz documento por *token* é preenchida com a frequência com que o *token* está presente no texto de determinado documento, em outras palavras, com a quantidade de vezes em que o *token* aparece no texto coletado do *site* da empresa. Nesse caso específico, para normalização, as quantidades são divididas pelo total de *tokens* obtidos para a mesma empresa, logo, utiliza-se a frequência relativa.[3]

Term Frequency - Inverse Document Frequency - TF-IDF

Essa abordagem, diferente da anterior, leva em consideração, além do documento a ser descrito, os demais da coleção. O principal objetivo do uso do TF-IDF é considerar a frequência relativa de *tokens* em um documento em relação à sua frequência no restante do conjunto. Bengfort *et al.* (2018) usam como exemplo um corpus sobre esportes. De acordo com os autores, *tokens* como "*umpire*"[4], "*base*"[5] e "*dugout*"[6] são mais comuns em textos que discutem sobre beisebol e podem ser úteis para caracterizá-los, enquanto outros *tokens*, como "*run*"(correr), "*score*"(pontuar) e "*play*"(jogar), que aparecem frequentemente ao longo de todo o corpus são menos importantes.

Ainda segundo Bengfort *et al.* (2018), essa forma de codificação acentua termos que são relevantes para uma instância específica. Assim, o TF-IDF é calculado por termo e por documento, de modo que a importância de um *token* para um documento é medida pela frequência do aparecimento do termo no documento ponderada pelo inverso da frequência do termo no corpus. O valor obtido através desse cálculo irá compor a matriz de documentos por *tokens*.[7]

Outro ponto a ser levado em consideração é o desbalanceamento dos dados, isto é, o fato de algumas classes possuírem mais dados do que outras, como pode ser verificado na Tabela 3.3.

4.3 Dados Desbalanceados

O desequilíbrio na quantidade de amostras em cada classe gera consequências no processo de aprendizagem que, geralmente, produz classificadores que tendem a rotular as novas amostras como pertencentes à classe majoritária.

Esse problema pode ser contornado através de métodos de amostragem. A abordagem adotada pode consistir na exclusão de observações da classe majoritária (*undersampling*), na replicação (ou geração sintética) de casos da classe minoritária (*oversampling*) ou na combinação dos dois.

---

[3] Para a execução foram utilizadas em sequência as funções *CountVectorizer* e *TfidfTransformer*, contando a última com o parâmetro *use_idf=False*.

[4] É o árbitro do beisebol.

[5] A base é um objeto fixado no campo, um dos elementos primordiais do beisebol.

[6] É uma espécie de banco de reservas.

[7] Para a execução foram utilizadas em sequência as funções *CountVectorizer* e *TfidfTransformer*, contando a última com o parâmetro *use_idf=True*.

Optou-se no estudo por não utilizar métodos de *undersampling*, pois não havia muitas observações disponíveis. Já métodos combinados chegaram a ser testados para os algoritmos menos custosos computacionalmente. Nesses casos observou-se um aumento do tempo de execução sem melhora do desempenho. Por conta disso, além da aplicação do modelo na base desbalanceada[8], foi testado um método de *oversampling* conhecido como *SMOTE* (*Synthetic Minority Over-sampling Technique*).

SMOTE é uma abordagem na qual a classe minoritária é aumentada através da geração de exemplos sintéticos. Para criar um dado, a diferença entre um vetor da amostra e o seu vizinho mais próximo é multiplicada por um número aleatório entre 0 e 1 e o valor resultante é adicionado ao vetor da amostra considerado inicialmente. Essa abordagem força a região de decisão da classe minoritária a se tornar mais ampla (Chawla *et al.*, 2002).[9]

Mais uma questão a ser analisada, devido às características do problema tratado, se refere ao fato da CNAE possuir uma estrutura hierárquica, o que possibilita o uso de métodos de classificação que levem esse aspecto em consideração.

4.4 Classificação Hierárquica

De acordo com Silla e Freitas (2011, p.33), "[...]classificação hierárquica pode ser vista como um tipo particular de problema de classificação estruturada, em que a saída do algoritmo de classificação é definida sobre uma taxonomia de classe[...]"(tradução da autora)[10]. Assim, a classificação hierárquica pode ser entendida como um subconjunto da classificação estruturada, sendo essa última mais ampla e abrangendo problemas de classificação onde existe alguma estrutura, podendo ser essa hierárquica ou não, entre as classes.

Ainda segundo Silla e Freitas (2011), a taxonomia, normalmente, é organizada em uma árvore ou em um *direct acyclic graph - DAG* (grafo direcionado acíclico) onde cada nó corresponde a uma classe e as arestas representam a relação entre elas. A grande diferença entre eles é que, no último, um nó pode ter mais de um nó pai, em outras palavras, um nó pode ter como origem dois ou mais nós.

Os dados aqui trabalhados podem ser organizados em uma árvore, apenas sendo capazes de gerar algum tipo de confusão quando se trata da Divisão 45 (Comércio e reparação de veículos automotores e motocicletas) da CNAE, pois ela está presente na PAC e na PAS. Entretanto, apesar deste trabalho considerar, após a coleta e por causa dela, a árvore podada (um máximo de 3 dígitos na PAC e 2 dígitos na PAS), a CNAE a 4 dígitos está disponível e, a esse nível, não existem intersecções entre a pesquisa de Serviços e a de Comércio. A Figura 4.1 ilustra o exposto.

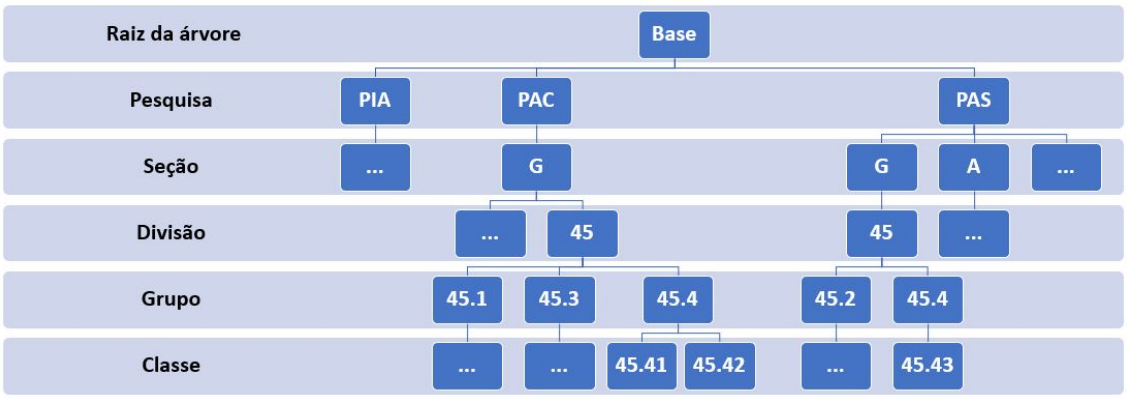
---

[8] Com e sem pesos inversamente proporcionais às frequências de classe ajustados, como pode ser visto no Apêndice A, através do parâmetro *class_weight*.

[9] A *SMOTE* foi implementada através do pacote *Imblearn* (Lemaître *et al.*, 2017).

[10] *"[...]hierarchical classification can be seen as a particular type of structured classification problem, where the output of the classification algorithm is defined over a class taxonomy[...]"*

Figura 4.1: Exemplo de parte da estrutura de árvore, oriunda da organização da CNAE, englobando a Seção G



Fonte: Autoria própria.
Nota: Os DVs das Classes CNAE foram omitidos.

Dito isso, pode-se considerar uma Classe G_{pac} com um nó 45_{pac}, que por sua vez tem um nó filho 45.4_{pac} de onde se originam os nós 45.41-2 (Comércio por atacado e a varejo de motocicletas, peças e acessórios) e 45.42-1 (Representantes comerciais e agentes do comércio de motocicletas, peças e acessórios). A mesma lógica pode ser aplicada a parte da Seção G existente na PAS, ficando caracterizada uma estrutura de árvore. Assim, entende-se que quando o modelo classifica na Divisão 45_{pas}, ele está caracterizando, na verdade, um conjunto que abarca a CNAE a 4 dígitos 45.43-9 (Manutenção e reparação de motocicletas) e a CNAE a 3 dígitos 45.2 (Manutenção e reparação de veículos automotores). Já na PAC, no caso da CNAE 45.4_{pac}, ele está indo além da CNAE a 3 dígitos, mas não alcançando o detalhamento da CNAE a 4 dígitos, pois o conjunto engloba apenas as Classes 45.41-2 e 45.42-1.[11]

Em conformidade com Silla e Freitas (2011), a taxonomia de classes pode ser definida como uma árvore estruturada sobre um conjunto parcialmente ordenado (C, $\prec$), onde $C = \{c_1, c_2, \ldots, c_M\}$ é um conjunto finito composto por todas as M classes presentes na taxonomia e $\prec$ representa a relação "*é uma*"(*is-a*) assimétrica, antirreflexiva e transitiva. Sendo essas propriedades definidas da seguinte forma:

- Existe apenas um nó, conhecido como raiz da árvore, de onde descendem todas a outras classes;
- $\forall c_i, c_j \in C$, se $c_i \prec c_j$ então $c_j \nprec c_i$;
- $\forall c_i \in C, c_i \nprec c_i$; e
- $\forall c_i, c_j, c_k \in C, c_i \prec c_j$ e $c_j \prec c_k$ implica $c_i \prec c_k$.

A estrutura hierárquica pode ser explorada de diversas maneiras na construção dos classificadores. Neste trabalho foi testado o uso de classificadores *flat* (planos) e também de classificadores locais (ou *top-down*).

[11] Os códigos e denominações das CNAEs utilizadas neste trabalho podem ser vistos no Anexo A.

Classificadores Planos

A classificação plana, segundo Silla e Freitas (2011), é a forma mais simples de lidar com problemas de classificação hierárquica e consiste em ignorar completamente a hierarquia de classes, prevendo apenas os nós finais. Ela se comporta como um algoritmo tradicional durante o treinamento e o teste.

Essa abordagem tem a desvantagem de se ter que construir um classificador para discriminar entre um extenso número de classes, sem explorar informações sobre relacionamentos de classe pai-filho.

Ao considerar esse tipo de classificador foram testados distintos algoritmos de aprendizado supervisionado, diferentes representações e bases de dados (com e sem a aplicação do *stemmer* e da *SMOTE*).

Classificadores Locais

A abordagem local transforma o problema de classificação hierárquica em um conjunto de problemas mais simples, para os quais já existem soluções testadas e validadas. Nessa abordagem um ou mais classificadores independentes são construídos para cada nível da hierarquia de classes, sendo criado algo similar a um conjunto de classificadores planos.

A ideia central desse método é que a hierarquia seja levada em consideração na perspectiva da informação local. Nesse tipo de abordagem, também conhecida como "*de cima para baixo*"(*top-down*), para cada nova observação do conjunto de teste, o sistema primeiro prevê a sua classe de primeiro nível (neste caso a pesquisa onde a empresa será alocada), depois usa essa classe prevista para limitar as opções de classes a serem previstas no segundo nível (sendo as únicas válidas as filhas das classes previstas no primeiro nível) e assim por diante, até a classe mais específica. Uma desvantagem desse tipo de abordagem é que, se nenhum procedimento for aplicado, um erro em um determinado nó superior é propagado para os seguintes, ou seja, para os nós mais específicos (Silla e Freitas, 2011).

Existem diferentes formas desse método ser implementado[12], neste trabalho utilizou-se um classificador local por nó pai (*local classifier per parent node*). Nesse caso, para cada nó pai na hierarquia de classes um classificador multiclasse[13] é treinado para distinguir entre seus nós filhos e, para evitar o problema de fazer previsões inconsistentes e respeitar as restrições naturais de pertencer a uma classe, os classificadores após o primeiro nível são treinados apenas com os filhos do seu nó pai.

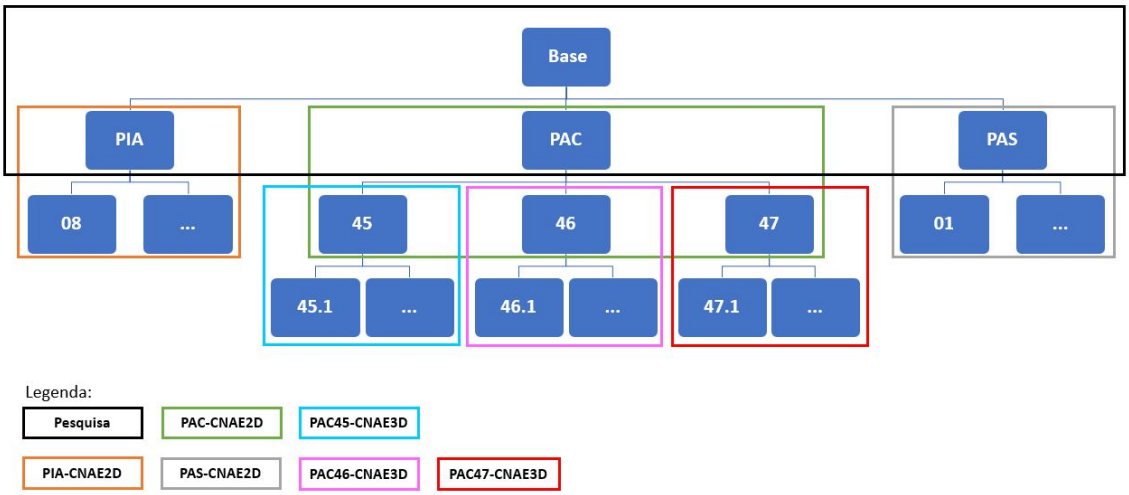
Vale ainda mencionar que optou-se por não considerar os nós referentes à Seção da CNAE. Tal decisão foi motivada pelo fato de, ao incluir esses nós, o número de modelos a serem treinados aumentar consideravelmente e, além disso, a modelagem dentro de cada Seção passar a contar, por vezes, com uma quantidade pequena de observações. Ao

[12] A descrição destas e mais detalhes podem ser vistos em Silla e Freitas (2011).

[13] Ou um que considere uma abordagem de classificação *uma contra todas* (*one-vs-rest*), ou seja, cada vez comparando uma classe com as restantes.

desconsiderar a Seção, 7 classificadores locais foram criados. Eles são exibidos na Figura 4.2.

Figura 4.2: Ilustração dos classificadores locais criados



Fonte: Autoria própria.

Na Figura 4.2, *Pesquisa* representa a modelagem executada no primeiro nó, isto é, a tentativa de classificar a base em cada uma das três pesquisas, PAC, PAS e PIA-Empresa. Já *PIA-CNAE2D* consiste na tarefa de classificar a parte da base pertencente a PIA-Empresa em uma das 24 classes representadas pela CNAE a 2 dígitos, detalhadas na Tabela 3.3. O mesmo explicitado para a *PIA-CNAE2D* vale para *PAS-CNAE2D* e *PAC-CNAE2D*. Além disso, *PAC45-CNAE3D* se refere ao exercício de tentar classificar a parte da base que pertence a PAC e possui CNAE a 2 dígitos 45_{pac}, em cada uma de suas classes a CNAE 3 dígitos, ou seja, classificar em 45.1, 45.3 ou 45.4_{pac}. Essa lógica também se aplica à *PAC46-CNAE3D* e à *PAC47-CNAE3D*.

Conceição da Silva Ferreira (2012) obteve bons resultados ao utilizar essa abordagem para classificar atividades econômicas de empresas através de textos, em língua portuguesa "de portugal", coletados da *web*. Vale ainda citar que em seu estudo os textos foram extraídos manualmente, o que limitou a quantidade de observações disponíveis. Assim, foram ajustados apenas três classificadores. O primeiro classificava as observações nas Seções G (Comércio; reparação de veículos automotores e motocicletas) e F (Construção) e em um terceiro grupo, denominado "Outras". O segundo classificava as observações da Seção G nas Divisões 45 (Comércio e reparação de veículos automotores e motocicletas), 46 (Comércio por atacado, exceto veículos automotores e motocicletas) e 47 (Comércio varejista). O terceiro classificava as observações da Seção F nas Divisões 41 (Construção de edifícios), 42 (Obras de infra-estrutura) e 43 (Serviços especializados para construção)[14]. Para os dois primeiros nós, Naive Bayes, a ser apresentado na Subseção 4.5.1, sem o uso de *stemming* exibiu o melhor resultado.

Voltando à Figura 4.2, destaca-se ainda que por nó pai foram testados diferentes algoritmos e também diferentes representações e bases de dados (com e sem a aplicação do

[14] A Seção F e suas Divisões 41, 42 e 43 não foram tratadas aqui, pois se encontram no âmbito da PAIC.

stemmer e da *SMOTE*). Os algoritmos considerados, tanto para o ajuste de classificadores planos quanto de locais, são apresentados a seguir.

4.5 Algoritmos de Aprendizado Supervisionado

Conforme mencionado na Seção 2.1, no aprendizado automático supervisionado o objetivo é prever uma variável de saída associada aos itens de entrada. Para tanto, o treinamento é realizado com dados rotulados. Assim, a ideia central é obter uma função adequada para a realização da tarefa.

Atualmente, existem muitos algoritmos, que consideram abordagens distintas, aptos a realizar a busca por funções válidas quando se visa o aprendizado supervisionado e a classificação. Este trabalho testou alguns deles[15].

4.5.1 Naive Bayes - NB

Naive Bayes é um algoritmo que geralmente se comporta bem com dados textuais. Ele é rápido de treinar e também ágil ao fazer previsões, portanto, válido para dados que têm alta dimensão. Além disso, também lida bem com matrizes esparsas. No geral, essas são características de classificadores lineares, ou seja, classificadores que usam uma combinação linear das variáveis preditoras para tomar uma decisão de classificação. Roelands (2017) obteve o melhor desempenho ao utilizar Naive Bayes para classificar atividades econômicas de empresas através de textos, em língua holandesa, coletados da *web*.

De acordo com Jurafsky e Martin (2021), ele é um classificador probabilístico, o que significa que para um documento d, de todas as classes $c \in C$, ele retorna a classe $\hat{c}$ cuja probabilidade a posteriori dado o documento é máxima. Ou seja:

$$\hat{c} = \arg\max_{c \in C} P(c|d) \tag{4.1}$$

O classificador tem como base o Teorema de Bayes. Esse teorema permite que a probabilidade condicional $P(u|v)$ seja decomposta da seguinte forma:

$$\mathrm{P}(u|v) = \frac{\mathrm{P}(v|u)\mathrm{P}(u)}{\mathrm{P}(v)} \tag{4.2}$$

Logo, a Equação 4.1 pode ser reescrita como:

$$\hat{c} = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} \frac{\mathrm{P}(d|c)\mathrm{P}(c)}{\mathrm{P}(d)} \tag{4.3}$$

[15] As funções e pacotes utilizados para o ajuste de cada algoritmo podem ser vistos no Apêndice A.

Podendo $P(d)$ ser omitido na Equação 4.3 já que o cálculo de $\frac{\mathrm{P}(d|c)\mathrm{P}(c)}{\mathrm{P}(d)}$ é realizado para cada classe possível, ou seja, questiona-se sempre sobre a classe mais provável para um mesmo documento d, não alterado assim $P(d)$.

Ainda segundo Jurafsky e Martin (2021), Naive Bayes é considerado um modelo generativo porque, a partir da Equação 4.3, pode-se entender que um documento é gerado da seguinte forma: Primeiro uma classe é amostrada de $P(c)$ e, então, as palavras são geradas por amostragem de $P(d|c)$.

Assim, seguindo os autores e voltando à classificação, determina-se a classe mais provável $\hat{c}$ para algum documento d escolhendo a classe que tem o maior valor quando calculado o produto da probabilidade a priori da classe, $P(c)$, e a probabilidade condicional do documento dada a classe, $P(d|c)$.

Sem perder a generalização, é possível representar o documento d como um conjunto de *tokens* $x_1, x_2, \ldots, x_p$:

$$\hat{c} = \arg\max_{c \in C} P(x_1, \ldots, x_p|c)P(c) \tag{4.4}$$

Em seguida, para tornar o cálculo viável, é preciso fazer duas simplificações. A primeira é de que a ordem na qual os *tokens* aparecem no documento não importa. Já a segunda é de que as probabilidades $P(x_j|c)$, onde $1 \leq j \leq p$, são independentes. Essa última tem relação direta com o uso do termo *naive* (ingênuo), pois a independência condicional entre os atributos dada a classe na maioria das vezes não se verifica. Apesar disso, o algoritmo é conhecido pelo bom desempenho em tarefas cotidianas.

Assim, a equação final para a classe escolhida por um classificador Naive Bayes é dada por:

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{x \in \mathcal{X}} P(x|c) \tag{4.5}$$

Como já exposto, Naive Bayes pode ser caracterizado como um classificador generativo. Esse tipo de classificador produz um modelo de como uma classe poderia gerar os dados de entrada. Logo, dada uma observação, ele retorna a classe com maior probabilidade de ter gerado a observação. Segundo James *et al.* (2021), Naive Bayes, por ser um modelo generativo, apresenta vantagens em relação à Regressão Logística, classificador discriminativo, quando há separação substancial entre as classes, pois nesse caso as estimativas de parâmetros para o modelo de Regressão Logística são instáveis. Além disso, se a distribuição dos preditores for aproximadamente normal em cada uma das categorias e o tamanho da amostra for pequeno, Naive Bayes pode ser mais acurado do que a Regressão Logística. Assim, por conta dos aspectos distintos dos dois classificadores, a Regressão Logística também foi utilizada e é apresentada na sequência.

4.5.2 Regressão Logística - RL

Conforme já mencionado, a Regressão Logística é considerada um classificador discriminativo. Esses classificadores aprendem quais características dos dados de entrada são mais úteis para discriminar entre as diferentes classes possíveis. Jurafsky e Martin (2021) utilizam uma metáfora visual para explicar a diferença entre os classificadores generativos e os discriminativos. Eles sugerem que o leitor imagine que o objetivo é distinguir entre imagens de gatos e imagens de cachorros e explicam que um modelo generativo teria como objetivo entender como cada um dos dois animais se parece. Já um modelo discriminativo tentaria aprender a distinguir as classes (talvez sem aprender muito sobre elas). Seguindo a notação utilizada na Subseção 4.5.1, tem-se que, diferente do classificador generativo, o classificador discriminativo, no contexto de classificação de textos, tenta calcular diretamente $P(c|d)$.

A Regressão Logística também pode ser caracterizada como um classificador linear, lida melhor com variáveis preditoras correlacionadas do que Naive Bayes e garante uma boa interpretabilidade dos resultados. Além disso, se comporta de forma superior ao classificador apresentado anteriormente quando não há uma separação substancial entre todas as classes.

