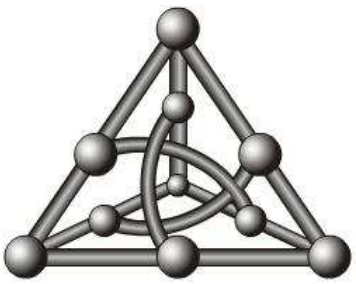
Charles Andre Profilio dos Santos

Uma Análise Exploratória da Influência dos Projetos Pedagógicos dos Cursos Superiores no Resultado do Enade por meio de Mineração de Textos e Aprendizado de Máquina

Campo Grande - MS
2022

Uma Análise Exploratória da Influência dos Projetos Pedagógicos dos Cursos Superiores no Resultado do Enade por meio de Mineração de Textos e Aprendizado de Máquina

FACOM
Faculdade de Computação
Universidade Federal de Mato Grosso do Sul



Mestrando: Charles Andre Profilio dos Santos

Orientadora: Profª. Drª. Liana Dessandre Duenha Garanhani

Coorientador: Prof. Dr. Bruno Magalhães Nogueira

Dissertação de mestrado apresentado ao Programa de Pós-Graduação em Ciência da Computação - Mestrado em Ciência da Computação - da Faculdade de Computação da Universidade Federal de Mato Grosso do Sul.

Campo Grande - MS
2022

Agradecimentos

Primeiramente a Deus por todas as bençãos e proteção a mim, minha família e amigos. Aos meus familiares por todo apoio incondicional durante estes últimos 2 anos, pois justamente ao iniciar o Mestrado houve o começo da pandemia do Covid-19, em especial meu irmão, e sua esposa, pela recepção, acolhimento e parceria em 2020 ao me mudar para Campo Grande - (MS), no qual enfrentamos com muita fé, alegria e esperança as etapas da pandemia. Aos meus pais por me apoiarem e terem fé na busca dos meus objetivos, e sempre orando pelo sucesso de seus filhos. Aos meus amigos que, mesmo longes, me davam apoio, alegria, esperanças durante etapas difíceis do mestrado. Aos meus queridos e bondosos orientadores Profa. Liana e Prof. Bruno, por todo o tempo, dedicação e confiança empregados a mim durante esta jornada, claro agradeço imensamente a eles por não terem desistidos da minha pessoa durante empecilhos enfrentados, pois se mudar para outra cidade e estados, enfrentar uma didática diferente do que eu já tinha visto, aprender sobre diversas coisas, não vou negar, demorei a assimilhar tudo que enfrentei, mas em nenhuma momento meus orientadores perderam a fé, são excelentes pessoas e profissionais.

À Faculdade de Computação (FACOM) da Universidade Federal de Mato Grosso do Sul (UFMS) e todos seus profissionais pelo excelentes serviços prestados tanto a comunidade acadêmica, como o público em geral. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Um curso de ensino superior é orientado pelo Projeto Pedagógico do Curso (PPC), que orienta a formação esperada para o egresso do curso, tanto no aspecto profissional quanto humanístico, de acordo com as diretrizes curriculares nacionais vigentes. Para avaliar os cursos de graduação e as instituições de ensino superior, o Ministério da Educação (MEC) utiliza alguns indicadores de qualidade, como o Exame Nacional de Desempenho dos Estudantes (Enade), sendo uma avaliação aplicada a cada três anos aos estudantes egressos de cada curso, que visa avaliar a qualidade do ensino de graduação no país por meio da atribuição de um conceito a cada curso avaliado. Tal conceito e os demais relatórios de avaliação resultantes do Enade auxiliam os gestores das instituições de ensino superior, coordenadores de curso e professores atuarem para a melhoria de seus projetos pedagógicos, infraestrutura física, recursos humanos e demais aspectos que impactem na formação do aluno. Este trabalho propõe uma análise dos projetos pedagógicos de cursos utilizando aprendizado de máquina, para auxiliar na compreensão de como o seu conteúdo impacta na avaliação dos cursos, mais especificamente, nos conceitos Enade Faixa e Enade Contínuo dos cursos. A análise foi aplicada sobre projetos pedagógicos dos cursos de Ciência da Computação e Sistemas de Informação, porém a metodologia é aplicável para outros cursos, mediante replicação do método sobre novos dados de treinamento. Os resultados experimentais demonstraram que é possível predizer o Conceito Enade Faixa com acurácia de $\approx 80\%$ e o Conceito Enade Contínuo com erro percentual absoluto médio de $\approx 11\%$.

Palavras-chave: Aprendizado de Máquina, Mineração de texto, Educação Superior.

Abstract

A higher education course is guided by the Pedagogical Project of the Course (PPC), which suggests the training expected for the graduates, both in the professional and humanistic aspects, according to the current national curriculum guidelines. To evaluate undergraduate courses and higher education institutions, the Ministry of Education (MEC) uses some quality indicators, such as the National Student Performance Exam (Enade), with an assessment applied every three years to students graduating from each course, which aims to assess the quality of undergraduate education in the country by assigning a concept to each evaluated course. This concept and the other evaluation reports resulting from Enade help the managers of higher education institutions, course coordinators and professors to act to improve their pedagogical projects, physical infrastructure, human resources and other aspects that impact student training. This work proposes an analysis of the pedagogical projects of courses using machine learning, to assist in the understanding of how its content impacts the evaluation, more specifically, the concepts Enade Track and Enade Continuous. The analysis was applied to PPC of Computer Science and Information Systems courses, however the the methodology applies to other courses, by replicating the method on new training data. The experimental results showed that it is possible to predict the Enade Range Concept with an accuracy of $\approx 80\%$, and the Enade Continuous concept with an average absolute percentage error of $\approx 11\%$.

Keywords: Machine Learning, Text Mining, Superior Education.

Lista de Figuras

Figura 2.1: Mapa conceitual dos Instrumentos de Avaliação. 26
Figura 2.2: Mapa conceitual do CONAES e ramificações. 28
Figura 2.3: Etapas do método de Mineração de Texto. 34
Figura 2.4: Exemplo de matriz atributo-valor. 36
Figura 2.5: Hierarquia de Aprendizagem de Máquina. 39
Figura 2.6: Exemplo SVM com duas classes. 43
Figura 2.7: Exemplo de SVM em duas ou mais dimensões. 44
Figura 2.8: Exemplo de SVR. 45
Figura 2.9: Exemplo do *k-Means*. 47
Figura 2.10: Exemplos de técnicas de amostragens. 48
Figura 2.11: Exemplo de árvore do método *Top-Down* em CH. . . . 50
Figura 2.12: Exemplo do Corte de Luhn. 52
Figura 2.13: Exemplo Matriz de Confusão Multiclasses. 56

Figura 4.1: Mapa estrutural da formação do projeto prático. . . . 65
Figura 4.2: Distribuição dos PPCs de acordo com o Conceito Enade Faixa. 69
Figura 4.3: Distribuição dos PPCs de acordo com o Conceito Enade Contínuo. 70
Figura 4.4: Exemplo estrutura de relação e dependência dos classificadores. 74
Figura 4.5: Estrutura Simplificada da Classificação. 75
Figura 4.6: Síntese da Estrutura de Treino/Validação da Classificação. 77
Figura 4.7: Estrutura da Árvore de Classificação. 78
Figura 4.8: Estrutura da Árvore de Regressão. 83

Figura 5.1: Resultado da Silhueta no laço de repetição para execução com conjunto completo e somente unigramas. . . 86
Figura 5.2: Agrupamentos gerados pelo *k-Means* utilizando Unigramas em uma representação 2D. 87

Figura 5.3: Matriz Multiclasses Final, referente ao treino/teste da classificação. 92

Figura A.1: Número de Instituições de Educação Superior, por Organização Acadêmica e Categoria Administrativa - Brasil - 2009-2019. 129

Figura A.2: Número de Cursos de Graduação, por Modalidade de Ensino e por Grau Acadêmico - Brasil - 2009-2019. . . 130

Lista de Tabelas

Tabela 2.1: Conceito Enade Contínuo e Conceito Enade Faixa. . . 29
Tabela 2.2: Exemplo de Matriz de Confusão. 56

Tabela 4.1: Distribuição dos PPCs dos cursos de Ciência da Computação e Sistemas de Informação de acordo com a categoria da instituição de ensino e do Conceito Enade Faixa do curso 68
Tabela 4.2: Hiperparâmetros do modelo MLP na execução de *GridSearchCV*, juntamente com a variação da *Stratified K-Fold*. 76
Tabela 4.3: Exemplo sintético da Estrutura gerada ao final da hierarquia de classificação. 79
Tabela 4.4: Exemplo sintético da estrutura gerada ao final da árvore de regressão. 84

Tabela 5.1: Síntese da Tabela Características dos *clusters* gerados pelo *k-Means* 88
Tabela 5.2: Acurácias referente ao treino/teste da classificação. . . 90
Tabela 5.3: *Micro-Avereged Precision e Recall* referente ao treino/-teste da classificação. 91
Tabela 5.4: Médias Finais das métricas Acurácia, *Micro-Avereged Precision e Recall* referente ao treino/teste da classificação. 92
Tabela 5.5: Características dos dados utilizados na Classificação. . 93
Tabela 5.6: Distribuição das classes de IES Privadas e Públicas por região. 95
Tabela 5.7: Médias finais referente ao treino/teste da regressão, contendo unigramas. 96

Tabela A.1: Parâmetros da Classificação com Unigramas, execuções 1 e 2. 115

Tabela A.2: Parâmetros da Classificação com Unigramas, execuções 3 e 4. 116
Tabela A.3: Parâmetros da Classificação com Unigramas, execução 5. 117

Tabela B.1: Modelos e seus Hiperparâmetros na execução de *GridSearchCV*, juntamente com a variação da *k-Fold Cross-Validation*, para a Regressão. 119
Tabela B.2: Parâmetros da Regressão utilizando Unigramas, repetições 1 e 2. 121
Tabela B.3: Parâmetros da Regressão utilizando Unigramas, repetição 3. 122

Tabela C.1: Características dos *clusters* gerados pelo *k-Means* utilizando Unigramas. 124

Tabela D.1: Estimativas médias do modelo MLP em cada execução. 126

Tabela E.1: Estimativas médias dos modelos para Classificador 1 e Regressores 2 e 3, utilizando o conjunto com unigramas. 127

0 Lista de Abreviaturas e Siglas

k-NN *k-Nearest Neighbors*

ABNT *Associação Brasileira de Normas Técnicas*

AM *Aprendizado de Máquina*

CAPES *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*

CC *Ciência da Computação*

CEC *Conceito Enade Contínuo*

CEF *Conceito Enade Faixa*

CEFET *Centro Federal de Educação Tecnológica*

Censup *Censo da Educação Superior*

CES *Câmara de Educação Superior*

CH *Classificação Hierárquica*

CL *Corte de Luhn*

CNCST *Catalogo Nacional de Cursos Superiores de Tecnologia*

CNE *Conselho Nacional de Educação*

CNPq *Conselho Nacional de Desenvolvimento Cientifico e Tecnológico*

CONAES *Comissão Nacional de Avaliação de Educação Superior*

COVID-19 *Corona Vírus Disease 2019*

CPF *Cadastro de Pessoas Físicas*

DA *Data Augmentation*

DCN *Diretrizes Curriculares Nacionais*

DEAES *Diretoria de Estatísticas e Avaliação da Educação Superior*

Dipec *Divisão de Projetos Pedagógicos de Cursos*

EaD *Ensino a Distância*

Enade *Exame Nacional de Desempenho dos Estudantes*

ENC *Exame Nacional de Cursos*

Enem *Exame Nacional do Ensino Médio*

FACOM *Faculdade de Computação*

IA *Inteligência Artificial*

IES *Instituição de Ensino Superior*

IF *Instituto Federal de Educação, Ciência e Tecnologia*

Inep *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*

LDB *Lei de Diretrizes e Bases da Educação Nacional*

MAE *Mean Absolute Error*

MAPE *Mean Absolute Percentage Error*

MEC *Ministério da Educação*

MLP *Multilayer Perceptron*

MLPR *Regressor Perceptron Multicamadas*

MS *Mato Grosso do Sul*

MSE *Mean Squared Error*

PCA *Principal Compenent Analysis*

PDF *Portable Document Format*

PDI *Plano de Desenvolvimento Institucional*

PPC *Projeto Pedagógico do Curso*

PPI *Projeto Pedagógico Institucional*

RBF *Radial Basis Function*

SI *Sistemas de Informação*

SINAES *Sistema Nacional de Avaliação da Educação Superior*

SVC *Support Vector Classification*

SVM *Máquinas de Vetores de Suporte*

SVR *Máquinas de Vetores de Suporte Aplicadas à Regressão*

TF-IDF *Termo de Frequência e Frequência Inversa de Documentos*

UFMS Universidade Federal de Mato Grosso do Sul

Sumário

1 INTRODUÇÃO **16**
1.1 OBJETIVOS 17
1.2 ORGANIZAÇÃO DO TEXTO 18

2 REFERENCIAL TEÓRICO **19**
2.1 INSTRUMENTOS DE GESTÃO EDUCACIONAL 19
2.1.1 Diretrizes Curriculares Nacionais 19
2.1.2 Projeto Pedagógico do Curso 20
2.1.3 Sistema Nacional de Avaliação da Educação Superior . 24
2.1.4 Instrumentos de Avaliação de um Curso de Graduação e sua Instituição 25
2.1.4.1 Projeto Pedagógico Institucional 26
2.1.4.2 Plano de Desenvolvimento Institucional . . . 27
2.1.5 Indicadores de Qualidade da Educação Superior 27
2.1.5.1 Exame Nacional de Desempenho de Estudantes (Enade) 28
2.1.5.2 Indicador de Diferença entre o Desempenho Observado e Esperado 30
2.1.5.3 Conceito Preliminar de Curso 30
2.1.5.4 Índice Geral de Cursos 31
2.1.5.5 Conceito Institucional 31
2.1.6 Censup e Portal e-Mec 32
2.1.7 Considerações Finais 32
2.2 TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL 33
2.2.1 Mineração de Textos 33
2.2.1.1 Conhecimento do Domínio 35
2.2.1.2 Pré-Processamento 35
2.2.1.3 Extração de Padrões 38
2.2.1.4 Pós-Processamento e Uso do Conhecimento . 39
2.2.2 Aprendizado de Máquina 39
2.2.2.1 Aprendizado Supervisionado 40