De acordo com Hastie *et al.* (2009), o modelo de Regressão Logística surge do desejo de modelar as probabilidades a posteriori das M classes, pertencentes ao conjunto $C = \{c_1, \ldots, c_M\}$ de todas as classes, por meio de funções lineares das variáveis preditoras x, garantindo que essas probabilidades somem um e pertençam ao intervalo $[0, 1]$. O modelo tem a forma[16]:

$$\begin{aligned} \log\left(\frac{\mathrm{P(C=1|X=}x\mathrm{)}}{\mathrm{P(C=M|X=}x\mathrm{)}}\right) &= \beta_{10} + \beta_1^T x \\ &\vdots \\ \log\left(\frac{\mathrm{P(C=}M-1\mathrm{|X=}x\mathrm{)}}{\mathrm{P(C=M|X=}x\mathrm{)}}\right) &= \beta_{(M-1)0} + \beta_{M-1}^T x \end{aligned} \tag{4.6}$$

onde $X = (X_1, \ldots, X_p)$ são as p variáveis preditoras e, de forma análoga, $x = (x_1, \ldots, x_p)$ são as p variáveis preditoras observadas. Além disso, $\beta_1^T, \ldots, \beta_{M-1}^T$ também são vetores de tamanho p.

[16] De acordo com a notação utilizada por Hastie *et al.* (2009), X é uma forma genérica de se referir a uma variável e x é uma forma genérica de se referir aos seus valores observados. Neste trabalho x é o vetor de *tokens* atribuído a uma determinada empresa.

O modelo é especificado em termos das *M-1* transformações *logit* em relação à classe *M*, sendo essa uma escolha arbitrária. Assim, pode-se mostrar que:

$$\begin{aligned} \mathrm{P(C{=}}m|\mathrm{X{=}}x) &= \frac{\exp\left(\beta_{m0} + \beta_m^T x\right)}{1 + \sum_{l=1}^{M-1} \exp\left(\beta_{l0} + \beta_l^T x\right)}, \quad para \quad m = 1, \ldots, M-1 \\ \mathrm{P(C{=}}m|\mathrm{X{=}}x) &= \frac{1}{1 + \sum_{l=1}^{M-1} \exp\left(\beta_{l0} + \beta_l^T x\right)}, \quad para \quad m = M \end{aligned} \tag{4.7}$$

Modelos de Regressão Logística são, geralmente, ajustados por máxima verossimilhança, usando a probabilidade condicional de C dado X. Como $P(C|X)$ especifica completamente a distribuição condicional, a distribuição multinomial é apropriada. A função log-verossimilhança para N observações é:

$$l(\theta) = \sum_{i=1}^{N} log p_{c_i}(x_i; \theta), \tag{4.8}$$

onde $\theta = \{\beta_{10}, \beta_1^T, \ldots, \beta_{(M-1)0}, \beta_{M-1}^T\}$ é o conjunto de parâmetros, i é a observação considerada de um conjunto de N observações e $p_m(x_i; \theta) = P(C = m|X = x_i; \theta)$.

Agora, considerando apenas duas classes, para fins de simplificação, é conveniente codificar c_i a partir de uma resposta y_i de 0 ou 1, de tal forma que $y_i = 1$ quando $c_i = 1$ e $y_i = 0$ quando $c_i = 2$. Assim, $p_1(x; \theta) = p(x; \theta)$ e $p_2(x; \theta) = 1 - p(x; \theta)$ e a função log-verossimilhança pode ser reescrita como:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} \{y_i log p(x_i; \beta) + (1 - y_i) log(1 - p(x_i; \beta))\} \\ &= \sum_{i=1}^{N} \{y_i \beta^T x_i - log(1 + e^{\beta^T x_i})\} \end{aligned} \tag{4.9}$$

onde $\beta = \{\beta_{10}, \beta_1\}$ e se assume que o vetor de entrada x_j inclui um termo constante 1 para acomodar o intercepto, nesse caso β_{10}.

Vale também mencionar que à expressão exibida na Equação 4.9 pode ser ainda adicionado um termo de regularização. Conforme Hastie *et al.* (2009), ao utilizar um método de regularização os coeficientes são restringidos ou, equivalentemente, reduzidos para zero. Os métodos de regularização contribuem para a redução da variância.

Uma nova versão da Equação 4.9 ao se considerar a regularização conhecida como L_1 (ou Lasso), sendo essa apenas uma das possíveis técnicas a serem implementadas, pode ser vista a seguir:

$$l(\beta) = \sum_{i=1}^{N} \{ y_i(\beta_{10} + \beta_1^T x_i) - log(1 + e^{\beta_{10} + \beta_1^T x_i}) \} - \lambda \sum_{j=1}^{p} |\beta_j| \tag{4.10}$$

onde λ é o coeficiente de regularização e o intercepto é apresentado apartado dos demais parâmetros por comumente não ser penalizado quando do uso da regularização L_1.

Na sequência, os valores de β que maximizam $l(\beta)$ devem ser encontrados. Quando não existe uma expressão analítica para esse estimador é necessário recorrer a métodos de otimização numérica.

Neste estudo, empregou-se uma abordagem de classificação *uma contra todas* (*one-vs-rest*), ou seja, cada vez comparando uma classe com as restantes, fazendo C=2 classes, quando a quantidade de classes era muito grande. Já quando a quantidade de classes era pequena utilizou-se, de fato, um modelo multiclasse, com C>2.

Segundo James *et al.* (2021), Métodos de Suporte Vetorial, por conta das suas características, são classificadores intimamente relacionados à Regressão Logística, geralmente se comportando melhor o primeiro quando as classes apresentam maiores separações. Métodos de Suporte Vetorial também são muito utilizados em tarefas de classificação de dados textuais. Assim como o Naive Bayes e a Regressão Logística, eles também podem ser caracterizados como classificadores lineares, a depender de como são configurados. Apesar dos algoritmos de Suporte Vetorial serem computacionalmente mais custosos do que os outros dois apresentados, eles também foram testados.

4.5.3 *Support Vector Machines* (Métodos de Suporte Vetorial - MSV)

Os algoritmos de Suporte Vetorial podem ser vislumbrados, de forma simplificada, como a busca pelo melhor hiperplano separador das classes, sendo esse obtido através da maximização da margem. A margem pode ser entendida como a menor distância entre o hiperplano e os pontos de treino. Já os pontos cujas distâncias ao hiperplano separador sejam iguais à margem são chamados de vetores de suporte. Deste modo, dado um espaço de dimensão *p* e um subespaço de dimensão *p-1*, um exemplo de hiperplano separador pode ser visto como:

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = 0 \tag{4.11}$$

Ao se considerar um problema de classificação binário, se um ponto com coordenadas $(x_1, \ldots, x_p)$, satisfazendo a Equação 4.11, produz $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p > 0$, recebe uma determinada classificação e, caso contrário, recebe outra.

No caso mais simples, quando as classes são perfeitamente separáveis por uma fronteira, o classificador fica conhecido como de margem máxima. Já quando as classes não são totalmente separáveis ou não há interesse nesse tipo de divisão, o classificador é chamado de margem flexível e os pontos que não podem ser separados corretamente podem ser ponderados para reduzir sua influência.

James *et al.* (2021) chamam os classificadores apresentados anteriormente de Classificadores de Suporte Vetorial e consideram os MSV a extensão desses, os vislumbrando como o resultado da ampliação do espaço característico (*feature space*) através de *kernels*. Assim, é possível obter limites não lineares entre as classes.

De acordo com os autores, o Classificador de Suporte Vetorial Linear pode ser representado como:

$$f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i \langle x, x_i \rangle, \tag{4.12}$$

onde existem N parâmetros α_i, $i = 1, \ldots, N$, um por observação treinada e $\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$ é produto interno de duas observações, x_i e $x_{i'}$. Logo, para estimar os parâmetros $\alpha_1, \ldots, \alpha_N$ e β_0, todo o necessário são $\binom{N}{2}$ produtos internos entre todos os pares de observações de treinamento.

A partir da Equação 4.12 é possível notar que para calcular a função $f(x)$ é preciso calcular o produto interno entre o novo ponto x e cada um dos pontos de treinamento x_i. No entanto, verifica-se que α_i diferente de zero apenas para os vetores de suporte.

Agora, seguindo ainda James *et al.* (2021), substitui-se o produto interno por uma generalização da seguinte forma:

$$K(x_{ij} x_{i'j}) = \sum_{j=1}^{p} x_{ij} x_{i'j}, \tag{4.13}$$

onde K é alguma função, intitulada *kernel*, que quantifica a similaridade de duas observações.

A Equação 4.13 exibe o *kernel* conhecido como linear, pois o Classificador de Suporte Vetorial é linear nas variáveis preditoras. Porém, ao invés disso, $\sum_{j=1}^{p} x_{ij} x_{i'j}$ poderia ser substituído por:

$$K(x_{ij} x_{i'j}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^z \tag{4.14}$$

O *kernel* exibido na Equação 4.14 é conhecido como *kernel* polinomial de grau z, onde z é um inteiro positivo. Quando o Classificador de Suporte Vetorial é combinado com um

kernel não linear, como em 4.14, o classificador resultante é conhecido como Método de Suporte Vetorial[17]. O *kernel* polinomial é apenas um exemplo de *kernel* não linear, existindo diversas outras possibilidades além dessa.

Conforme já mencionado, Métodos de Suporte Vetorial e a Regressão Logística apresentam similaridades. Segundo James *et al.* (2021), para o primeiro, observações posicionadas no lado correto da margem não afetam o classificador e, no caso da Regressão Logística, o seu valor é pequeno para observações que estão longe do limite de decisão. Por conta disso, algoritmos de Métodos de Suporte Vetorial tendem a fornecer resultados semelhantes aos da Regressão Logística, geralmente se comportando melhor quando as classes apresentam maiores separações. Os autores pontuam ainda que abordagens não lineares podem ser implementadas com os dois algoritmos, mas que, por razões históricas, o uso de *kernels* não lineares é muito mais difundido no contexto de Métodos de Suporte Vetorial do que no de Regressão Logística e outros métodos.

Berardi *et al.* (2015)[18] obtiveram bons resultados ao utilizar Métodos de Suporte Vetorial, adotando o *kernel* linear e uma metodologia que levava a hierarquia em consideração, para classificar atividades econômicas de empresas através de textos em língua inglesa. Por conta disso, e também objetivando testar uma perspectiva não linear, eles foram aplicados considerando uma abordagem de classificação *uma contra todas*.

Já os métodos de árvores não costumam ser uma boa escolha para tarefas que incluem dados que têm alta dimensão e um número de observações não muito grande. Eles, geralmente, têm um alto custo computacional e não geram benefícios suficientes quando comparados aos resultados obtidos com classificadores lineares. Ainda assim, Caterini (2018)[19] obteve os seus melhores resultados ao utilizar um desses algoritmos para classificar atividades econômicas de empresas através de dados textuais em língua italiana. Por isso, optou-se por testar o desempenho do Boosting quando implementada a classificação plana. Ele é apresentado a seguir.

4.5.4 Boosting

Modelos de árvores são baseados numa segmentação do espaço gerado pelas variáveis preditoras em várias regiões, com base em alguma medida de erro, em que a moda, no caso de variáveis respostas categorizadas, é utilizada como predição. Desse modo, decisões complexas e globais são aproximadas por uma série de decisões mais simples e locais. Como o conjunto de regras de divisão usado pode ser resumido em uma árvore, eles são conhecidos como métodos de árvores de decisão.

---

[17] Apesar de James *et al.* (2021) diferenciarem Classificadores de Suporte Vetorial de Métodos de Suporte Vetorial, neste trabalho ambos são tratados como Métodos de Suporte Vetorial.

[18] Vale a pena destacar que apesar de Berardi *et al.* (2015) terem obtido um bom desempenho ao utilizar Métodos de Suporte Vetorial, o foco dos autores não era comparar diferentes classificadores, mas sim confrontar os desempenhos de um mesmo classificador em diferentes conjuntos de variáveis preditoras.

[19] Destaca-se que neste caso, diferente dos outros trabalhos citados, o texto utilizado tinha como finalidade a descrição da atividade da empresa, isto é, foi fornecido por ela com esse fim. Além disso, vale mencionar que os modelos comparados no trabalho foram Florestas Aleatórias, Naive Bayes e Análise Linear Discriminante.

Árvores não são muito robustas, isto é, uma pequena mudança nos dados pode causar uma grande mudança na estimativa final da árvore. Visando contornar esse tipo de problema, adotam-se métodos que combinam várias árvores. Neste trabalho foi utilizado o Boosting.

No Boosting, as árvores são geradas sequencialmente a partir de um único conjunto de treinamento, ou seja, cada árvore cresce utilizando informações de árvores anteriores. Assim, cada nova árvore ajusta os resíduos que ficam da árvore anterior. A ideia básica é considerar um conjunto de classificadores fracos de modo a obter um classificador forte.

Para simplificar, na continuação, será considerado um problema de regressão e não de classificação, ou seja, uma questão na qual as variáveis a serem buscadas são quantitativas e não qualitativas como aqui proposto. De acordo com James *et al.* (2021), o Boosting pode ser descrito pelo seguinte algoritmo:

1. Faça $\hat{f}(x) = 0$ e $r_i = y_i$ para todo i no conjunto de treino, sendo $1 \leq i \leq N$ e N o número de observações.
2. Para $b = 1, 2, \ldots, B$, repita:
   (a) Ajuste uma árvore $\hat{f}^b$ com m divisões ($m + 1$ nós terminais[20]) para os dados de treino (X, r).
   (b) Atualize $\hat{f}$ adicionando a versão reduzida da nova árvore:
   $$\hat{f}(x) \longleftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{4.15}$$
   (c) Atualize os resíduos:
   $$r_i \longleftarrow r_i - \lambda \hat{f}^b(x_i) \tag{4.16}$$
3. Saída do modelo.
   $$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{4.17}$$

O propósito desse procedimento é que a aprendizagem ocorra vagarosamente. De tal forma que, dado o modelo atual, ajusta-se uma árvore de decisão aos seus resíduos, representados por r. Em outras palavras, ajusta-se uma árvore usando as variáveis preditoras, retratadas por X, e os resíduos atuais, ao invés do resultado Y, como resposta. Na sequência, a nova árvore é adicionada a função ajustada para atualizar os resíduos. O número de árvores é denotado por B no algoritmo exposto. Cada uma dessas novas árvores pode ser pequena, com apenas alguns nós terminais, determinados pelo parâmetro m do algoritmo. Com esse tipo de ajuste, melhora-se $\hat{f}$ em áreas onde ela não funciona bem. Já o parâmetro de encolhimento (*shrinkage parameter*) λ retarda ainda mais o processo,

[20] São chamados de nós terminais (*terminal nodes*) as regiões oriundas da segmentação do espaço gerado pelas variáveis preditoras.

permitindo que diferentes tipos de árvores ajustem os resíduos. Em geral, abordagens de aprendizado automático que aprendem lentamente tendem a executar bem a tarefa. Detalhes mais técnicos focados na tarefa de classificação podem ser obtidos em Hastie *et al.* (2009).

Os classificadores apresentados foram testados e os seus hiperparâmetros foram otimizados através da validação cruzada.

4.6 Validação Cruzada

A validação cruzada é um método de reamostragem cujo objetivo é extrair vários conjuntos de validação e treino do conjunto de treino inicial para avaliar o desempenho em dados independentes ou selecionar o nível de flexibilidade adequado para o método de aprendizagem utilizado.

De acordo com James *et al.* (2021), na validação cruzada de ordem k[21] (*k-fold cross validation*) o conjunto de treino, originalmente definido na Seção 3.4, é dividido em k grupos de tamanho aproximadamente igual. Inicialmente, o primeiro grupo é tratado como um conjunto de validação e o método é ajustado nos k-1 grupos restantes. As observações mal classificadas são quantificadas de acordo com a métrica escolhida no conjunto considerado como de validação. Esse procedimento é repetido k vezes, sendo um grupo diferente de observações tratado como conjunto de validação a cada iteração. Ao final, é obtida a média dos valores quantificados em cada conjunto de validação.

Para a execução do trabalho foi utilizada validação cruzada de ordem 3 (*3-fold cross validation)* e considerada a mesma estratificação mencionada na Seção 3.4. A escolha da ordem 3, apesar de não ser a mais indicada de acordo com a literatura a respeito do tema, foi pautada no custo computacional envolvido, já que esse tipo de abordagem envolve ajustar o mesmo método várias vezes usando diferentes subconjuntos dos dados, o que pode ser computacionalmente caro.

Em conjunto com a reamostragem, visando a otimização dos hiperparâmetros, ou seja, das informações a serem fornecidas para os algoritmos antes do processo de modelagem, foi empregado o *Random Search* (Busca Aleatória).

4.7 *Random Search* (Busca Aleatória)

Hiperparâmetro é como é denominado um parâmetro que, diferente de outros, não pode ser aprendido pelo modelo a partir dos dados de treino, devendo assim ser fornecido para o algoritmo (Jurafsky e Martin, 2021). Distintos valores desses são capazes de gerar diferentes desempenhos. Pode ser visto como exemplo de hiperparâmetro a técnica de regularização citada na Subseção 4.5.2 e a quantidade de árvores mencionada na Subseção 4.5.4.

Existem várias formas de se determinar os hiperparâmetros. A busca aleatória, diferente da abordagem *grid search* (busca em grade) que executa o modelo em toda extensão de

[21] Quando k é igual ao número de observações usa-se a denominação *leave-one-out cross validation.*

hiperparâmetros definidos pelo desenvolvedor, realiza o teste apenas para um grupo de configurações escolhidas por meio de amostragem. A quantidade de amostras selecionadas fica a cargo do usuário.

De acordo com Bergstra e Bengio (2012), para a maioria dos conjuntos de dados apenas alguns dos hiperparâmetros realmente importam, mas diferentes hiperparâmetros são importantes em diferentes conjuntos de dados. Esse fenômeno faz com que a busca em grade não seja a melhor escolha para configurar algoritmos para conjuntos de dados inéditos.

A busca aleatória foi empregada junto com a validação cruzada, sendo assim, cada combinação foi executada em 3 conjuntos de validação diferentes. Além disso, foram testadas pelo menos 30 combinações para cada combinação de aplicação ou não do *stemmer* e do uso ou não da *SMOTE* para balanceamento dos dados. Dessas, para a que apresentou o melhor desempenho, foram testadas pelo menos outras 30 combinações com o objetivo de refinar a busca na vizinhança dos valores de hiperparâmetros que geraram resultados mais satisfatórios. As combinações de hiperparâmetros usadas e a aplicação ou não do IDF foram diferentes dentro de cada uma das 4 combinações (*SMOTE* = [Sim, Não] e *STEMMER* = [Sim, Não]). Os valores adotados na busca aleatória podem ser vistos no Apêndice A.

As métricas utilizadas para avaliação dos resultados, escolha dos modelos e dos seus hiperparâmetros são apresentadas na sequência.

4.8 Métricas para Avaliação de Desempenho

Existem variadas métricas para avaliação de desempenho de algoritmos e a escolha de uma delas está intimamente relacionada às características da tarefa que se deseja realizar e à configuração da amostra disponível. A maioria das formas de avaliação para problemas de classificação estão baseadas na matriz de confusão. Um exemplo dessa matriz, ao se considerar uma classificação binária, pode ser visto a seguir.

		Valores Reais	
		Positivo	Negativo
Predição	Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)

A partir da matriz de confusão é possível obter as seguintes métricas:

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + FN + VN} \tag{4.18}$$

$$\text{Recall} = \frac{VP}{VP + FN} \tag{4.19}$$

$$\text{Precision} = \frac{VP}{VP + FP} \tag{4.20}$$

$$\text{F1 Score} = \frac{2 \cdot (Precision \cdot Recall)}{(Precision + Recall)} \tag{4.21}$$

De forma sucinta, tem-se que a acurácia corresponde à probabilidade de resultados corretos, *recall* reflete a capacidade do classificador de encontrar todas as amostras positivas, enquanto *precision* está relacionada com a sua habilidade de não rotular como positiva uma observação negativa. Já o *F1 score* combina *precision* e *recall* de modo a trazer um número único que indique a qualidade geral do modelo. Dito isso, se faz importante destacar que, como o problema em questão trata de 78 classes e não de apenas duas, as interpretações da matriz e das métricas oriundas dela devem ser realizadas considerando a substituição das categorias *Positivo* e *Negativo* pelas 78 exibidas na Tabela 3.3.

Vale ainda mencionar que em problemas com múltiplas classes os resultados podem ser fornecidos considerando a *Macro F1 score* e a *Micro F1 score*. A primeira retrata a média aritmética simples da *F1 score* em cada classe, ou seja, desconsidera diferenças de tamanho. Já a segunda calcula as métricas globalmente contando o total de verdadeiros positivos, falsos negativos e falsos positivos.

As medidas exibidas nas Equações 4.18, 4.19, 4.20 e 4.21 não são as mais adequadas para categorização hierárquica, uma vez que tratam da mesma forma os diferentes tipos de classificações incorretas. Assim, a partir dessas, que podem ser intituladas métricas "planas", uma empresa classificada na pesquisa incorreta receberia a mesma penalidade que uma empresa classificada incorretamente em nível de Grupo na PAC, por exemplo. Em outras palavras, a classificação incorreta para um irmão ou nó pai da categoria correta é penalizada da mesma forma que a classificação incorreta para um nó distante.

Objetivando contornar esse problema Kiritchenko *et al.* (2006) propuseram as seguintes métricas:

$$\text{Hierarchical Recall } (h\text{R}) = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{T}_i|} \tag{4.22}$$

$$\text{Hierarchical Precision } (h\text{P}) = \frac{\sum_i |\hat{P}_i \cap \hat{T}_i|}{\sum_i |\hat{P}_i|} \tag{4.23}$$

$$hF = \frac{2 \cdot (hP \cdot hR)}{(hP + hR)}, \tag{4.24}$$

onde $\hat{P}_i$ é um conjunto que engloba a classe prevista mais específica e todos os seus ancestrais para uma observação e $\hat{T}_i$ abarca a classe verdadeira mais específica e seus ancestrais para a mesma observação. Já o índice i representa cada uma das empresas alocadas na base.

As métricas apresentadas são versões estendidas das métricas conhecidas, *recall*, *precision* e *F1 Score*, adaptadas ao cenário de classificação hierárquica. De acordo com os desenvolvedores, a métrica hierárquica *h*F atende aos três requisitos desejáveis elencados pelos mesmos, a saber: a medida dá crédito à classificação parcialmente correta; a medida pune erros distantes mais fortemente; e a medida pune erros em níveis mais altos de uma hierarquia de forma mais severa.

Optou-se por utilizar as métricas exibidas nas Equações 4.22, 4.23 e 4.24 para avaliar os resultados obtidos a partir da base de teste, além das métricas apresentadas nas Equações 4.18, 4.19, 4.20 e 4.21, devido à facilidade de computar e o fato da árvore na qual a taxonomia é organizada, neste problema, não ser profunda. Contudo, não existe um consenso no que diz respeito às métricas a serem adotadas quando se trata de um problema de classificação hierárquica, sendo possível encontrar outras propostas diferentes, como em Kosmopoulos *et al.* (2015).