2.2.2.2 Aprendizado Semissupervisionado 40
2.2.2.3 Aprendizado Não Supervisionado 41
2.2.2.4 Aprendizado por Reforço 41
2.2.2.5 Algoritmos de Aprendizado de Máquina para Análise de Documentos 42
2.2.2.5.1 k-NN ou k-Vizinhos mais Próximos . 42
2.2.2.5.2 Máquinas de Vetores de Suporte . . 43
2.2.2.5.3 Máquinas de Vetores de Suporte para Regressão 44
2.2.2.5.4 *Perceptron* Multicamadas 45
2.2.2.5.5 ***k**-Means* 46
2.2.2.6 Amostragem dos Dados para Estimativas de Desempenho dos Algoritmos de Aprendizado de Máquina 47
2.2.2.6.1 Holdout 48
2.2.2.6.2 Cross-Validation 49
2.2.2.6.2.1 k-Fold Cross-Validation 49
2.2.2.6.2.2 Stratified k-Fold Cross-Validation 50
2.2.2.7 Classificação Hierárquica 50
2.2.2.8 *Data Augmentation* 51
2.2.2.9 Corte de Luhn 52
2.2.2.10 *Grid Search* para Otimização de Hiperparâmetros 53
2.2.2.11 Métricas para avaliação dos Modelos 53
2.2.2.11.1 Métricas para avaliação de modelos de regressão 54
2.2.2.11.1.1 Erro Médio Quadrático 54
2.2.2.11.1.2 Erro Médio Absoluto 54
2.2.2.11.1.3 Erro Percentual Absoluto Médio 55
2.2.2.11.2 Métricas para Modelos de Classificação 55
2.2.2.11.2.1 *Acurácia* 57
2.2.2.11.2.2 *Micro-Avereged* 57
2.2.3 Considerações Finais 58

3 TRABALHOS RELACIONADOS **59**
3.1 ANÁLISES DE DOCUMENTOS DE GESTÃO E AVALIAÇÃO EDUCACIONAL 59

3.2 APRENDIZADO DE MÁQUINA APLICADO À GESTÃO DO ENSINO DE GRADUAÇÃO 62
3.3 CONSIDERAÇÕES FINAIS 64

4 AVALIAÇÃO EXPERIMENTAL 65
4.1 O CONJUNTO DE DADOS 66
4.2 METODOLOGIA DE PRÉ-PROCESSAMENTO, PARAMETRIZAÇÃO E AVALIAÇÃO PARA AGRUPAMENTO 71
4.2.1 Pré-processamento para Agrupamento 71
4.2.2 Agrupamento dos PPCs 71
4.3 METODOLOGIA EXPERIMENTAL PARA CLASSIFICAÇÃO 72
4.3.1 Pré-processamentos para Classificação dos Dados . . . 72
4.3.2 Classificação dos PPCs de acordo com o Conceito Enade (Faixa) 73
4.4 METODOLOGIA DE PRÉ-PROCESSAMENTO, PARAMETRIZAÇÃO E AVALIAÇÃO PARA REGRESSÃO 80
4.4.1 Pré-Processamentos para Regressão 80
4.4.2 Modelos de Regressão para a Predição do Conceito Enade Contínuo a partir dos PPCs 80
4.5 Considerações Finais 84

5 ANÁLISE DOS RESULTADOS EXPERIMENTAIS 85
5.1 AVALIAÇÃO DO MODELO DE AGRUPAMENTO DOS PPC'S 85
5.2 AVALIAÇÃO DOS MODELOS DE CLASSIFICAÇÃO DOS PPC'S 89
5.3 AVALIAÇÃO DOS MODELOS DE PREDIÇÃO DO CONCEITO ENADE CONTÍNUO 95
5.4 CONSIDERAÇÕES FINAIS 97

6 CONSIDERAÇÕES FINAIS 98

REFERÊNCIAS BIBLIOGRÁFICAS 102

APÊNDICES 113

APÊNDICE A MODELOS E SEUS PARÂMETROS DE CLASSIFICAÇÃO 114

APÊNDICE B HIPERPARÂMETROS GRIDSEARCH E MELHORES PARÂMETROS DA REGRESSÃO 118
B.1 Parâmetros Candidatos dos Modelos na Execução do *GridSearch* 118
B.2 Modelos e seus Melhores Parâmetros Regressão 120

APÊNDICE C CARACTERÍSTICAS DOS AGRUPAMENTOS GERADAS PELO K-MEANS 123

APÊNDICE D ESTIMATIVAS GERADAS PELOS MODELOS DURANTE TREINO/VALIDAÇÃO NA CLASSIFICAÇÃO 125

APÊNDICE E ESTIMATIVAS GERADAS PELOS MODELOS DURANTE TREINO/VALIDAÇÃO NA REGRESSÃO 127

ANEXOS 128

ANEXO A NÚMEROS DO CENSUP 129

1 INTRODUÇÃO

O Conselho Nacional de Educação (CNE) é responsável por assessorar o Ministério da Educação (MEC) no desempenho das suas funções, cabendo-lhe deliberar, normatizar e avaliar as políticas nacionais da pasta, zelar pela qualidade do ensino, velar pelo cumprimento da legislação educacional e assegurar a participação da sociedade no aprimoramento da educação brasileira [1]. O CNE, juntamente com a Câmara de Educação Superior (CES), estabelecem as diretrizes curriculares dos cursos de graduação que norteiam a criação e aperfeiçoamento dos cursos superiores dentro território nacional. Não é objetivo do CNE e CES o engessamento dos programas e propostas pedagógicas das instituições de ensino superior; portanto, as diretrizes garantem à flexibilidade e diversidade nos programas oferecidos pelas diferentes instituições de ensino superior, de forma a melhor atender suas clientelas e peculiaridades regionais [2]. Todas as diretrizes curriculares dos cursos de graduação publicados desde 2001 pelo CNE/CES podem ser encontradas em [3].

Um curso de graduação tem como seu principal documento norteador o Projeto Pedagógico do Curso (PPC), o qual define o currículo acadêmico mediante um conjunto de ações de ensino, pesquisa e extensão, além de diversos aspectos institucionais, de infraestrutura, administrativos e sociais que conduzem a formação do egresso. O PPC define o caminho a ser trilhado pelo aluno durante a graduação e impacta diretamente na qualidade da sua formação e no seu desempenho após formado. Ainda assim, são poucos os trabalhos que utilizam o PPC como uma fonte primária de dados e insumo para análises. A proposta pedagógica dos PPCs é embasada nas diretrizes curriculares nacionais vigentes definidas pelo CNE/CES.

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), por meio da Comissão Nacional de Avaliação de Educação Superior (CONAES) juntamente com a Diretoria de Estatísticas e Avaliação da Educação Superior (DEAES), utiliza o SINAES (Sistema Nacional de Avaliação da Educação Superior) para a construção de um sistema nacional de avaliação que articula regulação e avaliação educativa. O SINAES emprega alguns indicadores para avaliação dos cursos de graduação, como o Exame Nacional de Desempenho dos Estudantes (Enade) [4]. Criado em 2004, o Enade tem o objetivo de avaliar o rendimento dos concluintes dos cursos de graduação em relação aos conteúdos programáticos previstos nas diretrizes

curriculares dos cursos. Os resultados do Enade, aliados às respostas do Questionário do Estudante, são insumos para o cálculo dos Indicadores de Qualidade da Educação Superior.

A análise manual de projetos pedagógicos é inviável devido a grande quantidade de dados a serem manipulados, motivando o uso de técnicas que permitam a automatização desse processo. Neste contexto, a *Mineração de textos* surge como uma ferramenta poderosa na gestão do conhecimento, a qual pode ser definida como um conjunto de técnicas e processos que buscam padrões, tendências e regularidades em documentos escritos em linguagem natural, com o objetivo de descobrir conhecimento nesses textos [5]. Da mesma forma, algoritmos de *Aprendizado de Máquina* também podem ser adequados a esta tarefa, pois permitem uma análise semiautomática de grandes volumes de dados textuais, extraindo conhecimento potencialmente útil para embasar potenciais conclusões sobre o domínio dos dados avaliados.