Capítulo 5

Apresentação do Modelo Final e Análise dos Resultados

Objetivando a implementação e a avaliação dos passos descritos no Capítulo 4, foram utilizadas as bases de treino e de teste mencionadas na Seção 3.4. A divisão em treino e teste tem como função evitar que os dados de teste impactem no treinamento dos modelos, fornecendo assim informação confiável sobre o desempenho dos algoritmos na realização da tarefa de classificação quando empregados na base de teste.

Neste capítulo, inicialmente, são apresentados os resultados obtidos através da validação cruzada nos dados de treino considerando classificadores locais e planos. Nesse passo, são testadas diferentes formas de representação, algoritmos, hiperparâmetros e o tratamento ou não para dados desbalanceados.

Na sequência, a melhor configuração oriunda da etapa anterior é considerada para o ajuste em toda a base de treino e são exibidos os dez *tokens* com maiores coeficientes por classe considerando a abordagem local e a plana. Essa exposição tem como objetivo garantir que os *tokens* de fato têm relação com o fenômeno estudado[1].

Na fase seguinte são expostos os resultados oriundos da base de teste considerando, mais uma vez, as duas abordagens e as métricas tradicionais e hierárquicas. Nesse passo, busca-se também exibir contrapontos entre os classificadores locais e o plano. Além disso, realiza-se uma investigação do histórico das empresas classificadas incorretamente e são fornecidos dados sobre o seu comportamento em 2020.

[1] Daas e Doef (2021) relataram um bom desempenho ao tentar prever se uma empresa era inovadora ou não a partir do texto oriundo da *web*, porém se depararam com um modelo final repleto de *tokens* de tamanho 2 com alto coeficiente, o que os fez inspecioná-los. Durante a averiguação observaram que muitos deles tinham como origem *e-mails* e *links* exibidos nas páginas da *web* das companhias, sugerindo que empresas inovadoras focavam mais em exibir esse tipo de informação. Como não era essa relação que estavam buscando, pois ela indicava que um *site* preenchido apenas com endereços de *e-mail* e *links* seria classificado como inovador, decidiram ignorar os *tokens* de tamanho igual a 2 e focaram em desenvolver um modelo que considerasse *tokens* de tamanho igual a 3 ou maiores.

5.1 Resultados Obtidos na Base de Treino

A partir da base de treino foram considerados 4 cenários, com as combinações da utilização ou não da *SMOTE* e a aplicação ou não do *stemmer*. Os melhores resultados obtidos, via validação cruzada, para cada uma delas e os hiperparâmetros escolhidos podem ser vistos no Apêndice B. A partir da Tabela B.1 até a Tabela B.7 constam os números alcançados através dos classificadores locais. Já na Tabela B.8 estão os oriundos da abordagem plana. A métrica utilizada foi a *Macro F1 score*.

A seguir, podem ser observados os melhores resultados encontrados, por algoritmo e via validação cruzada, ao se considerar cada tipo de classificador e também são apresentados alguns *tokens* por classe para os modelos finais ajustados em toda a base de treino. Quanto a esse último ponto, é possível verificar nas tabelas uma nota de rodapé referente à omissão de termos relacionados, ou possivelmente relacionados, a nomes próprios de empresas ou marcas. Imagina-se que esse fato tenha relação com a pouca quantidade de observações em determinadas classes, com as características da atividade e do setor no qual se insere a CNAE e, além disso, com a forma como a coleta foi realizada, descrita na Subseção 3.2.1, na qual se buscaram textos associados a produtos e serviços. Não foram encontrados indícios de que os modelos estivessem utilizando o nome de cada empresa para classificar. Vale lembrar que a retirada de *tokens* de acordo com a sua cobertura, conforme mencionado na Subseção 4.1.1, corrobora com a não ocorrência de *overfitting* (super-ajuste), bem como o uso da regularização.

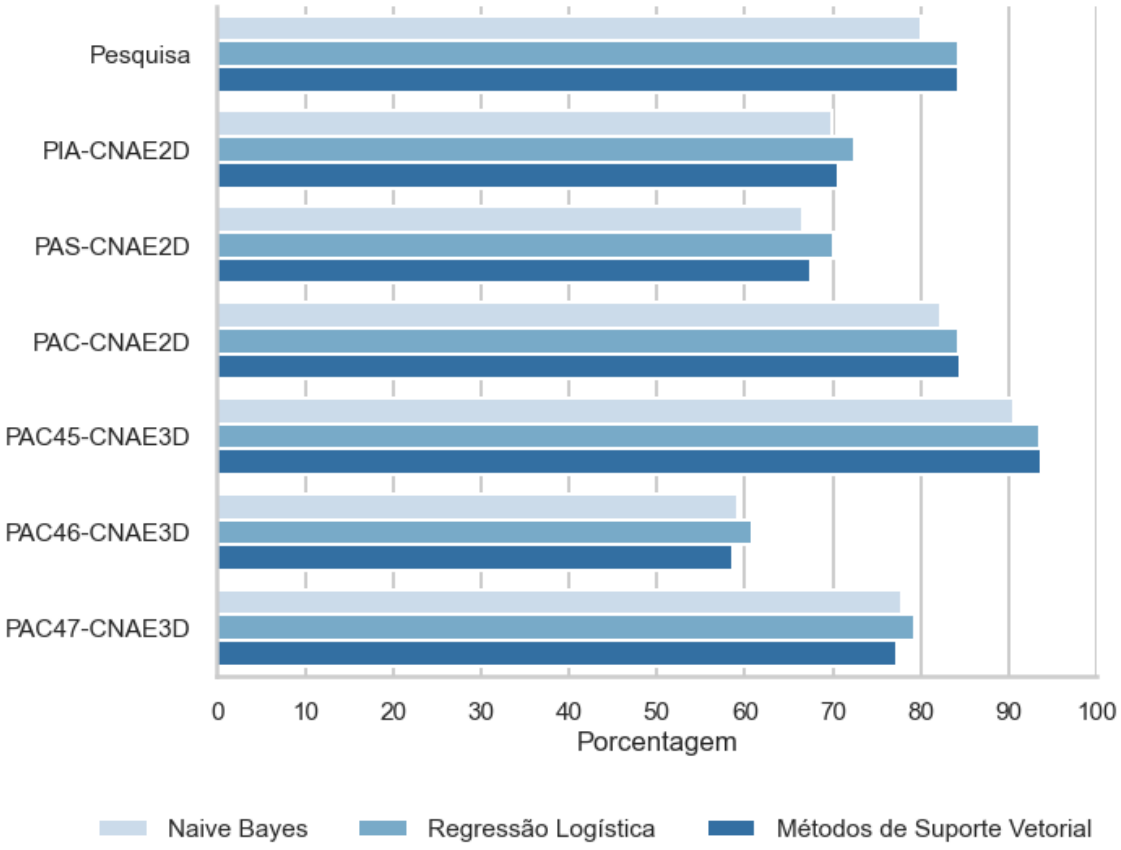
5.1.1 Resultados Obtidos com os Classificadores Locais

Conforme mencionado na Seção 4.4 e na Seção 4.5, ao se utilizar classificadores locais o nível Seção da CNAE foi desconsiderado. Além disso, para cada nó foram testados e otimizados 3 algoritmos, a saber, Naive Bayes, Regressão Logística e Métodos de Suporte Vetorial.

De acordo com os resultados obtidos, o uso da *SMOTE* gerou alguma melhora no desempenho da maioria dos algoritmos e na maior parte dos nós, sendo mais evidente quando Naive Bayes foi aplicado. Já o *stemmer*, na maior parte das vezes, não gerou ganho. Contudo, os resultados foram muito próximos. Dado o modelo final escolhido, pode-se verificar que o *stemmer* foi aplicado em 3 nós e a *SMOTE* utilizada em 4. Já o *kernel* que apresentou melhor desempenho ao se considerar Métodos de Suporte Vetorial variou de acordo com o nó. Além disso, a representação via TF-IDF se mostrou preferível.

Os melhores resultados alcançados, a partir da base de treino, para cada algoritmo e para cada nó são apresentados na Figura 5.1. As denominações utilizadas são as mesmas mencionadas na Figura 4.2, exibida na Seção 4.4.

Figura 5.1: Melhores resultados obtidos, através dos classificadores locais, considerando a métrica *Macro* F_1 *score*, a partir da base de treino, por nó, segundo os algoritmos testados



Fonte: Autoria própria.

Como pode ser visto, a Regressão Logística obteve um bom desempenho quando comparada com os demais algoritmos. A Regressão Logística possui uma fácil interpretabilidade, característica essa muito desejável, principalmente, no contexto das estatísticas oficiais. Por conta disso, esse algoritmo foi preferido para executar a tarefa nos dados de teste em todos os nós.

Antes da apresentação dos resultados, vale ainda destacar que, segundo James *et al.* (2021), deve-se ter muita cautela ao exibir o produto de procedimentos de regressão no cenário de alta dimensão por conta da multicolinearidade, o conceito de que as variáveis em uma regressão podem estar correlacionadas entre si. No contexto de alta dimensão, o problema da multicolinearidade é extremo, qualquer variável no modelo pode ser escrita como uma combinação de todas as outras. Isso, de acordo com os autores, significa, essencialmente, que nunca é possível saber exatamente quais variáveis (se houver) são realmente preditoras do resultado, bem como identificar os melhores coeficientes para uso na regressão. No máximo, espera-se atribuir grandes coeficientes para variáveis que estão correlacionadas com as que realmente predizem o resultado.

A seguir, são mostrados alguns detalhes de cada um dos modelos finais ajustados por nó. A configuração eleita para cada nó foi a que apresentou o melhor desempenho entre as 4 explicitadas no Apêndice B, tendo sido desconsiderado o desvio padrão pois

não ocorreram grandes variações entre eles. No Apêndice B também podem ser vistos os hiperparâmetros utilizados para o ajuste.

Nó: Pesquisa

Nessa tarefa utilizou-se a *SMOTE* e o *stemmer* não foi aplicado. Além disso, considerou-se a representação via TF-IDF. Os 10 *tokens* com maiores coeficientes, por pesquisa, no modelo de Regressão Logística ajustado podem ser vistos no Quadro 5.1.

Quadro 5.1: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por pesquisa - Nó: Pesquisa

PIA-Empresa	PAC	PAS
fabricacao	distribuidora	servicos
produtos	lojas	locacao
producao	ofertas	transporte
fabrica	distribuicao	cardapio
representantes	marcas	frota
industria	importacao	gestao
industrial	posto	andar
qualidade	produtos	hotel
linha	loja	restaurante
ind	acessorios	transportes

Fonte: Autoria própria.

Aparentemente, a maior parte dos *tokens* exibidos no Quadro 5.1 remetem às características de cada uma das pesquisas. Talvez o mais intrigante tenha sido o fato de *representantes* ter aparecido com alto coeficiente na PIA-Empresa e não na PAC. Essa é uma palavra que ocorre com frequência em denominações de CNAEs associadas ao comércio, relacionada a representantes comerciais. Outro ponto a ser destacado é que há indícios de que as variáveis preditoras oriundas do texto da *URL* foram úteis. É possível observar, na coluna referente à PIA-Empresa, o *token ind* e o Registro.br oferece uma categoria de domínio com esse mesmo nome para indústrias, sendo necessário o registro por Pessoas Jurídicas mas sem a necessidade de comprovação da atividade.

Em seguida, são apresentados os modelos ajustados para os demais nós dentro de cada uma das pesquisas. Os resultados são exibidos de acordo com os dígitos da CNAE, a denominação de cada uma delas pode ser vista no Anexo A. Além da denominação, outras explicações a respeito de cada classe podem ser consultadas em IBGE e CONCLA (2022) ou em IBGE e CONCLA (2020).

Nó: PAC-CNAE2D

Nesse passo não foi utilizada a *SMOTE* e o *stemmer* não foi aplicado. Somado a isso, considerou-se a representação via TF-IDF. Os 10 *tokens* com maiores coeficientes, por Divisão da CNAE na PAC, no modelo de Regressão Logística ajustado podem ser vistos no Quadro 5.2.

Quadro 5.2: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PAC - Nó: PAC-CNAE2D

45 (1)	46	47
pecas	distribuidora	lojas
veiculos	produtos	posto
autopecas	importacao	casa
pneus	distribuicao	loja
auto	logistica	rede
motos	agricolas	manipulacao
carro	sementes	farmacia
seminovos	papeis	postos
veiculo	industria	madeira
caminhoes	brasil	piso

Fonte: Autoria própria.

(1) Nessa classe 2 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas ou marcas (ou por haver essa possibilidade).

Ao que tudo indica, a maior parte dos *tokens* exibidos no Quadro 5.2 remetem às características de cada uma das Divisões consideradas na PAC. Um *token* que merece destaque é *industria*, ele aparece na PIA-Empresa e também na CNAE 46 (Comércio por atacado, exceto veículos automotores e motocicletas). Talvez uma razão para tanto esteja no fato da venda por atacado ser, em partes, caracterizada pela venda de mercadoria em grandes quantidades a varejistas, empresas, estabelecimentos agropecuários, cooperativas e a uma clientela institucional. A Divisão 46 conta, inclusive, com a Classe 46.63-0 (Comércio atacadista de máquinas e equipamentos para uso industrial; partes e peças) que abarca o comércio atacadista de máquinas e equipamentos para uso nos diversos ramos da indústria e o comércio atacadista de componentes não eletrônicos para máquinas e equipamentos para uso industrial, exceto para mineração e construção. *logistica* também não parece ser um *token* esperado na PAC, dado que esse tipo de atividade se concentra, principalmente, na Divisão 52 (Armazenamento e atividades auxiliares dos transportes) da PAS. Contudo, de acordo com IBGE e CONCLA (2020), a Divisão 46 "[...]compreende também as manipulações habituais do comércio atacadista - montagem, classificação e agrupamento de produtos em grande escala, fracionamento, acondicionamento e envasamento, redistribuição em recipientes de menor escala [...]"(IBGE e CONCLA, 2020, p.327) e, mais precisamente, o Grupo 46.1 (Representantes comerciais e agentes do comércio, exceto de veículos automotores e motocicletas) engloba "As atividades dos representantes e agentes do comércio que, sob contrato, comercializam a mercadoria em nome de terceiros ou fazem a intermediação entre vendedores e compradores de mercadorias no atacado [...]"(IBGE e CONCLA, 2020, p.327). Logo, os manuseios típicos do comércio atacadista e a intermediação relacionada ao Grupo 46.1 poderiam suscitar *tokens* como *distribuidora*, *distribuicao* e *logistica* na CNAE 46, fazendo com que a delimitação entre PAC e PAS seja tênue.

Nó: PAC45-CNAE3D

Nessa etapa a *SMOTE* não foi utilizada e o *stemmer* não foi aplicado. Ademais, considerou-se a representação via TF-IDF. Os 10 *tokens* com maiores coeficientes, por Grupo da Divisão 45_{pac} da CNAE, no modelo de Regressão Logística ajustado podem ser vistos no Quadro 5.3.

Quadro 5.3: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 3 dígitos da Divisão 45 da PAC - Nó: PAC45-CNAE3D

45.1 (1)	45.3	45.4 (2)
seminovos	produtos	motos
veiculos	autopecas	moto
novos	pneus	motocicletas
caminhoes	auto	abs
agendamento	empresa	capacete
vendas	aro	seminovas
onibus	suspensao	interesse
concessionaria	linha	phone
drive	automotivo	oficina
novo	fornecedores	capacetes

Fonte: Autoria própria.

(1) Nessa classe 3 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas ou marcas (ou por haver essa possibilidade). (2) Nessa classe 7 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas ou marcas (ou por haver essa possibilidade).

À vista humana, os *tokens* exibidos no Quadro 5.3 parecem estar associados aos Grupos CNAE. O Grupo 45.1 (Comércio de veículos automotores) engloba o comércio varejista de automóveis, utilitários, camionetas e similares, enquanto a venda de peças e acessórios para veículos automotores está no Grupo 45.3 (Comércio de peças e acessórios para veículos automotores). Já o comércio varejista de motocicletas e motonetas novas e usadas é representado pela CNAE 45.41-2 (Comércio por atacado e a varejo de motocicletas, peças e acessórios) e as atividades de representantes comerciais e agentes do comércio atacadista e varejista de motocicletas e motonetas, partes, peças e acessórios pela 45.42-1 (Representantes comerciais e agentes do comércio de motocicletas, peças e acessórios). Vale lembrar que as Classes 45.41-2 e 45.42-1 são as únicas abarcadas pelo Grupo 45.4 no contexto da PAC. Nesse Grupo foi necessária a omissão de vários *tokens* relacionados a nomes de marcas que dominam o setor e diversos dos seus modelos de motos.

Nó: PAC46-CNAE3D

Nessa tarefa foi utilizada a *SMOTE* e o *stemmer*. Além disso, foi considerada a representação via TF-IDF. Os 10 *tokens* com maiores coeficientes, por Grupo da Divisão 46

(Comércio por atacado, exceto veículos automotores e motocicletas) da CNAE, no modelo de Regressão Logística ajustado podem ser vistos no Quadro 5.4.

Quadro 5.4: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 3 dígitos da Divisão 46 da PAC - Nó: PAC46-CNAE3D

46.1 (1)	46.2	46.3	46.4	46.5 (2)
representaco	sement	aliment	medic	informa
drog	caf	frut	farmaceu	telecom
rep	pet	receit	saud	impres
co	soj	azeit	cirurg	conec
representaca	cooper	ingredi	hospital	notebook
assess	cour	carn	brinqued	soluco
transmis	nutrica	food	music	ti
etanol	gene	cervej	profiss	red
sobreposica	gat	vinh	tec	tecnolog
exteri	fl	kg	lent	dad
46.6	**46.7**	**46.8**	**46.9 (2)**	
maquin	eletr	lubrific	cest	
pec	ferrament	combusti	peix	
equip	construca	embal	propr	
bomb	parafus	sucat	atacad	
corre	mater	quim	and	
tra	mad	papel	importaca	
rol	vidr	aco	detail	
valvul	ferr	diesel	far	
agricol	lumin	residu	veterinari	
manutenca	painel	agricol	money	

Fonte: Autoria própria.

(1) Nessa classe 7 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas ou marcas (ou por haver essa possibilidade). (2) Nessas classes de 1 a 2 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas ou marcas (ou por haver essa possibilidade).

Ao que parece, os *tokens* exibidos no Quadro 5.4 remetem aos Grupos CNAE, apesar de ser difícil compreendê-los dada a utilização do *stemmer*. Por outro lado, o *stemmer* permitiu a exibição de mais raízes de palavras. As tabelas que exibem os termos oriundos dos modelos sem a sua aplicação contam muitas vezes com o plural e o singular (entre outras variações) de um mesmo vocábulo.

No Quadro 5.4 vale destacar a CNAE 46.1 (Representantes comerciais e agentes do comércio, exceto de veículos automotores e motocicletas) pois, diferente de todas as outras, para essa classe foi preciso substituir 7 *tokens* que se tratavam (ou incitavam dúvidas) de nomes próprios relacionados a empresas ou marcas. Além disso, também pode ser observado nessa CNAE o *token rep*, que poderia remeter à presença da categoria de domínio com esse mesmo nome oferecida pelo Registro.br a representantes comerciais.

Contudo, essa denominação só passou a ser disponibilizada a partir de 20/07/2020 e as *URLs* utilizadas têm como origem as pesquisas referentes ao ano de 2019, assim, apesar das empresas terem até o segundo semestre de 2020 para respondê-las, essa categoria dificilmente seria notada. Visa-se com essa observação destacar o quão delicado é fornecer qualquer afirmação com base nos *tokens* com maiores coeficientes.

Pode-se ainda notar que, apesar de terem sido retirados textos identificados em outras línguas, conforme explicitado na Subseção 3.2.2, ainda restaram palavras na língua inglesa, como *and*, *detail* e *money*, concentradas na CNAE 46.9 (Comércio atacadista não-especializado).

Nó: PAC47-CNAE3D

Nesse passo não se utilizou a *SMOTE* e o *stemmer* foi aplicado. Somado a isso, considerou-se a representação via TF-IDF. Os 10 *tokens* com maiores coeficientes, por Grupo da Divisão 47 (Comércio varejista) da CNAE, no modelo de Regressão Logística ajustado podem ser vistos no Quadro 5.5.

Quadro 5.5: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 3 dígitos da Divisão 47 da PAC - Nó: PAC47-CNAE3D

47.1	47.2	47.3 (1)	47.4
supermerc	chocolat	post	construca
ofert	salg	combusti	mater
receit	encomend	ole	tint
cooper	cha	km	eletr
acoug	caf	veicul	vidr
icon	tort	grup	mad
sup	frut	abastec	ferrament
text	win	red	revest
cooki	pad	dap	obr
money	natur	lubrific	industr
47.5	**47.6**	**47.7**	**47.8**
movel	livr	medic	mod
softw	esport	farmac	calc
impres	bik	saud	pet
cartuch	brinqued	ocul	fl
assistenc	papel	drog	joi
sof	curs	otic	feminin
tecnolog	bibl	perfum	plant
tud	nei	aparelh	pec
soluco	gift	lent	vest
ton	we	farmaceu	sucat

Fonte: Autoria própria.

(1) Nessa classe 1 *token* foi substituído pelo seguinte por se tratar de nome próprio relacionado a empresa ou marca (ou por haver essa possibilidade).

Ao que tudo indica, apesar de novamente não ser fácil a avaliação por conta da aplicação do *stemmer*, a maior parte dos *tokens* exibidos no Quadro 5.5 remetem às características de cada um dos Grupos presentes na Divisão 47 (Comércio varejista) da PAC. Talvez a classe de mais difícil avaliação seja a 47.8 (Comércio varejista de produtos novos não especificados anteriormente e de produtos usados) por englobar uma ampla gama de produtos no contexto do comércio varejista, tais como artigos do vestuário, joias, gás liquefeito de petróleo, artigos usados e outros artigos novos não especificados em outra CNAE. Outro ponto a ser notado são as semelhanças entre os Grupos da Divisão 46 (Comércio por atacado, exceto veículos automotores e motocicletas), no Quadro 5.4, e da Divisão 47. Pode-se verificar, por exemplo, muitos *tokens* iguais nas CNAEs 47.5 (Comércio varejista de equipamentos de informática e comunicação; equipamentos e artigos de uso doméstico) e 46.5 (Comércio atacadista de equipamentos e produtos de tecnologias de informação e comunicação), o que é esperado pois ambos abarcam equipamentos de informática, apesar de um Grupo tratar do comércio varejista e o outro do atacadista.

Nó: PIA-CNAE2D

Nessa etapa a *SMOTE* foi utilizada e o *stemmer* não foi aplicado. Ademais, foi considerada a representação via TF-IDF. Os 10 *tokens* com maiores coeficientes, por Divisão da CNAE na PIA-Empresa, no modelo de Regressão Logística ajustado podem ser vistos no Quadro 5.6.