1.1 OBJETIVOS

O objetivo geral deste trabalho é a aplicação de algoritmos de mineração de texto e aprendizado de máquina auxiliar na análise de como os projetos pedagógicos de cursos superiores influenciam no desempenho do estudante no Enade. Os resultados experimentais são úteis para gestores de IES e coordenadores de curso, como apoio inteligente à tomada de decisões pedagógicas, durante o processo de gestão e acompanhamento dos cursos. Embora os experimentos tenham sido realizados sobre projetos pedagógicos dos cursos de Ciência da Computação e Sistemas de Informação, que são os cursos da área de computação com maior oferta nas IES, a metodologia é aplicável para outros cursos, mediante replicação do método sobre novos dados de treinamento.

As principais contribuições deste trabalho são:

1. A elaboração de um conjunto de dados, formado por 223 projetos pedagógicos dos cursos de Sistemas de Informação e Ciência da Computação de 123 Instituições de ensino superior brasileiras, públicas e privadas.
2. Um modelo de agrupamento dos projetos pedagógicos dos cursos, com relação às similaridades, com objetivo de identificar quais características influenciam na formação de grupos.

3. Modelos de classificação para predizer o conceito Enade Faixa de de cursos, tendo com base os seus projetos pedagógicos.

4. Modelos para regressão para predizer o Conceito Enade Contínuo dos cursos, tendo como base os seus projetos pedagógicos.

1.2 ORGANIZAÇÃO DO TEXTO

O presente trabalho está dividido capítulos e seções, as quais possuem o objetivo de apresentar e explicar adequadamente cada fase do projeto.

- O Capítulo 2 contém o Referencial Teórico do projeto dividido em duas seções principais. A Seção 2.1 fornece o referencial teórico sobre instrumentos de gestão educacional, descrevendo os documentos e instrumentos de concepção, normatização e avaliação de cursos, bem como os órgãos governamentais a que correspondem estas atividades. A Seção 2.2 fornece os conceitos necessários e descreve técnicas e algoritmos que serão utilizados durante o desenvolvimento do projeto.

- O Capítulo 3 contém trabalhos prévios que abordam a análise, automatizada ou não, de documentos de gestão educacional para melhoria da qualidade do ensino de graduação.

- O Capítulo 4 contém a descrição dos experimentos realizados, incluindo a metodologia utilizada na obtenção do conjunto de dados e sua formação para cada tipo de aprendizado de máquina e a parametrização dos modelos.

- O Capítulo 5 contém os resultados finais e análises dos modelos de clusterização, classificação e regressão implementados.

- O Capítulo 6 conclui esta dissertação, com a análise de todo o trabalho desenvolvido, apresentação de dificuldades e trabalhos futuros.

2 REFERENCIAL TEÓRICO

Considerando que esse projeto é multidisciplinar, tendo a área de gestão educacional como parte do domínio do problema e a área de computação como parte do domínio da solução, dividiu-se o Referencial Teórico em duas seções majoritárias para facilitar a organização desta dissertação.

2.1 INSTRUMENTOS DE GESTÃO EDUCACIONAL

Nesta Seção, serão abordados os principais instrumentos de gestão educacional dos cursos de graduação, desde sua concepção, diretrizes e entidades regulamentadoras, até a avaliação dos egressos, cursos e instituições.

2.1.1 Diretrizes Curriculares Nacionais

A atual organização do ensino superior se iniciou em meados da década de 90, para ser mais preciso em 1995, por meio da Lei nº 9.131 [6]. A Lei em questão instituía o CNE, ficando responsável por realizar rupturas na estrutura jurídica da educação brasileira e, consequentemente, dando origem a CES. Em 1996 houve a instauração da Lei nº 9.394 [7] ou Lei de Diretrizes e Bases da Educação Nacional (LDB), no qual caracterizava a descentralização do MEC da figura de gestor dos cursos e tornava-se o promotor, deste modo, o governo apenas fomenta e dá as diretrizes, mas não atua diretamente na gestão das universidades [8, 9].

A proposta pela substituição dos currículos mínimos pelas DCNs (Diretrizes Curriculares Nacionais) ocorreu em 1997 por meio do Parecer 776 [10], no qual o CNE atribuía a CES a responsabilidade de dispor sobre as DCNs propostas pelo MEC. Os objetivos para a formação das DCNs eram: assegurar a liberdade das universidades, possibilitando a elas delimitar tempo e conteúdo; fazer com que as universidades pensem além da sala de aula; realizar o aprendizado teórico alinhado ao prático; estimular o aprendizado do aluno, criando assim ótimos profissionais; fazer com que a universidade faça uma autoavaliação sobre toda sua organização [8, 10].

A partir da publicação do Parecer CNE/CES nº 583 [2], aprovado em

04/04/2001, determinou-se que a definição da duração, carga horária e tempo de integralização dos cursos será objeto de um parecer e/ou uma resolução específica da Câmara de Educação Superior. Além disso, determinou que as DCNs dos cursos de graduação devem conter, obrigatoriamente:

1. Perfil do egresso ou profissional – conforme o curso, o projeto pedagógico deverá orientar o currículo para um perfil profissional desejado.
2. Competências, habilidades, atitudes.
3. Habilitações e ênfases.
4. Conteúdos curriculares.
5. Organização do curso.
6. Estágios e atividades complementares.
7. Acompanhamento e avaliação.

Ao longo dos anos as DCNs sofreram diversas alterações e atualizações, no qual pode-se averiguar todas elas, bem como, os pareceres de aprovação de cada curso no portal do MEC [11].

2.1.2 Projeto Pedagógico do Curso

O PPC é o documento que norteia a concepção de um curso de graduação e é o orientador de todas as decisões e ações de um curso. Um PPC tem o objetivo de delinear a identidade formativa do egresso, tanto na esfera humana quanto profissional, apresentar as concepções do curso juntamente com as orientações pedagógicas, contendo a matriz curricular e toda a estrutura acadêmica para sua aplicação [12, 13, 14].

O PPC é construído com base em atender às Diretrizes Nacionais Curriculares do curso. A fim de cumprir todas as exigências presentes nas DCN e apresenta alguns componentes comuns, como: Concepção do Curso; Estrutura do Curso, Procedimentos de Avaliação dos Processos de Ensino e Aprendizagem; Instrumentos Normativos de apoio. A estrutura de um PPC varia dependendo de cada curso e instituição, sendo que sua elaboração é complexa e demorada. Para tal, a seguir elaborou-se uma estrutura com os

principais tópicos que devem constar em um PPC, embasando-se no documento orientador do Inep [15] e um guia [16] da UFMS (Universidade Federal de Mato Grosso do Sul) para a elaboração dos PPCs de seus campus. Ressaltando que cada curso possui suas particularidades devido as exigências das DCNs, deste modo, para fins de conhecimento de tais peculiaridades deve-se acessar o portal do MEC [11] e averiguá-las.

1. Identificação do Curso

   Seção destinada a expor informações com relação ao curso, como: Denominação do Curso; Código e-MEC; Habilitação; Grau Acadêmico Conferido; Modalidade de Ensino; Regime de Matrícula; Tempo de Duração; Carga Horária Mínima; Número de Vagas Ofertas por Ingresso ou Número Médio de Vagas por Polo; Número de Entradas; Turno de Funcionamento ou Modelo de Funcionamento; Local de Funcionamento; Forma de Ingresso.