Quadro 5.6: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PIA-Empresa - Nó: PIA-CNAE2D

(Continua)

08	10	11	13
calcario	receitas	bebidas	tecidos
mineracao	alimentos	agua	textil
pedras	sabor	refrigerantes	poliester
brita	leite	cachaca	fios
pedreira	sorvetes	vinhos	texteis
pedra	arroz	cerveja	tapetes
areia	racoes	guarana	cortinas
iodo	queijo	mineral	elasticos
gesso	cafe	cervejas	cordas
extracao	nutricao	cervejaria	algodao
14	**15**	**16 (1)**	**17**
confeccoes	calcados	madeira	papel
uniformes	couro	portas	papelao
moda	tenis	compensados	papeis
meias	couros	paletes	embalagens
lingerie	bolsas	eucalipto	etiquetas
roupas	bota	molduras	fraldas
colecao	pasta	madeiras	incontinencia
jeans	pastas	pinus	ondulado
marca	botas	florestal	caixas

Quadro 5.6: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PIA-Empresa - Nó: PIA-CNAE2D

(Continuação)

14	15	16 (1)	17
marcas	sandalia	wood	absorventes
18 (1)	**19**	**20**	**21**
impressao	lubrificantes	fertilizantes	medicamentos
grafica	etanol	tintas	farmaceutica
cartoes	alcool	cosmeticos	saude
rotulos	safra	produtos	comprimidos
cores	oleo	quimica	farma
laminacao	acucar	limpeza	diagnostico
portfolio	biodiesel	aromas	animal
graficos	oleos	fispq	quimica
encadernacao	bioenergia	arla	farmacovigilancia
convites	programa	quimicos	medicamento
22	**23**	**24**	**25**
plasticas	concreto	fundicao	metalicas
borracha	vidros	aluminio	parafusos
pneus	ceramica	ferro	metalicos
pvc	vidro	tubos	aco
plasticos	cimento	perfis	ferramentas
filmes	cal	aco	caldeiras
plastico	abrasivos	ligas	moldes
pead	obras	metais	esquadrias
injecao	porcelana	steel	latas
bucha	ceramicos	tubo	fixadores
26	**27**	**28**	**29**
medicao	baterias	equipamentos	veiculos
suporte	eletricos	maquinas	carrocerias
eletronicos	transformadores	assistencia	componentes
antenas	luminarias	refrigeracao	onibus
eletronica	tensao	transportadores	retifica
relogio	bateria	bombas	caminhoes
audio	cabo	valvula	rodoviarios
medidores	energia	irrigacao	automotivo
telecomunicacoes	iluminacao	engrenagens	automotivas
rede	cabos	valvulas	veiculo
30 (1)	**31**	**32**	**33 (1)**
embarcacoes	moveis	brinquedos	manutencao
bicicletas	estofados	joias	servicos
estaleiro	colchoes	botoes	cesto
carrinhos	mobiliario	hospitalares	reparo
rodas	ambientes	lentes	motores
bike	colchao	escovas	containers

Quadro 5.6: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PIA-Empresa - Nó: PIA-CNAE2D

(Continuação)

30 (1)	31	32	33 (1)
moto	design	bolas	industriais
guidao	cadeiras	instrumentais	montagem
aro	cozinhas	implantes	opcoes
motos	divisorias	esteril	lubrificacao

Fonte: Autoria própria.

(1) Nessas classes de 1 a 2 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas ou marcas (ou por haver essa possibilidade).

Pode-se observar que os *tokens* exibidos no Quadro 5.6 parecem estar associados a cada uma das Divisões consideradas na PIA-Empresa. Um aspecto interessante é como o modelo foi capaz, em alguns casos, de caracterizar classes próximas, como a 13 (Fabricação de produtos têxteis) e a 14 (Confecção de artigos do vestuário e acessórios), com *tokens* distintos. Já em outros casos, um mesmo *token* apresentou um alto coeficiente em duas classes, como *aco* na 24 (Metalurgia) e na 25 (Fabricação de produtos de metal, exceto máquinas e equipamentos). É possível também notar a existência de *tokens* bastante específicos, como por exemplo, *fisqp*[2] e *arla*[3] na CNAE 20 (Fabricação de produtos químicos) e *pead*[4] na CNAE 22 (Fabrição de produtos de borracha e de material plástico).

Através dos Quadros 5.6 e 5.1 pode-se verificar a dificuldade em classificar a atividade econômica da empresa. Nota-se que uma palavra a princípio muito específica da PAS, como *servicos*, aparece na Divisão 33 (Manutenção, reparação e instalação de máquinas e equipamentos) da PIA-Empresa. Vale pontuar que a Divisão 33 compreende as atividades de manutenção, reparação e instalação de máquinas e equipamentos utilizados no processo de produção industrial, realizadas por unidades especializadas, normalmente sob contrato. Sendo *servicos* um *token* comum a diversas classes, espera-se que os demais sejam capazes de distingui-las. O mesmo se aplica a outras palavras, como *pneus* na Divisão 22 da PIA-Empresa e na Divisão 45_{pac}.

Nó: PAS-CNAE2D

Nessa tarefa foi utilizada a *SMOTE* e o *stemmer*. Além disso, considerou-se a representação via TF-IDF. Os 10 *tokens* com maiores coeficientes, por Divisão da CNAE na PAS, no modelo de Regressão Logística ajustado podem ser vistos no Quadro 5.7.

[2] Com base na análise de alguns textos, acredita-se que a sigla FISPQ, no contexto da CNAE 20, signifique Ficha de Informação de Segurança de Produtos Químicos.

[3] Com base na análise de alguns textos, acredita-se que a sigla ARLA, no contexto da CNAE 20, signifique Agente Redutor Líquido Automotivo.

[4] Com base na análise de alguns textos, acredita-se que a sigla PEAD, no contexto da CNAE 22, signifique Polietileno de Alta Densidade.

Quadro 5.7: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PAS - Nó: PAS-CNAE2D

(Continua)

01 (1)	38	45	49	50
agricol	residu	pint	transport	embarcaco
semen	recicl	automo	frot	marit
planti	cacamb	veicul	viaca	navegaca
plant	colet	pneu	onibu	passei
bovin	domini	funil	linh	bals
aviaca	destinaca	oficin	distribuica	lanch
ingredi	urban	reparaca	logis	fluv
certificaca	lix	pec	mudanc	apoi
florest	javascript	mecan	pass	navi
produca	ambient	frei	fret	transport
51 (1)	**52**	**53 (1)**	**55**	**56**
aeronav	estacion	entreg	hotel	cardapi
aere	logis	motoboy	suit	restaurant
voo	rodov	encomend	motel	refeico
tax	term	correi	pous	buffet
hangar	aduan	postal	acomodaco	pizz
fret	park	transport	hosped	alimentaca
hang	aeroport	utilitari	hot	delivery
farm	armazen	expr	reserv	gastronom
passag	navi	xpr	prai	bar
aer	portu	rap	resort	ped
58	**59**	**60 (1)**	**61**	**62**
graf	film	radi	internet	softw
livr	cinem	tv	telecom	tecnolog
edit	cin	fm	veloc	soluco
assin	audiovis	programaca	conexa	gesta
jorn	vide	polic	fibr	dad
colun	audi	notic	assin	erp
impress	estudi	atr	meg	sistem
impressa	produca	esport	rote	ti
revist	music	classificaca	plan	certific
editor	voz	entreten	centr	digit
63	**66**	**68**	**69**	**70 (1)**
conteud	segur	imovel	contabil	projet
plataform	invest	shopping	contavel	financ
sistem	cambi	lote	advog	consult
diari	corre	empreend	escritori	gesta
mid	risc	condomini	contabel	english
cloud	carto	loj	direit	econom
clipping	fund	condomin	atuaca	reestruturaca
dat	credit	apart	are	consulting

Quadro 5.7: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por CNAE a 2 dígitos da PAS - Nó: PAS-CNAE2D

(Continuação)

63	66	68	69	70 (1)
marketing	saud	sind	fiscal	otimizaca
rpa	carta	administraca	audit	inn
71	73	74	77	78 (1)
engenh	pesquis	fotograf	locaca	empreg
ensai	agenc	traduca	aluguel	vag
laboratori	comunicaca	fot	equip	trad
projet	campanh	mergulh	loc	recrut
topograf	mid	produt	frot	temporari
inspeca	marc	cas	andaim	rh
obr	portfoli	ambient	maquin	talent
analis	painel	album	empilh	merchandising
industr	ide	design	franque	candidat
sond	propagand	cortin	acessori	seleca
79	80	81	82	85
viag	seguranc	limp	event	curs
tur	vigilanc	serv	cobranc	alun
passei	monitor	prag	inventari	escol
cruz	alarm	esterilizaca	relacion	cnh
intercambi	veicul	paisag	leila	ingl
pacot	patrimon	jardim	benefici	ensin
destin	roub	zelad	ades	habilitaca
tour	eletron	profiss	credit	nataca
viaj	rastre	terceirizaca	recuperaca	autoescol
event	risc	terceir	contact	trein
90 (2)	93	95 (1)	96	
show	marin	tecn	funer	
music	academ	produt	lavand	
event	parqu	assistenc	velori	
ilustraca	fitnes	inf	lav	
dat	musculaca	suport	cemiteri	
band	club	fabric	higienizaca	
rock	unidad	consert	enxov	
viol	fisic	pec	depilaca	
playlist	ginas	ar	plan	
caqu	fit	tecnolog	bel	

Fonte: Autoria própria.

(1) Nessas classes de 1 a 3 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas ou marcas (ou por haver essa possibilidade). (2) Nessa classe 4 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas, marcas ou artistas (ou por haver essa possibilidade)

Aparentemente, os *tokens* exibidos no Quadro 5.7 remetem às Divisões CNAE, apesar de mais uma vez ser difícil compreendê-los dada a utilização do *stemmer*. É possível destacar a Divisão 90 (Atividades artísticas, criativas e de espetáculos), que assim como o Grupo 46.1 (Representantes comerciais e agentes do comércio, exceto de veículos automotores e motocicletas - no nó *PAC46-CNAE3D*), teve uma maior quantidade de *tokens* omitidos. Na CNAE 90 foram ocultados nomes, aparentemente, de artistas. Nota-se também a existência de algumas siglas como, por exemplo, *erp*[5] na CNAE 62 (Atividades dos serviços de tecnologia da informação), *cnh*[6] na CNAE 85 (Educação) e *rpa*[7] na CNAE 63 (Atividades de prestação de serviços de informação). Além disso, também pode-se destacar a presença do nome *javascript*, uma linguagem de programação, que dificilmente está associada à CNAE 38 (Coleta, tratamento e disposição de resíduos; recuperação de materiais). A presença desse *token* pode estar relacionada à estrutura das páginas da *web* coletadas.

A seguir, podem ser vistos os resultados obtidos através dos classificadores planos.

5.1.2 Resultados Obtidos com os Classificadores Planos

Conforme explicitado na Seção 4.5, foram testados e otimizados 4 algoritmos, Naive Bayes, Regressão Logística, Métodos de Suporte Vetorial e Boosting, objetivando prever as 78 categorias finais.

De acordo com os resultados obtidos, o uso da *SMOTE* gerou melhora quando Naive Bayes foi aplicado, entretanto no uso da Regressão Logística o aumento do número foi muito ligeiro. No caso desses dois modelos, o *stemmer* não trouxe ganhos. Já nos Métodos de Suporte Vetorial, a maior ocorrência foi do *kernel* linear e o seu melhor resultado foi obtido com o *stemmer* e sem aplicação da *SMOTE*. Por fim, o Boosting, diferente dos demais, apresentou um melhor desempenho quando a representação dos *tokens* ocorreu via frequência do termo e não TF-IDF.

Os melhores resultados obtidos, a partir da base de treino, para cada algoritmo podem ser vistos na Tabela 5.1.

Tabela 5.1: Melhores resultados obtidos, através dos classificadores planos, considerando a métrica *Macro* F_1 *score*, a partir da base de treino, segundo os algoritmos testados

Algoritmos	$Macro\,F_1\,score\,(\%)$
RL	59,90
MSV	58,64
NB	56,97
Boosting	55,54

Fonte: Autoria própria.

[5] Com base na análise de alguns textos, acredita-se que a sigla *ERP*, no contexto da CNAE 62, signifique *Enterprise Resource Planning* (Planejamento de Recursos Empresariais).

[6] Com base na análise de alguns textos, acredita-se que a sigla CNH, no contexto da CNAE 85, signifique Carteira Nacional de Habilitação.

[7] Com base na análise de alguns textos, acredita-se que a sigla *RPA*, no contexto da CNAE 63, signifique *Robotic Process Automation* (Automação Robótica de Processos).

A Tabela 5.1 mostra que, mais uma vez, a Regressão Logística apresentou um bom desempenho, tendo sido seguida pelos Métodos de Suporte Vetorial. Já o Boosting exibiu o pior resultado, apesar de ter sido o modelo que requereu o maior tempo de treino. Todas as observações feitas a respeito da Regressão Logística na Subseção 5.1.1 são também aqui aplicáveis.

Visando o ajuste do modelo, ao se considerar a abordagem plana, foi utilizada a *SMOTE* e o *stemmer* não foi aplicado. Somado a isso, considerou-se a representação via TF-IDF. Os 10 *tokens* com maiores coeficientes, por classes finais, no modelo de Regressão Logística ajustado podem ser vistos no Quadro 5.8.

Quadro 5.8: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por classe final - Classificação Plana

(Continua)

01 (1)	08	10	11
semen	mineracao	receitas	agua
aviacao	calcario	alimentos	cachaca
plantio	pedras	sabor	bebidas
agricola	pedreira	produtos	refrigerantes
lavanda	brita	producao	mineral
ingrediente	pedra	sorvetes	cervejaria
laranjeiras	areia	frigorifico	guarana
pulverizacao	extracao	arroz	vinhos
clonagem	gesso	beef	cerveja
mudas	iodo	animal	cervejas
13 (1)	**14**	**15**	**16**
tecidos	confeccoes	calcados	madeira
textil	uniformes	couro	portas
fios	meias	couros	compensados
texteis	lingerie	bolsas	paletes
poliester	colecao	bota	eucalipto
elasticos	marca	pastas	molduras
tapetes	moda	pares	divisorias
cordas	camiseta	calcado	wood
tecelagem	bones	pasta	pallets
tingimento	confeccao	rasteira	pinus
17	**18 (1)**	**19**	**20**
papel	impressao	lubrificantes	fertilizantes
papelao	cores	alcool	produtos
embalagens	cartoes	etanol	quimica
etiquetas	rotulos	acucar	cosmeticos
papeis	grafica	oleos	aromas
incontinencia	grafico	safra	tintas
produtos	laminacao	oleo	industria
fraldas	portfolio	usina	arla
ondulado	gravacao	bioenergia	gel

Quadro 5.8: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por classe final - Classificação Plana

(Continuação)

17	18 (1)	19	20
caixas	convite	biodiesel	compostos
21	**22**	**23**	**24**
farmaceutica	borracha	concreto	fundicao
medicamentos	plasticas	vidros	aluminio
comprimidos	pvc	ceramica	ligas
animal	pead	cimento	perfis
quimica	injecao	vidro	ferro
farmacovigilancia	plastico	postes	tubos
producao	plasticos	cal	fitas
diagnostico	plast	abrasivos	steel
bula	pneus	obras	mola
suplemento	plas	porcelana	trefilacao
25	**26 (1)**	**27**	**28**
metalicas	medicao	eletricos	equipamentos
metalicos	produtos	transformadores	maquinas
aco	automacao	baterias	transportadores
moldes	temperatura	tensao	engrenagens
esquadrias	suporte	luminarias	refrigeracao
caldeiras	relogio	fogoes	assistencia
cnc	eletronicos	resistencias	ind
tanques	antenas	produtos	carreta
ferramentas	eletronica	energia	cilindros
rebites	amplificadores	pdf	turbinas
29 (1)	**30 (1)**	**31 (1)**	**32**
retifica	estaleiro	moveis	brinquedos
carrocerias	bicicletas	estofados	produtos
componentes	carrinhos	mobiliario	bolas
automotivo	rodas	representantes	joias
automotivas	embarcacoes	colchao	escovas
veiculos	guidao	ambientes	instrumentais
onibus	naval	colchoes	botoes
ind	bike	lojista	ref
iatf	barcos	divisorias	brincos
juntas	mtb	projeto	implantes
33	**38 (1)**	**45_{pas} (1)**	**45.1 (1)**
manutencao	residuos	mecanica	seminovos
industrial	reciclagem	pintura	veiculos
montagem	lixo	reparacao	concessionaria
fabricacao	cacambas	funilaria	caminhoes
opcoes	coleta	seguradoras	agendamento
cesto	urbana	servicos	vendas

Quadro 5.8: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por classe final - Classificação Plana

(Continuação)

33	38 (1)	45_{pas} (1)	45.1 (1)
industriais	dominio	veiculo	consorcio
containers	tratamento	automotiva	novos
reparo	entulho	oficina	carro
medidor	javascript	correa	drive
45.3 (1)	45.4 (2)	46.1 (2)	46.2
autopecas	motos	representacoes	sementes
auto	moto	drogarias	cafes
pneus	capacete	rep	pet
pecas	motocicletas	co	soja
aro	abs	corretora	insumos
freio	oficina	representadas	nutricao
pneu	pecas	representacao	cosmetica
solas	seminovas	comercio	genetica
automotivo	motopecas	sobreposicao	cooperativa
automotivos	interesse	transmissores	flores
46.3 (1)	46.4	46.5 (2)	46.6 (1)
alimentos	distribuicao	informatica	maquinas
distribuidora	distribuidora	telecom	correias
azeite	importacao	conector	pecas
feijao	hospitalares	impressoras	codificacao
food	produtos	fabricantes	tratores
panificacao	estampados	suprimentos	reposicao
ceasa	rating	distribuicao	vacuo
banana	medicamentos	sku	solar
cereais	med	solucoes	valvulas
distribuicao	cosmeticos	incendio	laser
46.7	46.8	46.9 (1)	47.1
materiais	diesel	importacao	supermercados
ferragens	sucatas	atacadista	ofertas
ferramentas	combustiveis	venda	supermercado
parafuso	acido	peixes	lojas
produtos	importacao	caixa	acougue
broca	defensivos	natal	loja
led	distribuicao	cooperativa	receitas
civil	agricolas	cesta	contratada
rosca	parceiros	details	emporio
estoque	glp	marcas	atacado
47.2	47.3 (1)	47.4 (1)	47.5 (1)
chocolate	posto	construcao	lojas
salgados	postos	loja	tudo
cafe	rede	tintas	moveis

Quadro 5.8: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por classe final - Classificação Plana

(Continuação)

47.2	47.3 (1)	47.4 (1)	47.5 (1)
frutas	combustiveis	materiais	residencial
wine	troca	porcelanato	utilidades
carnes	gas	obra	toner
paes	osto	iluminacao	impressoras
cha	oleo	madeiras	decoracao
gelo	gasolina	pisos	unidade
loja	apontador	ferramentas	persianas
47.6	**47.7**	**47.8**	**49**
livraria	manipulacao	lojas	transporte
livros	farmacia	moda	transportes
biblia	oculos	garden	frota
livro	otica	incendio	viacao
esportes	drogaria	flores	onibus
papelaria	lentes	ouro	horarios
brinquedos	aparelhos	pet	logisticas
orcar	saude	relogio	fretamento
vista	auditivos	calcados	logistica
gift	aparelho	gas	mudanca
50	**51 (1)**	**52**	**53 (1)**
embarcacoes	aereo	estacionamentos	entregas
navegacao	aeronaves	logistica	encomendas
maritimo	voos	park	motoboy
passeios	taxi	terminal	correios
fluvial	aeronave	estacionamento	rapidas
apoio	fretamento	armazenagem	entrega
navios	hangaragem	rodovias	manuseio
transporte	voo	terminais	transporte
balsas	aerea	aeroporto	post
maritimas	aero	aduaneiros	rapidez
55	**56**	**58**	**59**
hotel	cardapio	jornal	filmes
motel	refeicoes	editorial	cinema
pousada	alimentacao	grafica	cine
suites	restaurante	livros	cinemas
suite	bar	editora	audiovisual
acomodacoes	buffet	raizes	vozes
hospedes	delivery	anuncie	audio
resort	restaurantes	prime	produtora
praia	gastronomia	planet	responder
luxo	pizza	conteudo	tv

Quadro 5.8: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por classe final - Classificação Plana

(Continuação)

60 (1)	61 (1)	62	63 (1)
radio	internet	solucoes	plataforma
tv	planos	software	conteudo
fm	telecom	gestao	sistema
programacao	velocidade	sistemas	horas
atras	assinante	erp	cloud
vivo	net	dados	diario
classificacao	conexao	certificado	clipping
ouca	fibra	tecnologia	data
noticias	telecomunicacoes	plataforma	rpa
buzios	cobertura	solucao	marketing
66	68	69	70 (1)
seguros	imoveis	contabilidade	consultoria
corretora	shopping	contabil	projetos
seguro	empreendimentos	advogados	gaia
cambio	imovel	escritorio	inn
fundos	apartamento	direito	consulting
credito	aluguel	contabeis	financeira
investimentos	condominios	atuacao	otimizacao
cartoes	administracao	advocacia	english
saude	administradora	auditoria	gestao
cartao	imobiliaria	false	lema
71	73	74 (1)	77
engenharia	comunicacao	traducao	locacao
ensaios	agencia	aguia	equipamentos
topografia	pesquisas	mergulho	aluguel
inspecao	pesquisa	albuns	andaimes
analises	midia	fotografia	locacoes
projetos	campanha	fotos	franquias
laboratorio	propaganda	usuario	franqueado
sondagem	marketing	design	adicionado
projeto	ideias	casamento	guindaste
areas	portfolio	foto	frotas
78 (1)	79	80	81
vagas	viagens	seguranca	limpeza
trade	turismo	vigilancia	esterilizacao
recrutamento	viagem	monitoramento	servicos
empregos	passeios	patrimonial	pragas
selecao	cruzeiros	veiculo	paisagismo
rh	intercambio	servicos	jardins
merchandising	hoteis	horas	zeladoria
talentos	destinos	porto	portaria
servicos	tour	roubo	terceirizacao

Quadro 5.8: Dez *tokens* com maiores coeficientes no modelo de Regressão Logística, por classe final - Classificação Plana

(Continuação)

78 (1)	79	80	81
temporarios	pacotes	risco	desinsetizacao
82	**85**	**90 (1)**	**93**
eventos	cursos	shows	marina
leilao	curso	eventos	academia
pack	ensino	rock	fitness
cobranca	escola	ilustracao	parque
inventario	cnh	viola	musculacao
recuperacao	professores	festival	treino
contact	ingles	banda	ginastica
banco	habilitacao	playlist	pier
despachante	treinamentos	musica	atracoes
rua	natacao	portfolio	clube
95 (1)	**96**		
condicionado	lavanderia		
cofres	funeraria		
clientes	higienizacao		
desk	plano		
speed	beleza		
meses	cemiterio		
tecnica	depilacao		
solucoes	velorio		
impressoras	tratamentos		
help	cremacao		

Fonte: Autoria própria.

(1) Nessas classes de 1 a 3 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas, marcas ou artistas (ou por haver essa possibilidade).

(2) Nessas classes de 5 a 6 *tokens* foram substituídos pelos seguintes por se tratarem de nomes próprios relacionados a empresas ou marcas (ou por haver essa possibilidade).

Ao que tudo indica, a maior parte dos *tokens* exibidos no Quadro 5.8 remetem às características de cada uma das classes, sendo muitos deles já exibidos quando da utilização de classificadores locais. Chama atenção nessa segunda abordagem o aparecimento em um contexto mais desagregado de termos antes evidentes, na primeira abordagem, em classes mais agregadas. Pode ser citado como exemplo o *token produtos*, exibido na PIA-Empresa e na PAC, no Quadro 5.1, e na Divisão 46 (Comércio por atacado, exceto veículos automotores e motocicletas), no Quadro 5.2, e nas CNAEs 10 (Fabricação de produtos alimentícios), 17 (Fabricação de celulose, papel e produtos de papel), 26 (Fabricação de equipamentos de informática, produtos eletrônicos e ópticos), 27 (Fabricação de máquinas, aparelhos e materiais elétricos), 32 (Fabricação de produtos diversos), 46.7 (Comércio atacadista de madeira, ferragens, ferramentas, material elétrico e material de construção) e 46.4 (Comércio atacadista de produtos de consumo não-alimentar) no Quadro 5.8. Entretanto,

esse mesmo termo que estava na CNAE 45.3 (Comércio de peças e acessórios para veículos automotores), no Quadro 5.3, e, no Quadro 5.8, não se encontra mais nesse grupo. Já no Quadro 5.6, ele estava presente na Divisão 20 (Fabricação de produtos químicos) e permaneceu nessa mesma classe no Quadro 5.8. Outras ocorrências semelhantes a essa também podem ser pontuadas, como é o caso de *lojas*, exibido na PAC, no Quadro 5.1, na Divisão 47 (Comércio varejista), no Quadro 5.2, e nos Grupos 47.1 (Comércio varejista não-especializado), 47.5 (Comércio varejista de equipamentos de informática e comunicação; equipamentos e artigos de uso doméstico) e 47.8 (Comércio varejista de produtos novos não especificados anteriormente e de produtos usados), no Quadro 5.8. Não tendo sido o *stemming* do termo *lojas*, *loj*, verificado no Quadro 5.5.

O Quadro 5.8 exibe também algumas siglas não elencadas quando da utilização dos classificadores locais, a saber, *cnc*[8] na CNAE 25 (Fabricação de produtos de metal, exceto máquinas e equipamentos), *iatf*[9] na CNAE 29 (Fabricação de veículos automotores, reboques e carrocerias) e *sku*[10] na CNAE 46.5 (Comércio atacadista de equipamentos e produtos de tecnologias de informação e comunicação). Além disso, pode-se ressaltar a presença do *token false* na CNAE 69 (Atividades jurídicas, de contabilidade e de auditoria). Após a análise de alguns textos é possível notar que, assim como *javascript* citado quando feita a análise do Quadro 5.7, ele tem como origem a estrutura *HTML* das páginas da *web* e não o texto contido nas mesmas.

Ainda com relação ao Quadro 5.8 é possível destacar, na Divisão 81 (Serviços para edifícios e atividades paisagísticas), o *token desinsetizacao*. De acordo com Pacheco (2021), a palavra *dedetização* apareceu, durante a Segunda Guerra Mundial, quando a substância Dicloro-difenil-tricloretano (popularmente chamada de DDT) passou a ser utilizada contra insetos. Entretanto, por causa da sua toxicidade, o DDT foi proibido. Devido a esse fato e a outras formas de combate terem surgido, hoje em dia, outros nomes são atribuídos à prática, como *desinsetização*. Na base de dados de treino aqui utilizada *dedetização* apareceu em 41 documentos e *desinsetização* em 36. Já em IBGE e CONCLA (2020), apenas a palavra *dedetização* pode ser encontrada associada à Classe 81.22-2 (Imunização e controle de pragas urbanas), todavia em IBGE e CONCLA (2022) (Sistema de busca online da CNAE) as duas palavras, *dedetização* e *desinsetização*, quando pesquisadas, levam o usuário à Classe 81.22-2. Essa é mais uma ilustração da dinamicidade da língua.

Na sequência, são detalhados os resultados obtidos na base de teste com os classificadores locais e o plano.

5.2 Resultados Obtidos na Base de Teste

Os modelos foram aplicados à base de teste possibilitando a visualização dos seus desempenhos em dados nos quais não foram treinados. Os resultados considerando as

[8] Com base na análise de alguns textos, acredita-se que a sigla CNC, no contexto da CNAE 25, signifique Controle Numérico Computadorizado.

[9] Com base na análise de alguns textos, acredita-se que a sigla *IATF*, no contexto da CNAE 29, signifique *International Automotive Task Force* (Força Tarefa Automotiva Internacional).

[10] Com base na análise de alguns textos, acredita-se que a sigla *SKU*, no contexto da CNAE 46.5, signifique *Stock Keeping Unit* (Unidade de Manutenção de Estoque).

abordagens plana e local e as métricas planas (tradicionais) e hierárquicas podem ser vistos na Tabela 5.2.

Tabela 5.2: Resultados obtidos na base de teste com os melhores algoritmos e configurações encontradas na base de treino, por tipo de métrica, segundo as abordagens testadas

		Métricas planas (macro)			Métricas hierárquicas		
Classificador	Abordagem	Precision	Recall	F_1	Precision	Recall	F_1
RL	Plana	0,6110	0,6083	0,6031	0,7457	0,7360	0,7408
RL	Local	0,6120	0,6014	0,6005	0,7485	0,7448	0,7466

Fonte: Autoria própria.

Através da Tabela 5.2 é possível verificar que as duas abordagens apresentaram um desempenho muito semelhante quando ambas as métricas, planas e hierárquicas, foram consideradas. No caso das métricas planas, a dois dígitos, os resultados foram idênticos. Já no caso das métricas hierárquicas, a abordagem local, com um modelo por nó, se mostrou ligeiramente superior. Além disso, o classificador plano apresentou acurácia de 0,6590 e os locais de 0,6570.

Apesar dos resultados semelhantes exibidos na Tabela 5.2, ao considerar as métricas planas, é possível observar diferenças nos desempenhos dos classificadores por classe. Essas informações foram disponibilizadas na Tabela 5.3. Nessa tabela o primeiro conjunto de linhas brancas engloba as classes pertencentes a PIA-Empresa, as linhas cinzas as classes pertencentes a PAC e o último conjunto de linhas brancas as classes que compõem a PAS.

Tabela 5.3: Resultados obtidos na base de teste considerando as métricas planas, por tipo de classificador, segundo as classes finais

(Continua)

	Classificador Plano			Classificadores Locais			
Classe	Precision	Recall	F_1	Precision	Recall	F_1	Nº Observações
08	0,7037	0,7600	0,7308	0,7600	0,7600	0,7600	25
10	0,7578	0,7860	0,7717	0,7358	0,8419	0,7852	215
11	0,7059	0,7742	0,7385	0,8148	0,7097	0,7586	31
13	0,5854	0,7059	0,6400	0,6667	0,6765	0,6715	68
14	0,5875	0,6351	0,6104	0,6308	0,5541	0,5899	74
15	0,6750	0,6923	0,6835	0,6667	0,6154	0,6400	39
16	0,6970	0,7931	0,7419	0,6765	0,7931	0,7302	29
17	0,6458	0,6078	0,6263	0,6471	0,6471	0,6471	51
18	0,4878	0,5263	0,5063	0,5000	0,4737	0,4865	38
19	0,6071	0,7727	0,6800	0,7273	0,7273	0,7273	22
20	0,6791	0,6947	0,6868	0,6769	0,6718	0,6743	131
21	0,5946	0,7857	0,6769	0,7308	0,6786	0,7037	28
22	0,6846	0,7391	0,7108	0,6757	0,7246	0,6993	138
23	0,6463	0,6463	0,6463	0,6512	0,6829	0,6667	82
24	0,5319	0,5556	0,5435	0,6000	0,5333	0,5647	45
25	0,5482	0,5759	0,5617	0,5640	0,6139	0,5879	158

Tabela 5.3: Resultados obtidos na base de teste considerando as métricas planas, por tipo de classificador, segundo as classes finais

(Continuação)

	Classificador Plano			Classificadores Locais			
Classe	Precision	Recall	F_1	Precision	Recall	F_1	Nº Observações
26	0,4091	0,4500	0,4286	0,4355	0,4500	0,4426	60
27	0,5974	0,5897	0,5935	0,6479	0,5897	0,6174	78
28	0,6310	0,6629	0,6466	0,6175	0,6348	0,6260	178
29	0,4722	0,5152	0,4928	0,5217	0,5455	0,5333	66
30	0,6154	0,3636	0,4571	0,5000	0,2273	0,3125	22
31	0,6900	0,7753	0,7302	0,7423	0,8090	0,7742	89
32	0,3804	0,4605	0,4167	0,4691	0,5000	0,4841	76
33	0,3043	0,3500	0,3256	0,2812	0,2250	0,2500	40
45.1	0,8690	0,9130	0,8905	0,8707	0,9275	0,8982	138
45.3	0,6951	0,6786	0,6867	0,6627	0,6548	0,6587	84
45.4	0,9310	0,8438	0,8852	0,8333	0,7812	0,8065	32
46.1	0,7500	0,3333	0,4615	0,5000	0,2222	0,3077	9
46.2	0,3654	0,6129	0,4578	0,4348	0,6452	0,5195	31
46.3	0,5395	0,5775	0,5578	0,5574	0,4789	0,5152	71
46.4	0,4327	0,3435	0,3830	0,4576	0,4122	0,4337	131
46.5	0,2941	0,2381	0,2632	0,2500	0,2857	0,2667	21
46.6	0,5373	0,4286	0,4768	0,4353	0,4405	0,4379	84
46.7	0,2791	0,2182	0,2449	0,3830	0,3273	0,3529	55
46.8	0,6234	0,4486	0,5217	0,5385	0,4579	0,4949	107
46.9	0,2000	0,0769	0,1111	0,1667	0,0769	0,1053	26
47.1	0,6418	0,7544	0,6935	0,6615	0,7544	0,7049	57
47.2	0,3846	0,4839	0,4286	0,4571	0,5161	0,4848	31
47.3	0,8571	0,7692	0,8108	0,8333	0,6410	0,7246	39
47.4	0,6435	0,5736	0,6066	0,6607	0,5736	0,6141	129
47.5	0,3281	0,2530	0,2857	0,3404	0,3855	0,3616	83
47.6	0,5455	0,5000	0,5217	0,5556	0,4167	0,4762	24
47.7	0,8070	0,6866	0,7419	0,8000	0,7164	0,7559	67
47.8	0,4891	0,4500	0,4688	0,4167	0,5500	0,4741	100
01	0,5714	0,4444	0,5000	0,7143	0,5556	0,6250	9
38	0,5667	0,6800	0,6182	0,5926	0,6400	0,6154	25
45_{pas}	0,6538	0,4857	0,5574	0,5455	0,3429	0,4211	35
49	0,8854	0,8615	0,8733	0,8926	0,8308	0,8606	260
50	0,6923	0,6923	0,6923	0,7333	0,8462	0,7857	13
51	0,6250	0,6250	0,6250	0,6250	0,6250	0,6250	8
52	0,6818	0,6667	0,6742	0,6932	0,6778	0,6854	90
53	0,6842	0,8125	0,7429	0,6087	0,8750	0,7179	16
55	0,9348	0,9416	0,9382	0,9338	0,9270	0,9304	137
56	0,8400	0,7730	0,8051	0,7730	0,7730	0,7730	163
58	0,6207	0,5806	0,6000	0,6000	0,5806	0,5902	31
59	0,7619	0,5926	0,6667	0,7500	0,5556	0,6383	27
60	0,7436	0,9355	0,8286	0,7105	0,8710	0,7826	31

Tabela 5.3: Resultados obtidos na base de teste considerando as métricas planas, por tipo de classificador, segundo as classes finais

(Continuação)

	Classificador Plano			Classificadores Locais			
Classe	Precision	Recall	F_1	Precision	Recall	F_1	Nº Observações
61	0,7705	0,7966	0,7833	0,7895	0,7627	0,7759	59
62	0,6438	0,6645	0,6540	0,6352	0,6516	0,6433	155
63	0,1042	0,1351	0,1176	0,1304	0,1622	0,1446	37
66	0,7692	0,8824	0,8219	0,7595	0,8824	0,8163	68
68	0,7397	0,8060	0,7714	0,7105	0,8060	0,7552	67
69	0,9286	0,9681	0,9479	0,9377	0,9610	0,9492	282
70	0,2667	0,1538	0,1951	0,2941	0,1923	0,2326	26
71	0,6939	0,7556	0,7234	0,6667	0,7111	0,6882	90
73	0,5319	0,6410	0,5814	0,5745	0,6923	0,6279	39
74	0,4583	0,4074	0,4314	0,3333	0,4074	0,3667	27
77	0,7544	0,5584	0,6418	0,7018	0,5195	0,5970	77
78	0,6486	0,6667	0,6575	0,6389	0,6389	0,6389	36
79	0,8276	0,8276	0,8276	0,7742	0,8276	0,8000	29
80	0,6545	0,8000	0,7200	0,6111	0,7333	0,6667	45
81	0,7093	0,7011	0,7052	0,6941	0,6782	0,6860	87
82	0,3924	0,3263	0,3563	0,3803	0,2842	0,3253	95
85	0,7231	0,8246	0,7705	0,7231	0,8246	0,7705	57
90	0,5000	0,1429	0,2222	0,5000	0,1429	0,2222	7
93	0,7500	0,9429	0,8354	0,7857	0,9429	0,8571	35
95	0,1818	0,1579	0,1690	0,2353	0,2105	0,2222	38
96	0,9032	0,8000	0,8485	0,9355	0,8286	0,8788	35

Fonte: Autoria própria.

Nota: O primeiro conjunto de linhas brancas engloba as classes pertencentes a PIA-Empresa, as linhas cinzas as classes pertencentes a PAC e o último conjunto de linhas brancas as classes que compõem a PAS.

Como pode ser visto, ambos os classificadores apresentaram dificuldades no reconhecimento do Grupo 46.9 (Comércio atacadista não-especializado), o que parece razoável já que a classe engloba basicamente tudo que não pode ser enquadrado em outras CNAEs da Divisão 46 (Comércio por atacado, exceto veículos automotores e motocicletas). Os classificadores também não foram bem-sucedidos nas Divisões 63 (Atividades de prestação de serviços de informação), 70 (Atividades de sedes de empresas e de consultoria em gestão empresarial), 90 (Atividades artísticas, criativas e de espetáculos) e 95 (Reparação e manutenção de equipamentos de informática e comunicação e de objetos pessoais e domésticos). Nessas classes a métrica F_1 obtida foi inferior a 0,25. No extremo oposto, com a métrica F_1 superior a 0,85 para ambos os classificadores, estão o Grupo 45.1 (Comércio de veículos automotores) e as Divisões 49 (Transporte terrestre), 55 (Alojamento) e 69 (Atividades jurídicas, de contabilidade e de auditoria).

Em outras categorias, o classificador plano e os locais exibiram um comportamento muito distinto. Pode-se destacar as classes 01 (Agricultura, pecuária e serviços relacionados)

e 46.7 (Comércio atacadista de madeira, ferragens, ferramentas, material elétrico e material de construção) como as que, de acordo com a métrica F_1, apresentaram o melhor resultado com o uso dos classificadores locais. Já as classes 30 (Fabricação de outros equipamentos de transporte, exceto veículos automotores), 33 (Manutenção, reparação e instalação de máquinas e equipamentos), 45_{pas} e 46.1 (Representantes comerciais e agentes do comércio, exceto de veículos automotores e motocicletas) exibiram uma métrica F_1 mais alta quando o classificador plano foi utilizado. Já para as Divisões 85 (Educação), 90 (Atividades artísticas, criativas e de espetáculos) e 51 (Transporte aéreo) o desempenho foi idêntico com as duas abordagens. Aparentemente, o classificador plano foi mais eficiente em classes que tendiam a ser mais difíceis de caracterizar nos nós superiores. Dois exemplos dessas podem ser vistos na Tabela 5.4.

Tabela 5.4: Empresas, de classes que evidenciaram a dificuldade encontrada pelos classificadores locais, classificadas incorretamente

Classe em 2019	Classe Predita						Total de empresas na classe
	Classificador Plano			Classificadores Locais			
	PIA-Emp	PAC	PAS	PIA-Emp	PAC	PAS	
33	14	2	10	11	4	16	40
45_{pas}	4	10	4	1	18	4	35

Fonte: Autoria própria.

A Divisão 33 (Manutenção, reparação e instalação de máquinas e equipamentos) engloba atividades de manutenção, reparação e instalação e pertence a PIA-Empresa[11], enquanto outras CNAEs também referentes à manutenção fazem parte da PAS, sendo esse o caso da Classe 62.09-2 (Suporte técnico, manutenção e outros serviços em tecnologia da informação), da Divisão 95 (Reparação e manutenção de equipamentos de informática e comunicação e de objetos pessoais e domésticos), do Grupo 45.2 (Manutenção e reparação de veículos automotores) e da Classe 45.43-9 (Manutenção e reparação de motocicletas). Já em relação a essas duas últimas, elas compõem a classe 45_{pas} e se misturam a classificações da PAC a medida que compartilham a Divisão 45 (Comércio e reparação de veículos automotores e motocicletas) e, até mesmo, dividem o mesmo Grupo 45.4 (Comércio, manutenção e reparação de motocicletas, peças e acessórios), como detalhado na Figura 4.1. A partir da Tabela 5.4 é possível notar que os classificadores locais classificaram mais empresas da Divisão 33 de forma inadequada, sendo na maioria das vezes atribuída uma CNAE da PAS. Enquanto isso, o classificador plano, por mais que também tenha cometido erros para uma quantidade considerável de empresas, tendeu a acertar o fato delas pertencerem à PIA-Empresa. O caso da 45_{pas} é parecido. Os classificadores locais, ao cometerem um erro, atribuíram mais CNAEs da PAC às empresas do que o classificador plano.

Durante a análise das companhias pertencentes à Divisão 33, classificadas de forma equivocada, foi possível notar que uma delas teve como classe predita pelos dois classificadores a Divisão 69 (Atividades jurídicas, de contabilidade e de auditoria). Após uma

[11] Conforme mencionado na Subseção 5.1.1 ao tratar do nó *PIA-CNAE2D*, a Divisão 33 compreende as atividades de manutenção, reparação e instalação de máquinas e equipamentos utilizados no processo de produção industrial, realizadas por unidades especializadas, normalmente sob contrato.

averiguação mais minuciosa, constatou-se que o *site* utilizado na coleta era referente a uma empresa que prestava serviços contábeis e não à selecionada pelo IBGE. Visto isso, foram examinadas todas as companhias com classe predita erroneamente como 69 por, ao menos, um dos dois classificadores.

A busca resultou em 21 empresas enquadradas no perfil exposto. Dessas, 3 foram classificadas incorretamente apenas pelo classificador plano e 1 apenas pelos locais. Já ambos classificaram na Divisão 69, inadequadamente, 17 companhias. Após a análise manual desse último grupo, foi constatado que 6 exibiram o *site* de uma firma de contabilidade e não da empresa selecionada. Assim, pode-se entender que apesar da retirada durante o pré-processamento, detalhada na Seção 3.3, de 227 companhias que indicaram possuir essa incongruência no preenchimento do questionário, aparentemente, ter funcionado para a maior parte dos casos, a limpeza ainda não foi suficiente.

Na sequência, a Tabela 5.5 mostra, por classificador, as classes que apresentaram 5 ou mais empresas classificadas incorretamente e que, dentre essas, mais de 40% se concentravam em uma única classe. Para a realização desta análise foram retiradas as 6 companhias, citadas anteriormente, com classificação predita 69. Os valores que constam entre parênteses na coluna "Percentual de Empresas (%)" são os números absolutos de empresas classificadas erroneamente na classe predita exibida. Já os valores à esquerda dos números entre parênteses são os percentuais que esses últimos representam do total de erros para determinada classe real.

Tabela 5.5: Classes que tiveram mais de 40% das empresas preditas inadequadamente em uma classe específica, considerando apenas as classes que apresentaram 5 ou mais companhias classificadas erroneamente

Classificador Plano		
Classe em 2019	Classe Predita	Percentual de Empresas (%)
66_{pas}	82_{pas}	62,50 (5)
79_{pas}	49_{pas}	60,00 (3)
46.3_{pac}	10_{pia}	53,33 (16)
78_{pas}	81_{pas}	50,00 (6)
55_{pas}	56_{pas}	50,00 (4)
45_{pas}	45.3_{pac}	44,44 (8)
47.2_{pac}	10_{pia}	43,75 (7)
11_{pia}	10_{pia}	42,86 (3)

(a) Principais equívocos ao se considerar o classificador plano

Classificadores Locais		
Classe em 2019	Classe Predita	Percentual de Empresas (%)
66_{pas}	82_{pas}	62,50 (5)
79_{pas}	49_{pas}	60,00 (3)
47.2_{pac}	10_{pia}	60,00 (9)
78_{pas}	81_{pas}	53,85 (7)
14_{pas}	47.8_{pac}	51,52 (17)
46.3_{pac}	10_{pia}	51,35 (19)
55_{pas}	56_{pas}	50,00 (5)
45_{pas}	45.3_{pac}	43,48 (10)

(b) Principais equívocos ao se considerar os classificadores locais

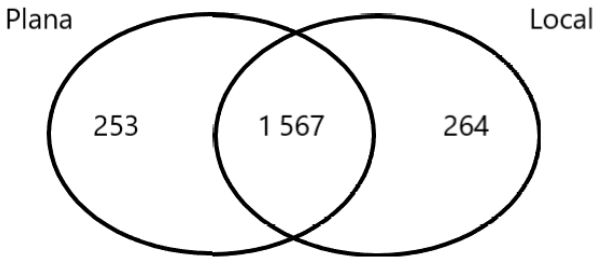
Fonte: Autoria própria.

Pode-se notar que os dois classificadores compartilharam a maior parte das combinações de classes em 2019 e classes preditas, sendo as únicas diferenças a classe 11 (Fabricação de bebidas), em 2019, com a classe predita 10 (Fabricação de produtos alimentícios) quando do uso do classificador plano e a classe 14 (Confecção de artigos do vestuário e acessórios), em

2019, com a classe predita 47.8 (Comércio varejista de produtos novos não especificados anteriormente e de produtos usados) quando os classificadores locais foram aplicados. Vale também destacar que metade das linhas apresentadas na Tabela 5.5 consistem em classes em 2019 e classes preditas pertencentes à PAS. Esse fato já era esperado, pois o papel do algoritmo é reconhecer padrões em dados previamente categorizados por humanos e, conforme explicitado na Subseção 3.1.1, a classificação em CNAEs da PAS é uma tarefa árdua para esses, assim é possível presumir que o reconhecimento não seja trivial também para as máquinas. Outro ponto a ser considerado, também abordado na Subseção 3.1.1, tem relação com a dinamicidade e a flexibilidade associadas aos serviços. Essas características podem ter feito com que as empresas classificadas com CNAEs da PAS tenham sentido de forma mais intensa os efeitos da defasagem temporal entre a variável resposta e as preditoras. Percebe-se, ainda, que mesmo quando a inadequação se refere também à pesquisa na qual a CNAE está alocada, as atividades não são tão distintas. Esse é o caso do Grupo 46.3 (Comércio atacadista especializado em produtos alimentícios, bebidas e fumo) e do Grupo 47.2 (Comércio varejista de produtos alimentícios, bebidas e fumo) em relação à Divisão 10 (Fabricação de produtos alimentícios). O mesmo também é válido para o Grupo 47.8 (Comércio varejista de produtos novos não especificados anteriormente e de produtos usados), que engloba artigos do vestuário e acessórios, calçados e artigos de viagem, joias e relógios, entre outros, e à Divisão 14 (Confecção de artigos do vestuário e acessórios).