2. Fundamentação Legal

   Esta seção se refere a todas legislações, normas, leis, etc, presentes nas DCNs, que regem o PPC. A instituição deve apenas acrescentar as normativas e legislações especificas do curso em questão, disponibilizado pelo CNE e pela IES.

3. Contextualização

   Seção destinada a apresentar a instituição de ensino, a história da formação do curso em questão, alguns subtópicos são comuns encontrar, como: Histórico da Instituição, Histórico da Unidade de Administração Setorial de Lotação do Curso, Histórico do Curso.

4. Necessidade Social do Curso

   Esta seção apresenta diversas informações das demandas regionais pelo curso para, assim, delimitar a real necessidade da mesorregião e sua demanda por profissionais destinados ao mercado de trabalho ou com conhecimentos teóricos avançados. Para elaboração deste Item, alguns subtópicos são essenciais, como: Indicadores Socioeconômicos da População da Mesorregião; Indicadores Socioambientais da Região; Análise da Oferta do Curso na Região.

5. Concepção do Curso

Seção destinada a respaldar toda a elaboração, ou formação, do curso. Cada instituição elabora esta seção com a finalidade de garantir a qualidade de ensino, bem como seguindo os princípios presentes nos Indicadores do Inep [15], ressaltando que alguns dos indicadores acima são específicos para alguns cursos. Subtópicos presentes nesta seção: Dimensões Formativas, incluindo Técnica, Política, Desenvolvimento Pessoal, Cultural, Ética, Social; Estratégias para o Desenvolvimento de Ações Interdisciplinares; Estratégias para Integração das Diferentes Componentes Curriculares; Perfil Desejado do Egresso; Objetivos; Metodologias de Ensino; Avaliação.

6. Administração Acadêmica do Curso

   Seção destinada a expor a maneira como será administrado o curso, englobando discentes, docentes, coordenação, ensino, estrutura, etc. Ramificações como Atribuições do Colegiado de Curso, Atribuições do Núcleo Docente Estruturante, perfil da Coordenação do Curso, Organização Acadêmico-Administrativa, Atenção aos Discentes, auxiliam na construção desta etapa.

7. Currículo

   Esta seção expõe a estrutura curricular respeitando o Indicadores do Inep [15], os quais exigem contemplar a interdisciplinaridade, flexibilidade, a acessibilidade metodológica, a compatibilidade da carga horária total, demonstrar a conexão entre teoria e a prática, a disponibilidade de disciplina de Libras e estrutura para a modalidade a distância, entre outros pontos. Alguns subtópicos auxiliam a elaboração desta etapa, como: matriz curricular do curso; quadro de semestralização; tabela de equivalência das disciplinas; lotação das disciplinas nas unidades da administração setorial; ementário; bibliografia básica e complementar; política de implantação da nova matriz curricular.

8. Políticas

   Engloba algumas politicas padrões de cada universidade, apresentando como exemplo a Capacitação do Corpo Docente, Inclusão de Pessoas com Deficiência, Inclusão de Cotistas e Atendimento aos Requisitos Legais e Normativos: Relações Étnico-Raciais, Direitos Humanos e Educação Ambiental.

9. Sistema de Avaliação

Seção para descrever, não somente os métodos de avaliação do estudante, mas também os procedimentos para autoavaliar a instituição e do curso. As seguintes ramificações podem ser encontradas neste item: Sistema de Avaliação do Processo Formativo; Sistema de Autoavaliação do Curso; Participação do Corpo Discente na Avaliação do Curso; Projeto Institucional de Monitoramento e Avaliação do Curso;

10. Atividades Acadêmicas Articuladas ao Ensino de Graduação

    Esta seção contempla as atividades acadêmicas vinculadas ao ensino de graduação, no qual o objetivo é garantir a junção entre o ensino teórico com o aprendizado pratico. A fim de assegurar o objetivo, essa seção possui algumas ramificações, no qual cada instituições deverá analisar as mais adequadas e possíveis para sua realidade. As ramificações são: atividades orientadas de ensino; atividades complementares; atividades de extensão; atividades obrigatórias; estágio obrigatório e não obrigatório; natureza do estágio; participação do corpo discente nas atividades acadêmicas; prática de ensino; trabalho de conclusão do curso. Todas essas ramificações estão presentes na Dimensão 1 do Inep [15], ressaltando que para alguns cursos as DCN exigem a presença de algumas.

11. Desenvolvimento de Materiais Didáticos

    Esta seção é opcional para cursos presencias, contudo para a modalidade EaD (Ensino a Distância) é obrigatória. Essa seção é elaborada pela Dipec (Divisão de Projetos Pedagógicos de Cursos) de cada instituição, a fim de respaldar todo o ensino a distância com materiais, docentes preparados, ambiente virtual adequados, entre outros.

12. Infraestrutura Necessária ao Curso

    Neste tópico, a instituição deve descrever todo o ambiente de ensino, como mobiliário, espaços, equipamentos e instrumentos da tecnologia da informação e comunicação. Deve-se descrever corretamente cada espaço presente na instituição, seguindo a Dimensão 3 do Inep [15], como salas de aula, dos professores, coordenação, biblioteca, entre outros.

13. Plano de Incorporação dos Avanços Tecnológicos ao Ensino de Graduação

    Apresenta as estratégias da instituição para incorporar o ensino aos novos meios tecnológicos, essa normativa é prevista no Indicador 1.16 do Inep [15].

14. Considerações Finais

    Este tópico deve contemplar uma síntese de todo o PPC, ou em outros termos, um resumo resoluto de toda proposta de formação corporizada no PPC.

15. Referências

    Apresenta as bibliografias aplicadas na concepção do PPC, seguindo todas as normas da ABNT (Associação Brasileira de Normas Técnicas). Observação, não citar neste item legislações presentes no Item 2 - Fundamentação Legal.

16. Anexos

    Este item engloba todos os documentos ou textos que não foram elaborados pelos autores, a exemplo: Resoluções de aprovações de Biotério, laboratórios, entre outros; Documentos de convênios firmados para estágio, utilização de sistemas locais e regime de saúde;

17. Apêndices

    Apresenta todos os documentos e textos elaborados pelos autores, por exemplo: Resoluções de aprovação para Estágio Obrigatório e Não-Obrigatório, modelo de Trabalho de Conclusão do Curso, Atividades Complementares, Atividade Orientada de Ensino, entre outras.

2.1.3 Sistema Nacional de Avaliação da Educação Superior

Em 2004, por recurso da Lei nº 10.861 [17], é instituído o SINAES (Sistema Nacional de Avaliação da Educação Superior) instrumento responsável por avaliar as instituições de ensino superior e seus cursos ofertados. Para tal, a supervisão e coordenação do SINAES é de responsabilidade da CONAES (Comissão Nacional de Avaliação da Educação Superior) em junção da DEAES (Diretoria de Estatísticas e Avaliação da Educação Superior), no qual são mantidos pelo Inep (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) em respaldo ao MEC [12, 18].

O SINAES [12, 19] foi criado mediante a Lei nº 10.861, com o intuito de avaliar os cursos de graduação e suas instituições. A avaliação das instituições, dos cursos e do desempenho dos estudantes, são os três pilares do

SINAES. Todos os aspectos que giram ao redor destes três princípios são avaliados por este sistema, sendo destaque: o ensino, a extensão, a pesquisa, o desempenho dos alunos, a responsabilidade social, o corpo docente, a gestão da instituição e as instalações.

O intuito do SINAES é: melhorar o mérito e o prestígio das instituições, áreas, cursos e programas, nos âmbitos de ensino, extensão, pesquisa, formação e gestão; melhorar o nível da educação superior e direcionar a expansão da oferta, ademais de promover a responsabilidade social das IES, prezando a identidade institucional e a liberdade de cada entidade.

Este sistema possui diversos mecanismo extras, como uma autoavaliação, avaliação externa, Enade, avaliação dos cursos de graduação e ferramentas de informação como o censo e o cadastro. A associação destes mecanismos permite a atribuição de conceitos, que variam de zero a cinco, a cada um dos âmbitos e ao conjunto de âmbitos avaliados. Sendo todos os resultado públicos e disponíveis no portal do Inep [20] de cada instituição e seus cursos.

2.1.4 Instrumentos de Avaliação de um Curso de Graduação e sua Instituição

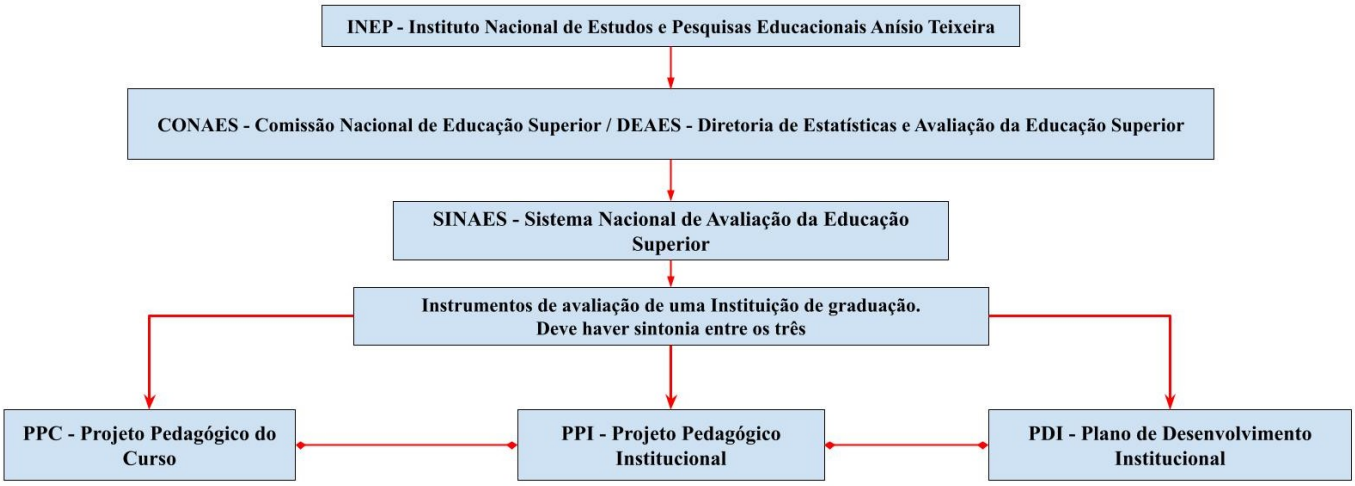
MEC [12] salienta que a avaliação da graduação acadêmica e profissional é compreendida como uma atividade composta que possibilita a captação da qualidade do curso na realidade da instituição, com o intuito de formar pessoas conscientes e profissionais responsáveis, possibilitando assim, realizarem transformações sociais.

Deste modo, o MEC juntamente com Inep [12, 15], criaram um instrumento de avaliação dos cursos de graduação, com o objetivo de obter informações com caráter qualitativo e quantitativo, alinhado com os dados coletados na avaliação presencial de especialistas, que ao final do processo irá gerar um conceito para a instituição.

Avaliar um curso de graduação no seu contexto institucional, envolve não somente o PPC, mas também o PPI (Projeto Pedagógico Institucional) e o PDI (Plano de Desenvolvimento Institucional). Esses três projetos devem apresentar sintonia entre eles, em outras palavras, dever haver sintonia entre o PPC e os objetivos da instituição (PPI e PDI).

Tais documentos, além de terem caráter técnico-burocrático, são ferramentas de ação politica e pedagógica, com o intuito de promover uma formação de qualidade. E ainda devem explicitar o posicionamento sobre sociedade, a educação e garantir o cumprimento das ações e politicas da instituição. Para avaliar a articulação entre o PPC, PPI e PDI, deve-se respeitar as características da presente instituição, bem como as peculiaridades de sua região. Na Figura 2.1 é apresentado a hierárquica de avaliação do PPC, PPI e PDI, bem como os órgãos responsáveis.

Figura 2.1: Mapa conceitual dos Instrumentos de Avaliação.



Fonte: Elaborado pelo autor (2022).

2.1.4.1 Projeto Pedagógico Institucional

Em sua sigla PPI [12, 13, 21], é um documento teórico-metodológico de suma importância, no qual define as politicas para a organização administrativa e também pedagógica para uma universidade. Este documento norteá as ações voltadas para a obtenção das missões e objetivos da instituição.

A elaboração deste documento necessita de uma discussão sobre alguns pontos em questão, como: análise sobre o mundo atual e o papel da instituição em épocas de mudanças globais e tecnológicas; o ensino, extensão e a pesquisa como componentes essências para a formação de pessoas com caráter profissional e humano; a produção e a socialização dos conhecimentos na procura pela junção entre a conjuntura real e desejada pela diferentes frentes administrativas, conceituais e pedagógicas.

2.1.4.2 Plano de Desenvolvimento Institucional

Ou PDI [12, 13], em sua abreviatura, é o documento para o planejamento e gestão que conceitua a identidade da instituição, no que se refere à sua filosofia de trabalho, à sua missão a que se compromete, as diretrizes pedagógicas que norteiam suas ações, seu sistema organizacional e às tarefas acadêmicas e científicas que elabora ou que pretende elaborar.

Para tal, qualquer PDI deve apresentar em sua estrutura alguns tópicos, sendo eles:

- Perfil institucional.
- Gestão institucional, englobando a organização administrativa, gestão pessoal e politica de atendimento aos discentes.
- Organização acadêmica, contendo a organização didático-pedagógica, disponibilidade de cursos e programas presenciais e a distância.
- Infra-estrutura.
- Aspectos financeiros, orçamentários, avaliação e acompanhamento do desenvolvimento da instituição.

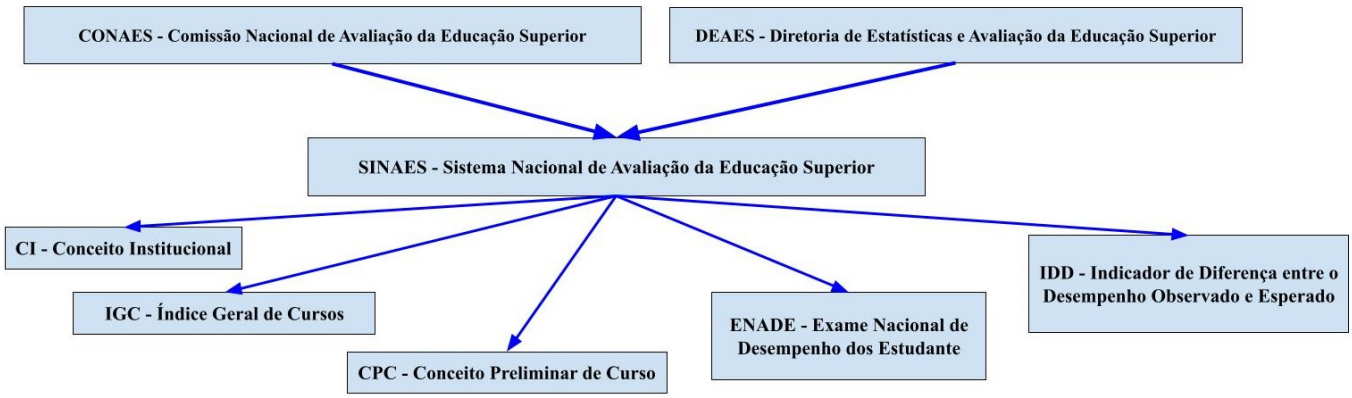
2.1.5 Indicadores de Qualidade da Educação Superior

Os Indicadores de Qualidade [22] são grandes ferramentas para avaliação da educação superior, são expressos em escala continua e com notas de 1 a 5 e possuem relação direta com o Ciclo Avaliativo do Enade. Ciclo Avaliativo do Enade corresponde a uma avaliação periódica dos cursos de graduação, sendo aplicada a cada três anos. Este ciclo é reparticionado em áreas de conhecimento, no qual cada grupo e eixo tecnológico é apresentado na lista abaixo.