Ainda objetivando entender melhor os resultados, foram comparadas, e exibidas no diagrama na Figura 5.2, as empresas classificadas incorretamente quando do uso de cada tipo de classificador.

Figura 5.2: Diagrama de Venn representando as empresas classificadas em classes finais incorretas de acordo com a abordagem plana e local



Fonte: Autoria própria.

De acordo com a Figura 5.2, 253 empresas foram classificadas de forma incorreta apenas quando aplicada a abordagem plana, tendo sido essas empresas classificadas corretamente através da abordagem local. O mesmo vale para as 264 empresas classificadas de maneira insatisfatória pela abordagem local. Já 1 567 companhias foram classificadas de forma incorreta pelos dois classificadores, não tendo sido, necessariamente, atribuídas às mesmas classes por ambos. As próximas análises realizadas tiveram como foco esse último grupo.

Objetivando entender se as empresas classificadas de forma errônea pelas duas abordagens, em algum momento, já apresentaram como atividade principal a fornecida por um

dos dois classificadores, foi analisado o histórico de CNAEs fornecido por essas companhias ao IBGE. As 6 empresas que informaram o *site* referente a uma firma de contabilidade foram retiradas da análise.

5.2.1 Investigação do Histórico das Empresas Classificadas Incorretamente

A análise contou com dados de classificações econômicas da PIA-Empresa, PAC e PAS, desde 2007 até o ano de 2018. A escolha do ano inicial foi pautada na implantação da versão 2.0 da CNAE.

Das 1 561 empresas investigadas, 168 não apresentaram nenhum histórico nessas pesquisas, ou seja, não possuíam nenhum questionário preenchido nos anos averiguados. Já 7 delas possuíam uma CNAE semelhante à atribuída pelos classificadores locais, e diferente da fornecida pelo classificador plano, em algum dos 12 anos, enquanto 11 empresas apresentaram uma CNAE igual à atribuída pelo classificador plano e diferente da fornecida pelos classificadores locais. Para finalizar, 75 companhias exibiram uma classificação no seu histórico similar às concedidas por ambos os classificadores. Dessas últimas, apenas uma contou com uma CNAE diferente atribuída por classificador, sendo ambas encontradas no seu histórico.

Aparentemente, para essas 93 empresas, um dos classificadores, ou ambos, detectaram um padrão de atividade econômica não tão distante do adequado, podendo se tratar de uma atividade secundária de destaque na empresa ainda atualmente ou, até mesmo, de uma atividade que voltou a ocupar o papel de principal nos anos de 2020 e 2021.

Apesar da análise de períodos anteriores funcionar bem, como mencionado na Subseção 3.1.1, os anos de 2020 e 2021 foram anos excepcionais, onde se viveu uma pandemia, e a análise histórica pode não ser tão adequada. Por conta disso, buscou-se uma comparação dos resultados da modelagem com a CNAE referente ao ano de 2020. Esse passo, da mesma forma que o anterior, foi executado na base com 1 561 empresas.

5.2.2 Empresas Classificadas Incorretamente e o seu Comportamento em 2020

Quando este trabalho começou a ser executado, a classificação econômica mais recente disponível para empresas nas Pesquisas Estruturais era referente ao ano de 2019. Entretanto, conforme explicitado na Seção 3.1, as informações são solicitadas, pelo IBGE, para as companhias no primeiro semestre de cada ano e são referentes ao ano anterior, sendo os dados publicados no ano seguinte ao do seu recebimento. Por conta disso, atualmente[12], os dados referentes ao ano de 2020 já foram coletados, mas ainda estão em processo de análise e compilação no Instituto.

Apesar dos dados ainda não estarem aptos para a divulgação, podendo apresentar inconsistências, a CNAE informada pela empresa em 2020 pode ser confrontada com a determinada pelos classificadores para as 1 561 companhias para as quais ambos foram

[12] Em novembro de 2021.

insatisfatórios. O objetivo dessa tarefa é ter ao menos uma noção dos erros oriundos da defasagem temporal entre a coleta dos dados textuais e a CNAE divulgada em 2019.

Das 1 561 empresas, 864 já tiveram os dados de 2020 coletados e possuem uma CNAE preenchida pelo informante. Dessas 864, 113 apresentaram uma classificação diferente da considerada no ano de 2019. Vale mencionar que a classificação fornecida pelo informante nem sempre é a mais adequada, sendo realizadas verificações manuais pelos técnicos do Instituto quando a CNAE informada difere da atribuída para a seleção da empresa, ponto esse detalhado na Subseção 3.1.1. As comparações de CNAEs, inclusive levando em conta a análise descrita na Subseção 5.2.1, constam na Tabela 5.6.

Tabela 5.6: Comparação das CNAEs do histórico e da classificação fornecida pelo informante em 2020 com o resultado obtido através de cada um dos classificadores - apenas para as empresas que foram classificadas de forma inadequada nas abordagens plana e local e tiveram uma CNAE informada em 2020 diferente da considerada em 2019

Classificação similar à de 2020?		Classificação existente no histórico?		Total de
Plano	Local	Plano	Local	Empresas
Não	Não	Não	Não	70
Não	Não	Sim	Sim	3
Não	Sim	Não	Não	3
Não	Sim	Não	Sim	1
Sim	Sim	Não	Não	24
Sim	Sim	Sim	Sim	8
Sim	Não	Não	Não	4

Fonte: Autoria própria.

Como pode ser visto, na primeira linha da Tabela 5.6 existem 70 empresas para as quais nenhum dos dois classificadores gerou uma CNAE similar à apresentada pelo informante em 2020 ou uma classificação exibida no histórico de respostas da companhia. Já a segunda linha evidencia que para 3 empresas os classificadores não obtiveram uma CNAE similar à enviada pelo informante em 2020, porém ambos chegaram a uma classificação já exibida pela companhia em anos anteriores. As linhas que seguem possuem uma interpretação semelhante. Logo, pode-se notar que a CNAE fornecida pelo classificador plano estaria correta para 36 empresas das 1 561 se fosse levada em consideração a CNAE preenchida pelo informante referente ao ano de 2020 e o mesmo vale para os classificadores locais, exibindo as duas abordagens 32 acertos em comum. Portanto, como já se esperava após as análises exibidas na Subseção 3.1.1, se não houvesse a defasagem temporal entre o dado textual e a classificação das empresas, provavelmente, os resultados do experimento seriam, ao menos, ligeiramente superiores.

Indo mais uma vez ao encontro das explanações realizadas na Subseção 3.1.1, vale também mencionar que das 40 empresas, cuja classificação predita por qualquer uma das abordagens foi igual à declarada pelo informante referente ao ano de 2020, 16 pertenciam a PAS, 13 a PAC e 11 a PIA-Empresa ao se considerar as classificações referentes ao ano de 2019. Assim, mais uma vez, fornece-se indícios de que a classificação de Serviços não é trivial e de que esta pode ter sido mais afetada pelo descompasso temporal.

Logo, apesar dos dois classificadores não terem fornecido uma classificação adequada para 1 561 empresas, ao que tudo indica, para 124 delas[13] houve a detecção de algum padrão útil.

[13] Noventa e três (93) companhias mencionadas na Subseção 5.2.1 e 31 constantes nas linhas da Tabela 5.6 referentes à classificação similar à de 2020 e divergente do histórico.

Capítulo 6

Conclusões

Este trabalho forneceu uma discussão a respeito da ciência de dados, abordando a sua história, principais conceitos, bem como facilidades e desafios criados por ela. Ele buscou ainda inseri-la no contexto dos órgãos oficiais de estatística, trazendo pontos e questionamentos oriundos de diversos grupos de trabalho organizados ao redor do mundo. Também descreveu o cenário do qual emergiu a tarefa de classificação a ser tratada, a atividade econômica principal de uma empresa, pontuando a sua função e importância. Ademais, devido aos dados disponíveis, tratou da estabilidade temporal do objeto a ser classificado visando garantir a viabilidade da execução da tarefa.

Desenvolveu-se durante o estudo uma forma automatizada de captura de informações textuais de *sites* de empresas que desempenhavam atividades distintas e de múltiplos portes e regiões brasileiras. Além disso, foi estudada a estrutura das *URLs* visando agregar informação aos textos recolhidos das páginas da *web*. Em todo processo, sempre que possível, foram utilizadas metodologias atualizadas e fornecidas descrições detalhadas dada a inviabilidade da disponibilização da base de dados criada devido ao sigilo estatístico, garantido por Lei, das informações prestadas ao IBGE.

A extração de dados executada neste trabalho foi um exercício importante, uma vez que não se encontrou relato de nenhuma outra tentativa que objetivasse a detecção da atividade econômica principal, considerando empresas tão diversas e a língua portuguesa "brasileira". Através dela foi possível verificar a viabilidade da coleta e também observar percalços como o fato da *URL* informada no questionário não pertencer à respondente, a existência de *URLs* referentes a páginas com textos em línguas diferentes do português e a dificuldade de extração de textos bem estruturados que favorecessem o uso de abordagens que levassem em conta o contexto no qual as palavras estavam inseridas, como o uso de *bigrams* e *word embeddings*. Apesar das dificuldades, mesmo sem um desempenho excelente, os diversos algoritmos testados (Naive Bayes, Regressão Logística, Métodos de Suporte Vetorial e Boosting) considerando uma representação do tipo saco de palavras se mostraram aptos para extrair padrões dos textos capazes de identificar a CNAE principal de uma empresa.

O estudo trouxe ainda a comparação do uso de um classificador plano e de vários locais. As duas abordagens apresentaram resultados agregados muito próximos, entretanto ao

se comparar classe a classe as diferenças foram expressivas, aparentando ter ocorrido um melhor desempenho do classificador plano em categorias que tendiam a ser mais difíceis de caracterizar nos nós superiores. Os diversos resultados obtidos através das aplicações locais, exibidas com mais detalhes no Apêndice B, foram úteis também para demonstrar os diferentes graus de dificuldade encontrados pelos algoritmos a depender do nó tratado.

Vale a pena ainda pontuar que este trabalho não pretendeu, de forma alguma, fornecer uma solução totalmente automatizada para a busca da classificação econômica da atividade principal de empresas. Entende-se que padrões não trivialmente estabelecidos por humanos dificilmente serão de fácil detecção pelas máquinas, principalmente, em um contexto de aprendizado supervisionado, uma vez que a ideia básica da concepção do último é reconhecer os exemplares a que foram previamente expostos. Ainda assim, haja vista o cenário onde ainda não se vislumbrou como trabalhar com classificações flexíveis fornecendo garantias, como transparência e comparabilidade, para o usuário, a CNAE tal como concebida hoje tem extrema relevância, sendo peça chave nos planos amostrais. Por isso, é importante que a classificação econômica seja sempre a mais atualizada possível, mesmo que estas mudanças não sejam instantaneamente aplicadas às pesquisas, para cumprir suas regras de estabilidade.

O estudo aqui desenvolvido pode ser muito útil como mecanismo de previsão de quais empresas devem ser, preferencialmente, analisadas pelos técnicos do Instituto, bem como os textos extraídos da *web* podem servir como fonte de informação adicional para o analista, já que existe a defasagem temporal entre a data que o questionário é analisado e o ano de referência da informação prestada pelo respondente. Além disso, a metodologia pode ser usada para a atribuição automática de uma CNAE às unidades cuja alteração de sua atividade não gerará um alto impacto nas estatísticas resultantes.

6.1 Limitações e Trabalhos Futuros

Este estudo teve como objetivo primário avaliar se seria possível obter a CNAE principal de uma empresa utilizando como variáveis preditoras o texto oriundo da sua página da *web* e o da própria *URL*. No entanto, devido à quantidade de *URLs* utilizadas e da sua origem nas Pesquisas Estruturais por Empresa, não foi possível mensurar o desempenho da metodologia em toda estrutura da CNAE, ficando o trabalho limitado a 78 classes.

A limitação citada pode encontrar solução em pesquisas futuras inspiradas nas desenvolvidas pelo Eurostat, no projeto ESSnet Big Data II[1]. O Grupo possui, entre diversas outras, uma linha de trabalho chamada de *Enterprise Characteristics*. O seu objetivo é usar *web scraping*, mineração de dados textuais e técnicas de inferência para coletar e processar dados de companhias, a fim de melhorar e atualizar as informações existentes. Entre os estudos desenvolvidos por eles, encontram-se métodos que podem ser aplicados para corrigir inconsistências em *URLs* existentes em cadastros de empresas e também

[1] ESSnet Big Data II é um projeto desenvolvido pelo Sistema Estatístico Europeu (*European Statistical System - ESS*) que tem como objetivo a integração de megadados à produção regular de estatísticas oficiais. Metodologias, resultados e metadados disponibilizados pelos diversos países participantes da iniciativa podem ser encontrados em Eurostat (2022b).

para a obtenção da *URL* via mecanismos de busca na *web*, podendo esse último ponto promover a possibilidade de se avaliar mais categorias de CNAEs. Uma outra maneira de aumentar a base disponível seria o estabelecimento de acordos com outras Instituições, o que possivelmente geraria um trabalho de integração de dados. Ou ainda, usar as redes sociais como fonte de dados, pois, conforme mencionado na Subseção 3.1.2, algumas companhias já preenchem nos questionários endereços referentes, por exemplo, a seu Instagram e Facebook no campo em que o *site* é solicitado. Outra solução poderia ser trabalhar também com os textos coletados em língua inglesa.

Ainda com relação ao trabalho desenvolvido pela ESSnet Big Data II, os métodos que visam promover a correção de inconsistências em *URLs* existentes em cadastros de empresas poderiam auxiliar em uma outra limitação deste estudo, a relacionada às *URLs* referentes a empresas de contabilidade e não as selecionadas pelas pesquisas. A construção, ainda que apenas de regras determinísticas que compreendessem a busca dos nomes fantasia, endereços e outros dados cadastrais das empresas nos textos obtidos das suas páginas da *web*, poderia garantir a consistência dos dados sem que fosse necessário averiguar a classificação final da companhia, como executado aqui. Além disso, provavelmente, outros tipos de erros, que não os vislumbrados neste trabalho[2], poderiam ser detectados e os algoritmos apresentariam melhor desempenho. Apesar dessa parecer uma tarefa simples, esse tipo de busca, principalmente se executada de forma mais flexível, através de expressões regulares (*regex*) em grandes conjuntos de dados textuais, requereria maiores recursos computacionais.

Como trabalho futuro pode-se ainda sugerir mudanças na forma como as páginas a serem coletadas foram determinadas. No passo onde se procurou páginas que estabelecessem quem era a empresa poderia ser adicionada às palavras buscadas o nome da própria companhia. Conjectura a autora que páginas associadas ao propósito de contar a história da empresa apresentarão textos mais estruturados e maior estabilidade no que diz respeito à atividade econômica, diferente de domínios focados nos produtos vendidos e/ou serviços prestados pela empresa que, por exemplo, podem ser adaptados a um cenário pandêmico. A coleta de textos estruturados, que retratem uma sequência lógica, permitiria o uso de ferramentas que explorassem o contexto no qual as palavras foram inseridas (como *word embeddings*) bem como abririam outras possibilidades, como a execução da redução de dimensionalidade da base através da filtragem de classes gramaticais. Ainda assim, não se pode ignorar que o uso desse ferramental ficaria limitado aos desafios impostos pela língua portuguesa. Para essa, diferente do que para a língua inglesa, a quantidade de ferramentas disponíveis é reduzida e seus desempenhos inferiores.

Ainda no escopo de mudanças na coleta, poderiam ser analisados diversos conjuntos de variáveis preditoras, como fizeram Berardi *et al.* (2015). Conforme já mencionado no decorrer deste trabalho, através de um estudo experimental os autores objetivaram extrair percepções interessantes sobre quais dados são mais úteis para execução da tarefa de classificação da atividade econômica principal de uma empresa. Para tanto, foram consideradas características endógenas, obtidas de determinados campos da estrutura *HTML* (como *Título*, *URL*, *Tamanho da URL* e *Body*), e características exógenas, externas ao *site* da companhia (como as obtidas do Facebook e Twitter). Os autores avaliaram

[2] Detalhados na Subseção 3.1.2.

a capacidade preditiva de cada uma delas e de diversas combinações. Outro ponto que também pode ser sugerido é a inclusão da análise de imagens disponíveis nas páginas da *web* das empresas.

Mesmo com os textos coletados tal como foram, existem diversos passos do pré-processamento que poderiam ser modificados para se tentar obter um melhor ajuste dos modelos. Pode-se citar como exemplo a seleção das variáveis preditoras. Seria possível testar diversos cortes a partir da cobertura calculada ou utilizar métodos que buscassem medir a dependência entre as variáveis (por exemplo, *mutual information*). Uma outra forma, talvez mais complexa, seria executar a retirada de *tokens* através do uso de modelos de tópicos, o que poderia possibilitar a exclusão de conjuntos específicos[3]. Algumas dessas rotinas poderiam ser, inclusive, implementadas por nó, no caso do uso de classificadores locais. Além disso, seria factível considerar uma árvore distinta da exibida na Figura 4.1. A PAS, por exemplo, pauta a sua divulgação em 7 agrupamentos desconsiderados por este estudo. Indo mais além, seria possível também desprezar a divisão entre as Pesquisas Estruturais por Empresas, de forma a se olhar apenas a estrutura da CNAE. Vale mencionar que a abordagem por nó, via classificadores locais, é altamente sensível às decisões no topo da hierarquia e, uma vez que o erro é cometido próximo ao topo, ele não pode ser recuperado, independente de quão bom sejam os classificadores nos outros níveis. Soma-se ainda às diversas possibilidades de trabalhos futuros, o uso de um classificador conhecido como Global. De acordo com Silla e Freitas (2011), nesse tipo de abordagem, um único modelo é construído a partir do conjunto de treino, considerando a hierarquia de classes como um todo durante uma única execução do algoritmo. Muitas vezes esse tipo de classificador consiste em modificações de algoritmos tradicionalmente utilizados para classificação plana.

Destaca-se ainda que a 4ª revisão da *ISIC*, atualmente utilizada, foi aprovada pela Divisão de Estatísticas das Nações Unidas em 2006 e, de acordo com a própria, desde essa aprovação a globalização e a digitalização mudaram a forma como muitas atividades econômicas fornecem bens e serviços, novas atividades ganharam importância na economia global enquanto outras perderam e rápidas e dinâmicas mudanças ocorreram no contexto da tecnologia da informação. Além disso, a maior conscientização do impacto da economia sobre o meio ambiente criou atividades especializadas para protegê-lo (UNSD, 2022a). Ciente desses fatos, a Divisão de Estatísticas das Nações Unidas está propondo uma nova revisão da *ISIC* para refletir mais de perto a realidade das atividades econômicas atuais na classificação. Essas alterações, e outras que ocorrerão no futuro, exigirão novos testes da metodologia proposta neste trabalho. Possivelmente, uma classificação mais atual faça com que os algoritmos sejam capazes de identificar melhor os padrões de cada atividade, aperfeiçoando o seu desempenho.

Outro ponto que ainda merece atenção, apesar da já recebida durante este trabalho, diz respeito à defasagem temporal entre a extração dos dados da *web* e a CNAE informada pela empresa ao IBGE, não podendo esquecer do contexto pandêmico no qual se insere. Este estudo reuniu argumentos, apresentados na Subseção 3.1.1, que sustentavam a hipótese

[3] Poderiam, por exemplo, ser visados grupos que retratassem componentes oriundo da estrutura *HTML* da página extraída e não o texto de fato. Esse tratamento talvez pudesse evitar o aparecimento de *tokens* como *javascript*, na CNAE 38, nos Quadros 5.7 e 5.8. Ainda assim, nada garante que o modelo seria capaz de gerar esse tipo de agrupamento.

da validade da sua execução e que puderam ser verificados com base nos resultados apresentados. Entretanto, a partir da investigação do histórico e do comportamento das empresas classificadas incorretamente, detalhados nas Subseções 5.2.1 e 5.2.2, fica evidente que a defasagem temporal reflete nos resultados. Logo, para uma melhor avaliação, se faz necessário testar diferentes tempos de coleta dos textos com relação ao ano de referência da classificação econômica.

Ainda no que concerne à questão temporal, existe a necessidade de se avaliar a estabilidade do modelo desenvolvido no longo prazo. Isso é, se o modelo utilizado para identificar uma atividade econômica hoje pode ser o mesmo empregado para detectar essa atividade no futuro. Essa é uma questão relacionada à mudança de conceito, mencionada na Seção 2.2. Há de se considerar, por exemplo, se a incorporação/surgimento de novas tecnologias mudará o conjunto de *tokens* capazes de caracterizar determinadas CNAEs, se uma mudança de estação não poderia modificar os *tokens* relacionados às atividades de vestuário, ou ainda, se o surgimento de uma grande empresa de um segmento específico não poderia alterar *tokens* de várias CNAEs a ela relacionadas por conta de nomes e modelos associados aos seus produtos e/ou serviços. Essas questões, provavelmente, irão variar de acordo com o perfil da página da qual o texto é coletado. Já outro ponto, também ligado à mudança na forma como um conceito é expresso por uma população, diz respeito à mutabilidade da língua. As mudanças oriundas desta dinamicidade, exemplificada ao se tratar do *token desinsetizacao*, provavelmente seriam sentidas ao se considerar um tempo maior do que as citadas anteriormente e dificilmente teriam relação com a página da qual o texto foi extraído.

É importante destacar também que o estudo se limitou aos casos em que um *site* era associado a apenas uma CNAE principal. Conforme detalhado na Subseção 3.1.2, ao avaliar as bases da PIA-Empresa, PAC e PAS, ano de referência 2019, foi possível observar um grupo de CNPJs diferentes, concentrados nos estratos certo e gerencial, com mesma *URL* informada e atividades econômicas principais distintas. Além dessas empresas, também não foram abarcadas pelo estudo as que declararam ter sofrido mudanças estruturais ou outras formas de ligação entre empresas e as companhias pertencentes à PIA-Empresa que apresentaram CNAEs distintas em diferentes ULs. Tratar desses pontos ampliaria o escopo do método proposto.