Os agrupamentos de cursos de bacharelado e licenciatura derivam da tabela de áreas do conhecimento disponibilizado pelo Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) e os eixos tecnológicos são de referencias do MEC, disponível no Catalogo Nacional de Cursos Superiores de Tecnologia (CNCST).

- **Ano I**: Cursos de bacharelado nas áreas de conhecimento de Ciências Agrárias, Ciências da Saúde, Engenharias, Arquitetura e Urbanismo, cursos na área de Ambiente e Saúde, Produção Alimentícia, Recursos Naturais, Militar e de Segurança.
- **Ano II**: Cursos de bacharelado nas áreas de conhecimento de Ciências Biológicas, Ciências exatas e da Terra, Linguística, Letras, Artes, Ciências Humanas e Ciências da Saúde (Licenciaturas), Controle e Processos Industriais, Informação e Comunicação, Infraestrutura e Produção Industrial.
- **Ano III**: Cursos de bacharelado nas áreas de conhecimento de Ciências Sociais Aplicadas, Gestão e Negócios, Apoio Escolar, Hospitalidade e Lazer, Produção Cultural e Design.

Figura 2.2: Mapa conceitual do CONAES e ramificações.



Fonte: Elaborado pelo autor (2022).

Na Figura 2.2, apresenta-se a estrutura hierárquica dos indicadores de avaliação, antemão os órgãos e o SINAES foram fundamentados neste capítulo. O SINAES utiliza alguns Indicadores para conceitualizar uma instituição, sendo eles: o Enade, IDD, CPC, IGC e CI.

2.1.5.1 Exame Nacional de Desempenho de Estudantes (Enade)

Conhecido popularmente por via de sua abreviatura, o Enade é um dos indicadores mais popular do SINAES. Instituído mediante a Lei nº 10.861 [17], este exame substituiu o antigo Exame Nacional de Cursos (ENC), vulgo Provão, que foi aplicado entre 1996 a 2003 atendendo a Lei

nº 9.131/1995 [6]. O Enade [23] avalia os cursos e suas instituições a fim de exibir o desenvolvimento do ensino dos estudante ingressantes e egressos, em relação aos conteúdos previstos nas DCNs.

O Enade ocorre a cada três anos para cada área de estudo (apresentada na Seção 2.1.5), ao final do processo a instituição recebe o Conceito Enade, sendo calculado por via da média das notas dos alunos participantes e os questionários de avaliação preenchidos por eles. As notas variam de 1 a 5, no qual 1 e 2 indicam resultados insatisfatórios e uma necessidade de mudanças, para nota 3 é um valor mediano e, por fim, notas 4 e 5 indicam uma excelência de ensino [23].

Este cálculo sobre a média dos alunos, para ser mais preciso, considera as informações em relação ao desempenho dos alunos nos exame de formação geral e específica, juntamente com o quantidade de estudantes que participaram da prova de cada curso por instituição. Deste modo, o componente de conhecido específico possui um peso de 75% em relação a nota final e a formação geral apenas 25%. O Inep estabelece que todas as médias devem ser padronizadas para um produto mais justo do conceito, assim, considera-se o desempenho médio nacional e o desvio padrão, no qual é calculado a partir dos resultados das IES participantes. O conceito final do Enade é disponibilizado apenas para as IES que tiveram dois ou mais alunos participantes.

Ao final do processo, são gerados duas versões do Conceito Enade para o curso, sendo que diversos valores Conceito Enade Contínuo são associados ao mesmo Conceito Enade Faixa, como mostrado na Tabela 2.1 e apresentados a seguir.

Tabela 2.1: Conceito Enade Contínuo e Conceito Enade Faixa.

Conceito Enade Contínuo	Conceito Enade Faixa
De 0 a 0,94	1
De 0,95 a 1,94	2
De 1,95 a 2,94	3
De 2,95 a 3,94	4
De 3,95 a 5,0	5

Fonte: Elaborado pelo autor (2022).[1]

- **Conceito Enade Contínuo**: média sobre todos os elementos avaliados, sendo está a nota final de cada IES, no qual seus valores variam de 0 à 5 com valores flutuantes.

[1]Baseado a partir da Nota Técnica Nº 5/2020/CGCQES/DAES [24].

- **Conceito Enade Faixa**: é o valor arredondado do Conceito Enade Contínuo, sendo este o conceito principal utilizado pelas IES na divulgação, com valores inteiros de 0 à 5.

O Inep disponibiliza em seu portal [25, 26], diversas fontes a respeito do processo de avaliação, como: Notas técnicas do Conceito Enade, de vários anos; Valores do indicador para cada curso, em cada edição; Provas, Gabaritos, Questionários e Resultados.

2.1.5.2 Indicador de Diferença entre o Desempenho Observado e Esperado

O IDD [27] é um dos indicadores de qualidade que expõe um valor agregado recebido pelo curso referente ao desenvolvimento dos alunos egressos, no qual é considerado o desempenho destes alunos no Enade e no Exame Nacional do Ensino Médio (Enem), como uma métrica de aproximação de seu conhecimento e aptidões desenvolvidas ao ingressar na graduação.

Em outras palavras, na avaliação do IDD o resultados expressa a diferença entre o desempenho médio do egressos em relação aos resultados médios de outras instituições, no qual há semelhanças entre seus participantes. Apresenta resultado com notas de 1 a 5, todos as demais informações estão disponíveis no portal do Inep [28].

2.1.5.3 Conceito Preliminar de Curso

Ou CPC [29, 30], é um indicador de qualidade que avalia o curso em notas de 1 a 5, no qual combina diferentes conceitos relativos aos cursos de graduação em um só conceito. É formado por oito elementos, delimitados em quatro esferas no intuito de avaliar a qualidade dos cursos. As esferas são:

I - Desempenho dos Estudantes: utiliza a nota referente ao Enade.

II - Valor Agregado pelo Processo Formativo Oferecido pelo Curso: nota referente ao IDD.

III - Corpo Docente: embasado em dados adquiridos por meio do Censo da Educação Superior, aplicado no mesmo ano do Enade, em relação a titulação e regime de trabalhos dos professores associados aos cursos avaliados.

IV - Percepção Discente sobre as Condições do Processo Formativo: é realizado um levantamento, mediante ao Questionário do Estudante, em busca de informações referentes a organização didática-pedagógica, à infraestrutura física da instituição e novas oportunidades de ampliação do ensino acadêmico e profissional.