Por fim, vale ressaltar que a coleta de dados da *web* pode auxiliar não só na atribuição da CNAE principal de uma empresa, como também em outros aspectos relacionados à atividade econômica, como na identificação de atividades emergentes através do uso de modelos tópicos, ou ainda, em assuntos não relacionados à CNAE como na detecção de comércio eletrônico, de empresas inovadoras ou sustentáveis. Essas informações podem ser utilizadas para incrementar as estatísticas já existentes ou para gerar novos resultados para os usuários.

Apêndice A

Funções Utilizadas para o Ajuste dos Algoritmos e Valores de Hiperparâmetros Aplicados na Busca Aleatória

A maioria dos algoritmos foi implementada através do pacote *Scikit-Learn*. Apenas o Boosting contou com o uso do pacote *LightGBM*. Naive Bayes foi ajustado através da função *MultinomialNB*, a Regressão Logística da função *LogisticRegression*, os Métodos de Suporte Vetorial através da função *SVC*. Já o Boosting por meio da função *LGBMClassifier*.

Os valores utilizados[1] na busca aleatória podem ser vistos no Quadro A.1.

Quadro A.1: Hiperparâmetros e representações testadas ao se considerar cada classificador

(Continua)

Classificador	Testes Executados
NB	IDF = [Sim, Não] alpha = Uniforme Contínua(0,1)
RL	IDF = [Sim, Não] C = Uniforme Contínua(0, 20) class_weight = [balanced, none] penalty = [l1, l2] solver = [liblinear, lbfgs] multi_class = [multinomial, ovr]
MSV	IDF = [Sim, Não] kernel = [linear, poly, rbf, sigmoid] C = Uniforme Contínua(0, 20) gamma = Uniforme Contínua(0, 5)

[1] Para hiperparâmetros não mencionados no Quadro A.1 foi considerado o *default* de cada função empregada.

Quadro A.1: Hiperparâmetros e representações testadas ao se considerar cada classificador

(Continuação)

Classificador	Testes Executados
Boosting	class_weight = [balanced, none] IDF = [Sim, Não] num_leaves = Uniforme Discreta(10, 85) max_depth = [3, 4, 5, 6, 7, 8, 9, 10] learning_rate = Uniforme Contínua(0,001, 0,1) n_estimators = Uniforme Discreta(50, 300) class_weight = [balanced, none]

Fonte: Autoria própria.

Nota: *class_weight* só foi utilizado quando não aplicada a SMOTE.

Onde[2]:

- NB_alpha: é o parâmetro de suavização aditivo, que ajuda a prevenir zeros.
- RL_C: é o parâmetro de regularização, sendo a intensidade da regularização inversamente proporcional a "C".
- RL_class_weight: o modo "balanced" usa os valores de cada classe para ajustar automaticamente pesos inversamente proporcionais às frequências, buscando assim evitar que o algoritmo favoreça os grupos com mais observações.
- RL_penalty: especifica a norma utilizada na penalização. O algoritmo "lbfgs" só suporta a penalização "l2".
- RL_solver: algoritmo utilizado no problema de otimização. "liblinear" foi empregado quando *multi_class = ovr* e "lbfgs" quando *multi_class = multinomial*.
- RL multi_class: se a opção escolhida for "ovr", então um problema binário é instituído para cada rótulo. Para "multinomial", a perda minimizada é o ajuste de perda multinomial em toda a distribuição de probabilidade.
- MSV_kernel: especifica o *kernel* utilizado pelo algoritmo na separação das classes.
- MSV_gamma: coeficiente do *kernel* para as opções "rbf", "poly" e "sigmoid".
- Boosting_num_leaves: define o número máximo de nós por árvore.
- Boosting_max_depth: controla a distância máxima entre o nó raiz de cada árvore e um nó folha.
- Boosting_learning_rate: taxa de aprendizagem.
- Boosting_n_estimators: número de árvores.

[2] Hiperparâmetros com mesma denominação e mesmo significado para classificadores distintos foram apresentados com referência a apenas um dos classificadores.

Apêndice B

Melhores Resultados Obtidos com cada Algoritmo

Tabela B.1: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: Pesquisa

(Continua)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Base ⟶ Pesquisa				
NB	Não	Não	alpha=0,007982 IDF=Não	0,7949 ± 0,0036
	Não	Sim	alpha=0,013337 IDF=Não	0,8004 ± 0,0028
	Sim	Não	alpha=0,012730 IDF=Não	0,7825 ± 0,0041
	Sim	Sim	alpha=0,048892 IDF=Não	0,7918 ± 0,0041
RL (1)	Não	Não	C=1,083790 IDF=Sim class_weight=balanced	0,8423 ± 0,0006
	Não	Sim	C=1,034904 IDF=Sim	0,8425 ± 0,0014
	Sim	Não	C=3,616011 IDF=Sim class_weight=balanced	0,8399 ± 0,0006
	Sim	Sim	C=1,585860 IDF=Sim	0,8353 ± 0,0008

Tabela B.1: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: Pesquisa

(Continuação)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Base → Pesquisa				
SVM	Não	Não	C=17,398171 kernel=rbf IDF=Sim class_weight=None gamma=0,998105	0,8425 ± 0,0003
	Não	Sim	C=17,870483 kernel=rbf IDF=Sim gamma=0,946583	0,8428 ± 0,0008
	Sim	Não	C=6,376696 kernel=rbf IDF=Não class_weight=None gamma=1,045366	0,8394 ± 0,0015
	Sim	Sim	C=7,072828 kernel=rbf IDF=Não gamma=0,632535	0,8367 ± 0,0017

Fonte: Autoria própria.

(1) Foi utilizado o *solver* lbfgs.

Tabela B.2: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC-CNAE2D

(Continua)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Pesquisa (PAC) → Divisão (PAC) - CNAE 2 dígitos				
NB	Não	Não	alpha=0,059187 IDF=Sim	0,8205 ± 0,0164
	Não	Sim	alpha=0,134699 IDF=Sim	0,8219 ± 0,0161
	Sim	Não	alpha=0,069791 IDF=Sim	0,8182 ± 0,0128
	Sim	Sim	alpha=0,137302 IDF=Sim	0,8203 ± 0,0161

Tabela B.2: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC-CNAE2D

(Continuação)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Pesquisa (PAC) ⟶ Divisão (PAC) - CNAE 2 dígitos				
RL (1)	Não	Não	C=2,838910 IDF=Sim class_weight=balanced	0,8432 ± 0,0115
	Não	Sim	C=2,919024 IDF=Sim	0,8406 ± 0,0109
	Sim	Não	C=2,777096 IDF=Sim class_weight=balanced	0,8397 ± 0,0131
	Sim	Sim	C=2,545693 IDF=Sim	0,8389 ± 0,0116
SVM	Não	Não	C=1,031337 kernel=linear IDF=Sim class_weight=balanced	0,8438 ± 0,0081
	Não	Sim	C=0,696176 kernel=linear IDF=Sim	0,8405 ± 0,0068
	Sim	Não	C=18,533635 kernel=sigmoid IDF=Sim class_weight=balanced gamma=0,0413035	0,8420 ± 0,0078
	Sim	Sim	C=0,676748 kernel=linear IDF=Não	0,8365 ± 0,0116

Fonte: Autoria própria.

(1) Foi utilizado o *solver* lbfgs.

Tabela B.3: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC45-CNAE3D

(Continua)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Divisão 45 (PAC) - CNAE 2 dígitos ⟶ Grupo (PAC) - CNAE 3 dígitos				
NB	Não	Não	alpha=0,038342 IDF=Sim	0,8935 ± 0,0129

Tabela B.3: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC45-CNAE3D

(Continuação)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Divisão 45 (PAC) - CNAE 2 dígitos → Grupo (PAC) - CNAE 3 dígitos				
NB	Não	Sim	alpha=0,142996 IDF=Não	0,9056 ± 0,0219
	Sim	Não	alpha=0,038312 IDF=Não	0,8974 ± 0,0222
	Sim	Sim	alpha=0,060886 IDF=Não	0,9031 ± 0,0251
RL (1)	Não	Não	C=19,395252 IDF=Sim class_weight=balanced	0,9358 ± 0,0051
	Não	Sim	C=17,803052 IDF=Não	0,9314 ± 0,0071
	Sim	Não	C=4,341370 IDF=Sim class_weight=balanced	0,9325 ± 0,0101
	Sim	Sim	C=1,457189 IDF=Sim	0,9297 ± 0,0148
SVM	Não	Não	C=2,327658 kernel=linear IDF=Sim class_weight=None	0,9362 ± 0,0033
	Não	Sim	C=7,970299 kernel=linear IDF=Sim	0,9319 ± 0,0018
	Sim	Não	C=5,906962 kernel=linear IDF=Não class_weight=balanced	0,9272 ± 0,0105
	Sim	Sim	C=15,631626 kernel=sigmoid IDF=Sim gamma=0,337022	0,9303 ± 0,0064

Fonte: Autoria própria.

(1) Foi utilizado o *solver* lbfgs.

Tabela B.4: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC46-CNAE3D

(Continua)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Divisão 46 (PAC) - CNAE 2 dígitos ⟶ Grupo (PAC) - CNAE 3 dígitos				
NB	Não	Não	alpha=0,029628 IDF=Sim	0,5520 ± 0,0089
	Não	Sim	alpha=0,252721 IDF=Sim	0,5759 ± 0,0287
	Sim	Não	alpha=0,002072 IDF=Sim	0,5587 ± 0,0047
	Sim	Sim	alpha=0,035169 IDF=Não	0,5910 ± 0,0278
RL (1)	Não	Não	C=4,370969 IDF=Sim class_weight=balanced penalty=l2	0,5869 ± 0,0269
	Não	Sim	C=3,949914 IDF=Sim penalty=l2	0,5950 ± 0,0192
	Sim	Não	C=5,373968 IDF=Sim class_weight=balanced penalty=l2	0,5966 ± 0,0276
	Sim	Sim	C=3,881384 IDF=Sim penalty=l2	0,6077 ± 0,03156
SVM	Não	Não	C=2,102957 kernel=linear IDF=Sim class_weight=balanced	0,5646 ± 0,01243
	Não	Sim	C=0,254238 kernel=sigmoid IDF=Sim gamma=4,807182	0,5868 ± 0,0181
	Sim	Não	C=5,3746837 kernel=sigmoid IDF=Sim class_weight=None gamma=0,393834	0,5667 ± 0,0171

Tabela B.4: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC46-CNAE3D

(Continuação)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Divisão 46 (PAC) - CNAE 2 dígitos → Grupo (PAC) - CNAE 3 dígitos				
SVM	Sim	Sim	C=10,592987 kernel=sigmoid IDF=Não gamma=0,093264	0,5667 ± 0,0252

Fonte: Autoria própria.

(1) Foi utilizado o *solver* liblinear.

Tabela B.5: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC47-CNAE3D

(Continua)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Divisão 47 (PAC) - CNAE 2 dígitos → Grupo (PAC) - CNAE 3 dígitos				
NB	Não	Não	alpha=0,0108497 IDF=Sim	0,7587 ± 0,0144
	Não	Sim	alpha=0,157630 IDF=Sim	0,7773 ± 0,0133
	Sim	Não	alpha=0,012695 IDF=Não	0,7494 ± 0,0190
	Sim	Sim	alpha=0,023089 IDF=Não	0,7665 ± 0,0111
RL (1)	Não	Não	C=6,718823 IDF=Sim class_weight=balanced penalty=l2	0,7762 ± 0,0144
	Não	Sim	C=18,535532 IDF=Sim penalty=l2	0,7798 ± 0,0159
	Sim	Não	C=10,302002 IDF=Sim class_weight=balanced penalty=l2	0,7928 ± 0,0182
	Sim	Sim	C=11,340738 IDF=Sim penalty=l2	0,7889 ± 0,0193

Tabela B.5: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAC47-CNAE3D

(Continuação)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Divisão 47 (PAC) - CNAE 2 dígitos ⟶ Grupo (PAC) - CNAE 3 dígitos				
SVM	Não	Não	C=3,059353 kernel=sigmoid IDF=Sim class_weight=None gamma=1,615835	0,7581 ± 0,0221
	Não	Sim	C=10,488046 kernel=sigmoid IDF=Sim gamma=0,549346	0,7559 ± 0,0230
	Sim	Não	C=0,583245 kernel=sigmoid IDF=Sim class_weight=balanced gamma=1,697874	0,7734 ± 0,0231
	Sim	Sim	C=13,417074 kernel=sigmoid IDF=Sim gamma=0,640128	0,7654 ± 0,0146

Fonte: Autoria própria.

(1) Foi utilizado o *solver* liblinear.

Tabela B.6: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PIA-CNAE2D

(Continua)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Pesquisa (PIA) ⟶ Divisão (PIA) - CNAE 2 dígitos				
NB	Não	Não	alpha=0,014889 IDF=Sim	0,6659 ± 0,0025
	Não	Sim	alpha=0,109001 IDF=Sim	0,6984 ± 0,0019
	Sim	Não	alpha=0,043785 IDF=Sim	0,6232 ± 0,0022
	Sim	Sim	alpha=0,129188 IDF=Sim	0,6859 ± 0,0040

Tabela B.6: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PIA-CNAE2D

(Continuação)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Pesquisa (PIA) → Divisão (PIA) - CNAE 2 dígitos				
RL (1)	Não	Não	C=6,500593 IDF=Sim class_weight=balanced penalty=l2	0,7225 ± 0,0021
	Não	Sim	C=5,367719 IDF=Sim penalty=l2	0,7239 ± 0,0020
	Sim	Não	C=13,982237 IDF=Sim class_weight=balanced penalty=l2	0,7120 ± 0,0012
	Sim	Sim	C=1,379516 IDF=Sim penalty=l2	0,7156 ± 0,0017
SVM	Não	Não	C=2,939987 kernel=linear IDF=Sim class_weight=None	0,6809 ± 0,0064
	Não	Sim	C=3,022732 kernel=sigmoid IDF=Sim gamma=0,304336	0,6889 ± 0,0065
	Sim	Não	C=1,744822 kernel=linear IDF=Sim class_weight=balanced	0,7012 ± 0,0074
	Sim	Sim	C=0,481314 kernel=linear IDF=Sim	0,7054 ± 0,0085

Fonte: Autoria própria.

(1) Foi utilizado o *solver* liblinear.

Tabela B.7: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAS-CNAE2D

(Continua)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Pesquisa (PAS) ⟶ Divisão (PAS) - CNAE 2 dígitos				
NB	Não	Não	alpha=0,012942 IDF=Sim	0,6319 ± 0,0062
	Não	Sim	alpha=0,968407 IDF=Sim	0,6661 ± 0,0061
	Sim	Não	alpha=0,010516 IDF=Sim	0,6314 ± 0,0033
	Sim	Sim	alpha=0,229816 IDF=Sim	0,6641 ± 0,0065
RL (1)	Não	Não	C=2,526199 IDF=Sim class_weight=balanced penalty=l2	0,6945 ± 0,0022
	Não	Sim	C=2,459526 IDF=Sim penalty=l2	0,6984 ± 0,0038
	Sim	Não	C=6,436504 IDF=Sim class_weight=balanced penalty=l2	0,6959 ± 0,0064
	Sim	Sim	C=1,306568 IDF=Sim penalty=l2	0,7007 ± 0,0068
SVM	Não	Não	C=9,454726 kernel=linear IDF=Sim class_weight=balanced	0,6642 ± 0,0067
	Não	Sim	C=1,463434 kernel=sigmoid IDF=Sim gamma=1,033298	0,6662 ± 0,0051
	Sim	Não	C=9,636953 kernel=rbf IDF=Sim class_weight=balanced gamma=0,051751	0,6747 ± 0,0007

Tabela B.7: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem hierárquica - Nó: PAS-CNAE2D

(Continuação)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
Nó: Pesquisa (PAS) → Divisão (PAS) - CNAE 2 dígitos				
SVM	Sim	Sim	C=1,674750 kernel=linear IDF=Sim	0,6674 ± 0,0065

Fonte: Autoria própria.

(1) Foi utilizado o *solver* liblinear.

Tabela B.8: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem plana

(Continua)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
NB	Não	Não	alpha=0,012978 IDF=Sim	0,5413 ± 0,0031
	Não	Sim	alpha=0,118592 IDF=Sim	0,5697 ± 0,0024
	Sim	Não	alpha=0,013247 IDF=Sim	0,5334 ± 0,0019
	Sim	Sim	alpha=0,198588 IDF=Sim	0,5639 ± 0,0035
RL (1)	Não	Não	C=4,811670 penalty=l2 IDF=Sim class_weight=balanced	0,5966 ± 0,0007
	Não	Sim	C=2,681655 penalty=l2 IDF=Sim	0,5990 ± 0,0015
	Sim	Não	C=5,168965 penalty=l2 IDF=Sim class_weight=balanced	0,5926 ± 0,0035
	Sim	Sim	C=1,919590 penalty=l2 IDF=Sim	0,5925 ± 0,0060

Tabela B.8: Melhores resultados obtidos na base de treino, com cada algoritmo, e suas configurações ao se considerar a abordagem plana

(Continuação)

Classificador	STEMMER	SMOTE	Configuração	Validação Cruzada (*F1_macro*)
SVM	Não	Não	C=0,841080 kernel=linear IDF=Sim class_weight=balanced	0,5838 ± 0,0035
	Não	Sim	C=6,582741 kernel=sigmoid IDF=Sim gamma=0,052693	0,5821 ± 0,0047
	Sim	Não	C=0,801103 kernel=linear IDF=Sim class_weight=balanced	0,5864 ± 0,0049
	Sim	Sim	C=7,734721 kernel=linear IDF=Sim	0,5732 ± 0,0028
Boosting	Não	Não	max_depth=3 learning_rate=0,031359 n_estimators=248 num_leaves=67 IDF=Não class_weight=balanced	0,5499 ± 0,0061
	Não	Sim	max_depth=3 learning_rate=0,025845 n_estimators=339 num_leaves=60 IDF=Não	0,5554 ± 0,0030
	Sim	Não	max_depth=3 learning_rate=0,025492 n_estimators=287 num_leaves=61 IDF=Não class_weight=balanced	0,5551 ± 0,0065
	Sim	Sim	max_depth=9 learning_rate=0,012958 n_estimators=243 num_leaves=20 IDF=Não	0,5449 ± 0,0021

Fonte: Autoria própria.

(1) Foi utilizado o *solver* liblinear.

Anexo A

Códigos e Denominações de parte da Classificação Nacional de Atividades Econômicas - CNAE

Quadro A.1: Estrutura detalhada de parte da CNAE 2.0: Códigos e denominações

(Continua)

Seção	Divisão	Grupo	Classe	Denominação
A				**Agricultura, pecuária, produção florestal, pesca e aquicultura**
	01			Agricultura, pecuária e serviços relacionados
	02			Produção Florestal
B				**Indústrias extrativas**
	05			Extração de carvão mineral
	06			Extração de petróleo e gás natural
	07			Extração de minerais metálicos
	08			Extração de minerais não-metálicos
	09			Atividades de apoio à extração de minerais
C				**Indústrias de transformação**
	10			Fabricação de produtos alimentícios
	11			Fabricação de bebidas
	12			Fabricação de produtos do fumo
	13			Fabricação de produtos têxteis
	14			Confecção de artigos do vestuário e acessórios
	15			Preparação de couros e fabricação de artefatos de couro, artigos para viagem e calçados
	16			Fabricação de produtos de madeira
	17			Fabricação de celulose, papel e produtos de papel
	18			Impressão e reprodução de gravações
	19			Fabricação de coque, de produtos derivados do petróleo e de biocombustíveis
	20			Fabricação de produtos químicos

Quadro A.1: Estrutura detalhada de parte da CNAE 2.0: Códigos e denominações

(Continuação)

Seção	Divisão	Grupo	Classe	Denominação
	21			Fabricação de produtos farmoquímicos e farmacêuticos
	22			Fabricação de produtos de borracha e de material plástico
	23			Fabricação de produtos de minerais não-metálicos
	24			Metalurgia
	25			Fabricação de produtos de metal, exceto máquinas e equipamentos
	26			Fabricação de equipamentos de informática, produtos eletrônicos e ópticos
	27			Fabricação de máquinas, aparelhos e materiais elétricos
	28			Fabricação de máquinas e equipamentos
	29			Fabricação de veículos automotores, reboques e carrocerias
	30			Fabricação de outros equipamentos de transporte, exceto veículos automotores
	31			Fabricação de móveis
	32			Fabricação de produtos diversos
	33			Manutenção, reparação e instalação de máquinas e equipamentos
E				**Água, esgoto, atividades de gestão de resíduos e descontaminação**
	37			Esgoto e atividades relacionadas
	38			Coleta, tratamento e disposição de resíduos; recuperação de materiais
	39			Descontaminação e outros serviços de gestão de resíduos
G				**Comércio; reparação de veículos automotores e motocicletas**
	45			Comércio e reparação de veículos automotores e motocicletas
		45.1		Comércio de veículos automotores
			45.11-1	Comércio a varejo e por atacado de veículos automotores
			45.12-9	Representantes comerciais e agentes do comércio de veículos automotores
		45.2		Manutenção e reparação de veículos automotores
			45.20-0	Manutenção e reparação de veículos automotores
		45.3		Comércio de peças e acessórios para veículos automotores
			45.30-7	Comércio de peças e acessórios para veículos automotores

Quadro A.1: Estrutura detalhada de parte da CNAE 2.0: Códigos e denominações

(Continuação)

Seção	Divisão	Grupo	Classe	Denominação
		45.4		Comércio, manutenção e reparação de motocicletas, peças e acessórios
			45.41-2	Comércio por atacado e a varejo de motocicletas, peças e acessórios
			45.42-1	Representantes comerciais e agentes do comércio de motocicletas, peças e acessórios
			45.43-9	Manutenção e reparação de motocicletas
	46			Comércio por atacado, exceto veículos automotores e motocicletas
		46.1		Representantes comerciais e agentes do comércio, exceto de veículos automotores e motocicletas
		46.2		Comércio atacadista de matérias-primas agrícolas e animais vivos
		46.3		Comércio atacadista especializado em produtos alimentícios, bebidas e fumo
		46.4		Comércio atacadista de produtos de consumo não-alimentar
		46.5		Comércio atacadista de equipamentos e produtos de tecnologias de informação e comunicação
		46.6		Comércio atacadista de máquinas, aparelhos e equipamentos, exceto de tecnologias de informação e comunicação
		46.7		Comércio atacadista de madeira, ferragens, ferramentas, material elétrico e material de construção
		46.8		Comércio atacadista especializado em outros produtos
		46.9		Comércio atacadista não-especializado
	47			Comércio varejista
		47.1		Comércio varejista não-especializado
		47.2		Comércio varejista de produtos alimentícios, bebidas e fumo
		47.3		Comércio varejista de combustíveis para veículos automotores
		47.4		Comércio varejista de material de construção
		47.5		Comércio varejista de equipamentos de informática e comunicação; equipamentos e artigos de uso doméstico
		47.6		Comércio varejista de artigos culturais, recreativos e esportivos
		47.7		Comércio varejista de produtos farmacêuticos, perfumaria e cosméticos e artigos médicos, ópticos e ortopédicos

Quadro A.1: Estrutura detalhada de parte da CNAE 2.0: Códigos e denominações

(Continuação)

Seção	Divisão	Grupo	Classe	Denominação
		47.8		Comércio varejista de produtos novos não especificados anteriormente e de produtos usados
H				**Transporte, armazenagem e correio**
	49			Transporte terrestre
	50			Transporte aquaviário
	51			Transporte aéreo
	52			Armazenamento e atividades auxiliares dos transportes
	53			Correio e outras atividades de entrega
I				**Alojamento e alimentação**
	55			Alojamento
	56			Alimentação
J				**Informação e comunicação**
	58			Edição e edição integrada à impressão
	59			Atividades cinematográficas, produção de vídeos e de programas de televisão; gravação de som e edição de música
	60			Atividades de rádio e de televisão
	61			Telecomunicações
	62			Atividades dos serviços de tecnologia da informação
	63			Atividades de prestação de serviços de informação
K				**Atividades financeiras, de seguros e serviços relacionados**
	66			Atividades auxiliares dos serviços financeiros, seguros previdência complementar e planos de saúde
L				**Atividades imobiliárias**
	68			Atividades imobiliárias
M				**Atividades profissionais, científicas e técnicas**
	69			Atividades jurídicas, de contabilidade e de auditoria
	70			Atividades de sedes de empresas e de consultoria em gestão empresarial
	71			Serviços de arquitetura e engenharia; testes e análises técnicas
	73			Publicidade e pesquisa de mercado
	74			Outras atividades profissionais, científicas e técnicas
N				**Atividades administrativas e serviços complementares**
	77			Aluguéis não-imobiliários e gestão de ativos intangíveis não-financeiros
	78			Seleção, agenciamento e locação de mão-de-obra
	79			Agências de viagens, operadores turísticos e serviços de reservas

Quadro A.1: Estrutura detalhada de parte da CNAE 2.0: Códigos e denominações

(Continuação)

Seção	Divisão	Grupo	Classe	Denominação
	80			Atividades de vigilância, segurança e investigação
	81			Serviços para edifícios e atividades paisagísticas
	82			Serviços de escritório, de apoio administrativo e outros serviços prestados às empresas
P				**Educação**
	85			Educação
R				**Artes, cultura, esporte e recreação**
	90			Atividades artísticas, criativas e de espetáculos
	92			Atividades de exploração de jogos de azar e apostas
	93			Atividades esportivas e de recreção e lazer
S				**Outras atividades de serviços**
	95			Reparação e manutenção de equipamentos de informática e comunicação e de objetos pessoais e domésticos
	96			Outras atividades de serviços pessoais

Fonte: IBGE e CONCLA (2020).

Referências

[Bagno 1999] Marcos Bagno. *Preconceito Lingüístico - o que é, como se faz*. 49ª ed. Edições Loyola, 1999 (citado na pg. 11).

[Bengfort *et al.* 2018] Benjamin Bengfort, Rebecca Bilbro e Tony Ojeda. *Applied Text Analysis with Python. Enabling Language-Aware Data Products with Machine Learning*. 1ª ed. O'REILLY, 2018 (citado nas pgs. 47, 49).

[Berardi *et al.* 2015] Giacomo Berardi, Andrea Esuli, Tiziano Fagni e Fabrizio Sebastiani. "Classifying websites by industry sector: a study in feature design". Em: *Proceedings of the 30th Annual ACM Symposium on Applied Computing* (2015), pgs. 1053–1059 (citado nas pgs. 37, 60, 99).

[Bergstra e Bengio 2012] James Bergstra e Yoshua Bengio. "Random search for hyper-parameter optimization". Em: *Journal of Machine Learning Research* 13 (2012), pgs. 281–305 (citado na pg. 63).

[Blei e Smyth 2017] David M. Blei e Padhraic Smyth. "Science and data science". Em: *Proceedings of the National Academy of Sciences of the United States of America* 114 (2017), pgs. 8689–8692 (citado nas pgs. 8, 9).

[Brasil 1974] Brasil. *Lei nº 6.183, de 11 de dezembro de 1974. Dispõe sobre os Sistemas Estatístico e Cartográfico Nacionais, e dá outras providências.* 1974. url: http://www.planalto.gov.br/ccivil_03/LEIS/1970-1979/L6183.htm (acesso em 24/06/2021) (citado na pg. 15).

[Brasil 2018] Brasil. *Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD).* 2018. url: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm (acesso em 20/01/2022) (citado na pg. 1).

[IBGE 2021a] Coordenação de Cadastro e Classificações. *Estatísticas do Cadastro Central de Empresas 2019*. Rel. técn. Rio de Janeiro, BRA: Instituto Brasileiro de Geografia e Estatística, 2021. url: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101833.pdf (citado nas pgs. 20, 21).

[StatCan 2021] Statistics Canada. *Home. Using new and existing data for official statistics. Where our data comes from.* 2021. url: https://www.statcan.gc.ca/en/our-data/where/web-scraping (acesso em 16/01/2022) (citado na pg. 2).

[Castilho 2017] Ataliba T. de Castilho. *Como as Línguas Nascem e Morrem? O que são Famílias Lingüísticas?* 2017. URL: https://www.museudalinguaportuguesa.org.br/wp-content/uploads/2017/09/Como-as-linguas-nascem-e-morrem.pdf (acesso em 10/06/2021) (citado na pg. 11).

[Caterini 2018] Giacomo Caterini. "Classifying firms with text mining". Em: *DEM Working Papers* 9 (2018) (citado na pg. 60).

[Chawla *et al.* 2002] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall e W. Philip Kegelmeyer. "Smote: synthetic minority over-sampling technique". Em: *Journal of Artificial Intelligence Research* 16 (2002), pgs. 321–357 (citado na pg. 50).

[Chollet e Allaire 2017] François Chollet e J.J. Allaire. *Deep Learning with R*. 1ª ed. Manning, 2017 (citado na pg. 10).

[Conceição da Silva Ferreira 2012] Maria da Conceição da Silva Ferreira. "Classificação Hierárquica da Atividade Económica das Empresas a partir de Texto da Web". Diss. de mestr. Porto, Portugal: Universidade do Porto, 2012 (citado na pg. 53).

[Daas e Doef 2021] Piet J. H. Daas e Suzanne van der Doef. "Using website texts to detect innovative companies". Em: *Center for Big Data Statistics* Working paper nº 01-21 (2021) (citado na pg. 67).

[UNSD 2022a] United Nations Statistics Division. *Revised new structure of ISIC. Note on the main changes to ISIC Rev. 4*. 2022. URL: https://unstats.un.org/unsd/classifications/ISIC/Main_changes_in_ISIC_14_Jan_2022.pdf (acesso em 08/02/2022) (citado na pg. 100).

[UNSD 2022b] United Nations Statistics Division. *United Nations. Department of Economic and Social Affairs Statistics Division. Classifications. Economic statistics. ISIC*. URL: https://unstats.un.org/unsd/classifications/Econ/ISIC.cshtml (acesso em 27/06/2022) (citado na pg. 17).

[UNSD 2022c] United Nations Statistics Division. *United Nations. Department of Economic and Social Affairs Statistics Division. Fundamental Principles of National Official Statistics*. URL: https://unstats.un.org/fpos/ (acesso em 19/05/2022) (citado na pg. 12).

[Eurostat 2020] Eurostat. *ESSnet Big Data II. Deliverable K5: First draft of the methodological report*. Rel. técn. 2020. URL: https://ec.europa.eu/eurostat/cros/sites/default/files/WPK_Deliverable_K5_First_draft_of_the_methodological_report_2020_06_17_Finalv3.pdf (citado na pg. 46).

[Eurostat 2021] Eurostat. *Guidance on the impact of the Covid-19 crisis on annual business statistics*. Rel. técn. 2021. URL: https://ec.europa.eu/eurostat/documents/10186/10693286/Guidance-on-the-impact-of-the-COVID-19-crisis-on-annual-business-statistics.pdf/81560f12-44a7-a6d5-3eb5-639a925b0a67?t=1609855074295 (citado na pg. 32).

[Eurostat 2022a] Eurostat. *European Commission. Eurostat. CROS. ESSnet Big Data II. ESSnet Big Data. WPC Enterprise characteristics. WPC ESSnet Webscraping policy draft*. 2022. URL: https://ec.europa.eu/eurostat/cros/content/WPC_ESSnet_Webscraping_policy_draft_en (acesso em 16/01/2022) (citado na pg. 2).

[Eurostat 2022b] Eurostat. *European Commission. Eurostat. CROS. ESSnet Big Data II. ESSnet Big Data. WPC Enterprise characteristics. WPC Experimental statistics*. 2022. URL: https://ec.europa.eu/eurostat/cros/content/wpc-experimental-statistics_en#Netherlands (acesso em 12/01/2022) (citado na pg. 98).

[Feijó e Valente 2006] Carmem Feijó e Elvio Valente. "As estatísticas oficiais e o interesse público". Em: *II Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais- CONFEST* (IBGE, Rio de Janeiro, 21–25 de ago. de 2006). 2006 (citado na pg. 16).

[IBGE 2020a] Instituto Brasileiro de Geografia e Estatística. *IBGE. Estatísticas Experimentais. Pesquisa Pulso Empresa: Impacto da Covid-19 nas empresas*. 2020. URL: https://www.ibge.gov.br/estatisticas/investigacoes-experimentais/estatisticas-experimentais/28291-pesquisa-pulso-empresa-impacto-da-covid-19-nas-empresas.html?edicao=28390&t=o-que-e (acesso em 18/11/2021) (citado nas pgs. 33, 34).

[IBGE 2021b] Instituto Brasileiro de Geografia e Estatística. *IBGE. CNAE - Classificação Nacional de Atividades Econômicas*. 2021. URL: https://www.ibge.gov.br/estatisticas/metodos-e-classificacoes/classificacoes-e-listas-estatisticas/9078-classificacao-nacional-de-atividades-economicas.html?=&t=conceitos-e-metodos (acesso em 30/06/2021) (citado na pg. 17).

[IBGE 2021c] Instituto Brasileiro de Geografia e Estatística. *Institucional. Códigos e Princípios. Princípios Fundamentais das Estatísticas Oficiais*. 2021. URL: https://www.ibge.gov.br/institucional/codigos-e-principios.html?option=com_content&view=article&id=16148 (acesso em 23/06/2021) (citado na pg. 12).

[IBGE e CONCLA 2020] Instituto Brasileiro de Geografia e Estatística e Comissão Nacional de Classificação. *Classificação Nacional de Atividades Econômicas - Subclasses para Uso da Administração Pública - Versão 2.3*. Rel. técn. Volume 29. Rio de Janeiro, BRA: Instituto Brasileiro de Geografia e Estatística, 2020. URL: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101721.pdf (citado nas pgs. 17, 18, 70, 71, 87, 121).

[IBGE e CONCLA 2021] Instituto Brasileiro de Geografia e Estatística e Comissão Nacional de Classificação. *Documentação. Cronologia. CNAE.* 2021. URL: https://concla.ibge.gov.br/documentacao/cronologia/cnae-cronologia.html (acesso em 29/06/2021) (citado na pg. 17).

[IBGE e CONCLA 2022] Instituto Brasileiro de Geografia e Estatística e Comissão Nacional de Classificação. *Busca online CNAE.* 2022. URL: https://cnae.ibge.gov.br/ (acesso em 14/01/2022) (citado nas pgs. 70, 87).

[Han *et al.* 2011] Jiawei Han, Micheline Kamber e Jian Pei. *Data Mining - Concepts and Techniques.* 3ª ed. Elsevier, 2011 (citado na pg. 10).

[Hastie *et al.* 2009] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning - Data Mining, Inference and Prediction.* 2ª ed. Springer, 2009 (citado nas pgs. 56, 57, 62).

[Holcombe 1997] Randall G. Holcombe. "A theory of the theory of public goods". Em: *The Review of Economics and Statistics* 10.1 (1997), pgs. 1–22 (citado na pg. 12).

[IBGE 2020b] Coordenação de Índices de Preços. *Sistema Nacional de Índices de Preços ao Consumidor - Métodos de Cálculo. Série Relatórios Metodológicos.* Rel. técn. Volume 14. Rio de Janeiro, BRA: Instituto Brasileiro de Geografia e Estatística, 2020. URL: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101767.pdf (citado na pg. 12).

[IBGE 2004] Coordenação de Indústria. *Pesquisa Industrial Anual - Empresa. Série Relatórios Metodológicos.* Rel. técn. Volume 26. Rio de Janeiro, BRA: Instituto Brasileiro de Geografia e Estatística, 2004. URL: https://biblioteca.ibge.gov.br/visualizacao/livros/liv4178.pdf (citado nas pgs. 35, 36).

[NIC.br 2022a] Núcle de Informação e Coordenação do Ponto BR. *Home. Domínio. Categorias.br.* 2022. URL: https://registro.br/dominio/categorias/ (acesso em 05/01/2022) (citado na pg. 37).

[NIC.br 2022b] Núcle de Informação e Coordenação do Ponto BR. *Home. Domínio. Estatísticas.* 2022. URL: https://registro.br/dominio/estatisticas/ (acesso em 05/01/2022) (citado na pg. 37).

[IBGE 2021d] Instituto Brasileiro de Geografia e Estatística. *Institucional. O IBGE.* 2021. URL: https://www.ibge.gov.br/institucional/o-ibge.html (acesso em 03/06/2021) (citado na pg. 14).

[IBGE 2021e] Instituto Brasileiro de Geografia e Estatística. *Você foi procurado pelo IBGE? Para que servem as pesquisas do IBGE?* 2021. URL: https://respondendo.ibge.gov.br/voce-foi-procurado-pelo-ibge/para-que-servem-as-pesquisas-do-ibge.html (acesso em 03/06/2021) (citado na pg. 15).

[Instituto de Tecnologia da Georgia 2021] Instituto de Tecnologia da Georgia. *Home. Jeff Wu.* 2021. URL: https://www.isye.gatech.edu/users/jeff-wu (acesso em 10/06/2021) (citado na pg. 8).

[James *et al.* 2021] Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* 2ª ed. Springer, 2021 (citado nas pgs. 55, 58–62, 69).

[Jordan e Mitchell 2015] M. I. Jordan e T. M. Mitchell. "Machine learning: trends, perspectives, and prospects". Em: *Science (New York, N.Y.)* 349 (2015), pgs. 255–260 (citado na pg. 9).

[Jurafsky e Martin 2021] Daniel Jurafsky e James H. Martin. *Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 3ª ed. Em fase de elaboração, 2021 (citado nas pgs. 54–56, 62).

[Kelleher e Tierney 2018] John D. Kelleher e Brendan Tierney. *Data Science.* The MIT Press, 2018 (citado na pg. 10).

[Kiritchenko *et al.* 2006] Svetlana Kiritchenko, Stan Matwin, Richard Nock e A. Fazel Famili. "Learning and evaluation in the presence of class hierarchies: application to text categorizationn". Em: *Conference of the Canadian Society for Computational Studies of Intelligence* (2006), pgs. 395–406 (citado na pg. 64).

[Kosmopoulos *et al.* 2015] Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras e Ion Androutsopoulos. "Evaluation measures for hierarchical classification: a unified view and novel approaches". Em: *Data Mining and Knowledge Discovery* 29 (2015), pgs. 820–865 (citado na pg. 65).

[Lemaître *et al.* 2017] Guillaume Lemaître, Fernando Nogueira e Christos K. Aridas. "Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning". Em: *Journal of Machine Learning Research* 18.17 (2017), pgs. 1–5. URL: http://jmlr.org/papers/v18/16-365.html (citado na pg. 50).

[ONS 2020] Office for National Statistics. *Home. About us. Transparency and governance. Data Strategy. Data Policies. Web Scraping Policy.* 2020. URL: https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/webscrapingpolicy#toc (acesso em 16/01/2022) (citado na pg. 2).

[Naur 1974] Peter Naur. *Concise Survey of Computer Methods.* Petrocelli, 1974 (citado na pg. 8).

[O'Neil 2020] Cathy O'Neil. *Algoritmos de Destruição em Massa: Como o Big Data Aumenta a Desigualdade e Ameaça a Democracia.* Trad. por Rafael Abraham. Rua do Sabão, 2020 (citado na pg. 13).

[OCDE 2019] Organização para Cooperação e Desenvolvimento Econômico. *OECD/LEGAL/0449/Recommendation of the Council on Artificial Intelligence*. 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (acesso em 10/06/2021) (citado na pg. 10).

[Pacheco 2021] Mariana Pacheco. *Dedetizar ou detetizar?* Ed. por Brasil Escola. URL: https://brasilescola.uol.com.br/gramatica/dedetizar-ou-detetizar.htm (acesso em 25/10/2021) (citado na pg. 87).

[Pedregosa *et al.* 2011] F. Pedregosa *et al.* "Scikit-learn: machine learning in Python". Em: *Journal of Machine Learning Research* 12 (2011), pgs. 2825–2830 (citado na pg. 48).

[Pomikálek 2011] Jan Pomikálek. "Removing Boilerplate and Duplicate Content from Web Corpora". Tese de dout. Brno, República Tcheca: Faculty of Informatics, Masaryk University, 2011 (citado na pg. 37).

[Press 2013] Gil Press. *A Very Short History of Data Science*. 2013. URL: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/?sh=b15ae3255cfc (acesso em 09/06/2021) (citado na pg. 8).

[Puts e Daas 2021] Marco J. H. Puts e Piet J. H. Daas. "Machine learning from the perspective of official statistic". Em: *The Survey Statistician* 84 (2021), pgs. 12–17 (citado na pg. 13).

[Roelands 2017] Maarten Roelands. "Classifying Economic Activity with Business Websites Using Text Mining Techniques". Diss. de mestr. Tilburg, The Netherlands: Tilburg University, jul. de 2017 (citado na pg. 54).

[Rolland 2017] Asle Rolland. "The concept and commodity of official statistics". Em: *Statistical Journal of the IAOS* 33 (2017), pgs. 373–385 (citado nas pgs. 11, 12).

[Samuelson 1954] Paul A. Samuelson. "The pure theory of public expenditure". Em: *The Review of Economics and Statistics* 36.4 (1954), pgs. 387–389 (citado na pg. 12).

[IBGE 2021f] Coordenação de Serviços e Comércio. *Pesquisa Anual da Indústria da Construção 2019 - Notas técnicas*. Rel. técn. Volume 29. Rio de Janeiro, BRA: Instituto Brasileiro de Geografia e Estatística, 2021. URL: https://biblioteca.ibge.gov.br/visualizacao/periodicos/54/paic_2019_v29_notas_tecnicas.pdf (citado nas pgs. 22, 23).

[IBGE 2021g] Coordenação de Serviços e Comércio. *Pesquisa Anual de Comércio 2019 - Notas técnicas*. Rel. técn. Volume 31. Rio de Janeiro, BRA: Instituto Brasileiro de Geografia e Estatística, 2021. URL: https://biblioteca.ibge.gov.br/visualizacao/periodicos/55/pac_2019_v31_notas_tecnicas.pdf (citado nas pgs. 23, 24, 36).

[IBGE 2021h] Coordenação de Serviços e Comércio. *Pesquisa Anual de Serviços 2019 - Notas técnicas*. Rel. técn. Volume 21. Rio de Janeiro, BRA: Instituto Brasileiro de Geografia e Estatística, 2021. URL: https://biblioteca.ibge.gov.br/visualizacao/periodicos/150/pas_2019_v21_notas_tecnicas.pdf (citado nas pgs. 24, 25, 36).

[IBGE 2021i] Coordenação de Serviços e Comércio. *Pesquisa Industrial - Empresa 2019 - Notas técnicas*. Rel. técn. Volume 38 - número 1. Rio de Janeiro, BRA: Instituto Brasileiro de Geografia e Estatística, 2021. URL: https://biblioteca.ibge.gov.br/visualizacao/periodicos/1719/pia_2019_v38_n1_empresa_notas_tecnicas.pdf (citado nas pgs. 19, 22, 36).

[Silla e Freitas 2011] Carlos N. Silla e Alex A. Freitas. "A survey of hierarchical classification across different application domains". Em: *Data Mining and Knowledge Discovery* 22 (2011), pgs. 31–72 (citado nas pgs. 50–52, 100).

[Steven Bird e Loper 2009] Ewan Klein Steven Bird e Edward Loper. *Natural Language Processing with Python*. O'Reilly Media Inc, 2009. URL: https://www.nltk.org/book/ (citado na pg. 37).

[United Nations 2021] *The Handbook on Management and Organization of National Statistical Systems*. 4ª ed. 2021. URL: https://unstats.un.org/capacity-development/handbook/UN%5C%20Handbook%5C%20beta%5C%20v2.1.pdf (acesso em 31/05/2022) (citado nas pgs. 16, 35).

[Tigre 2019a] Paulo Bastos Tigre. "A nova economia dos serviços". Em: *Inovação em Serviços na Economia do Compartilhamento*. Ed. por Saraiva Educação. 1ª ed. 2019, pgs. XIX–XXVII (citado na pg. 32).

[Tigre 2019b] Paulo Bastos Tigre. "Trajetórias e oportunidades das inovações em serviços". Em: *Inovação em Serviços na Economia do Compartilhamento*. Ed. por Saraiva Educação. 1ª ed. 2019, pgs. 2–21 (citado na pg. 2).

[Tigre e Pinheiro 2019] Paulo Bastos Tigre e Alesssandro Maia Pinheiro. *Inovação em Serviços na Economia do Compartilhamento*. 1ª ed. Saraiva Educação, 2019 (citado na pg. 31).

[Woods 2016] Tyler Woods. *'Mathwashing,' Facebook and the Zeitgeist of Data Worship*. 2016. URL: https://technical.ly/brooklyn/2016/06/08/fred-benenson-mathwashing-facebook-data-worship/ (acesso em 23/06/2021) (citado na pg. 13).

[Yung 2021] Wesley Yung. "The evolution of official statistics in a changing world". Em: *Harvard Data Science Review* 3.4 (2021) (citado na pg. 13).

[Yung, Karkimaa *et al.* 2018] Wesley Yung, Jukka Karkimaa *et al.* "The use of machine learning in official statistics". Em: *UNECE Machine Learning Team report* (2018) (citado nas pgs. 12, 14).

[Yung, Tam *et al.* 2022] Wesley Yung, Siu-Ming Tam *et al.* "A quality framework for statistical algorithms". Em: *Statistical Journal of the IAOS* 38 (2022), pgs. 291–308 (citado na pg. 14).

[Zeelenberg 2016] Kees Zeelenberg. *Big Data and Methodological Challenges in Official Statistics - Discussion Paper*. Rel. técn. 8. Statistics Netherlands (CBS), 2016 (citado na pg. 2).

[Zhou *et al.* 2021] Nengfeng Zhou *et al. Bias, Fairness, and Accountability with AI and ML Algorithms*. 2021. URL: https://arxiv.org/abs/2105.06558 (acesso em 28/06/2022) (citado na pg. 13).