2.1.5.4 Índice Geral de Cursos

O IGC [31, 32] é um indicador com o intuito de avaliar a instituição e, para tal, sua nota é calculada utilizando a média dos conceitos CPC de seus cursos de graduação e a média dos conceitos obtidos na avaliação de seus programas de pós-graduação (Mestrado e Doutorado) aplicado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a média de distribuição dos alunos entre os diferentes tipo de níveis de ensino. O IGC apresenta suas notas entre 1 a 5.

2.1.5.5 Conceito Institucional

O CI [33, 34], nada mais é, que uma avaliação *in loco* realizada por especialista do MEC/Inep, seguindo a recomendação do documento [15]. O objetivo da visita dos especialista à instituição é a analise sobre os Instrumentos de Avaliação (PDI, PPI e PPC), a gestão, as politicas de ensino (graduação, pós-graduação, pesquisa e extensão), a estrutura da instituição, acessibilidade, departamentos (sala dos professores, de aula, secretária, etc). Também é averiguado a autoavaliação da própria instituição, sendo, geralmente, aplicada pela Comissão Própria de Avaliação. Deste modo, após a visita é disponibilizado a nota CI, que varia de 1 a 5.

2.1.6 Censup e Portal e-Mec

Vale ressaltar toda grandiosidade da esfera do ensino superior, contendo uma gama de instituições e diversidades de cursos de graduação. Para tal, anualmente, o MEC e o Inep realizam o Censo da Educação Superior (Censup) [35], com objetivo de coletar informações estatísticas confiáveis sobre os cursos, discentes, docentes, demandas por ofertas de vagas nas instituições, grau de evasão, entre outros. O último Censup, foi disponibilizado em 2019, porém devido a grave pandemia da COVID-19, no ano de 2020, o sistema educacional foi afetado drasticamente, se recuperando lentamente no ano de 2021. Há projeções que entre os anos de 2021 e 2022, novos dados do Censup sejam disponibilizados.

O Censup do ano de 2019 contabilizou um total de 2.608 Instituições de Ensino Superior (IES), distribuídas da seguinte forma: 108 Universidades públicas e 90 privadas; 11 Centros Universitários públicos e 283 privados; 143 Faculdades públicas e 1.933 privadas; e 40 Institutos Federais (IF) e Centros Federais de Educação Tecnológica (CEFET). Em relação aos cursos de graduação, em 2019 contabilizou-se um total de 40.427 cursos entre bacharelados, licenciaturas e tecnólogos, nas modalidades presencial ou a distância. O histórico, toda a distribuição e evolução de 2009 à 2019, referente as instituições pode ser analisado no Anexo A.1 e em relação aos cursos, no Anexo A.2. A análise completa do MEC/Inep, encontra-se disponível em [36].

Outra fonte de informações sobre IES e cursos é o portal do E-Mec [37], que disponibiliza dados como quantidade de cursos, os indicadores de qualidade atualizados e outras informações pertinentes. Um dos objetivos do portal é fazer cumprir a Lei de Acesso à Informação [38], a qual indica a obrigatoriedade dos órgãos públicos disponibilizarem suas informações de forma simples, direta e atualizada, para toda a população. Deste modo, o portal supre informações desatualizadas do Censup, como novas IES, novos cursos ou cursos extintos.

2.1.7 Considerações Finais

Nesta subseção, foram definidos diversos instrumentos para gestão educacional no âmbito dos cursos de graduação. Em resumo, foram apresentados os órgãos responsáveis por normatizar, instruir, fomentar e fazer a gestão de toda essa grande esfera da educação no Brasil. A fundamentação sobre as

DCNs, os instrumentos gerados a partir dela, como os PPCs dos cursos nas Instituições de Ensino, os sistemas de avaliação como o SINAES e os seus instrumentos e/ou indicadores de avaliação de um curso e/ou instituição de ensino, como são o Enade, IDD, CPC, IGC e CI. Além disso, também foram descritos as principais fontes oficiais de informação sobre instituições de ensino, cursos e indicadores de qualidade.

2.2 TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL

Esta seção majoritária aborda os conceitos essenciais de mineração de textos, juntamente com modelos de aprendizado de máquinas necessários para compreensão do projeto. Na Seção 2.2.1 é contextualizada mineração de textos; Seção 2.2.2 define aprendizado de máquina e suas subáreas; Seção 2.2.2.5 apresenta alguns algoritmos populares e utilizados neste projeto; Seção 2.2.2.6 contêm métodos de amostragem de dado, no qual possibilita a produção de resultados confiáveis e a generalização correta do modelo; Seção 2.2.2.8 discorre sobre *Data Augmentation* e alguns de suas técnicas para aumento do conjunto de dados; Seção 2.2.2.9 discute uma técnica para recorte de dados que não oferecem informações interessante nos documentos; Seção 2.2.2.10 expõe técnicas para se adquirir parâmetros ideias para os modelos de AM (Aprendizado de Máquina); Seção 2.2.2.11.1 exterioriza algumas das métricas populares para avaliarem modelos de regressão; Seção 2.2.2.11.2 apresenta as métricas relacionadas à classificação.

2.2.1 Mineração de Textos

A grande quantidade de dados disponível na internet em tempo atuais é gigantesca. A IBM [39] cita que 90% dessas dados foram gerados desde 2015, sendo que diariamente a sociedade cria 2,5 quintilhões de bytes de dados que são despejados na internet. Com o crescimento populacional, aliado à criação de novos dispositivos, sensores e novas tecnologias, a perspectiva é que essa taxa aumente significativamente com o passar dos anos. Com essa grandiosidade de dados disponíveis na internet, a *Mineração de Dados* vem auxiliando no gerenciamento e organização desses dados de forma simples e automatizada.

Hotho et al. [40] define *Mineração de Textos* como o uso de métodos

para extração de informações úteis de documentos textuais, permitindo a categorização ou estruturação de uma coleção de textos. Já Morais e Ambrósio [41] definem a Mineração de Textos como a técnica que ampara a descoberta da informação inovadora por via de coleções textuais. A Mineração de Texto (ou *Text Mining*) é uma derivação da Mineração de Dados (ou *Data Mining*), com foco em análise e processamento de documentos textuais. A diferença entre estas duas áreas relaciona-se com a organização dos dados a serem analisados, sendo que a primeira utiliza dados não estruturados, enquanto a segunda necessita estruturação dos dados de entrada.

Nogueira [5] salienta que a Mineração de Textos é produzida em etapas e que ao final do processo o desenvolvedor adquire conhecimento em relação aos dados examinados. Este processo pode ser instanciado em determinada aplicação, dependendo de seus requisitos.

Figura 2.3: Etapas do método de Mineração de Texto.

Fonte: Rezende [42]

Rezende [42] categoriza as etapas da Mineração de Textos, como **conhecimento do domínio**, **pré-processamento**, **extração de padrões**, **pós-processamento** e **utilização do conhecimento**. Estas etapas foram ilustradas na Figura 2.3 e serão descritas nas seções seguintes.

2.2.1.1 Conhecimento do Domínio

Esta é a etapa de identificação ou definição do problema que será o foco da aplicação do método de mineração de texto. Fayyad et al. [43] afirmam que o êxito no processo de extração do conhecimento depende, em certa porcentagem, da participação de um especialista do domínio da aplicação e o apoio aos analistas na busca por padrões. Deste modo, antes do inicio do processo deve-se, inicialmente, realizar uma análise criteriosa para se obter experiência sobre o domínio.

Nogueira [5] complementa o assunto, informando que nesta etapa delimita-se quais os documentos que serão utilizados, sendo que o usuário deverá escolher documentos que sejam relevantes para o seu método e sua aplicação. Estes documentos podem ser obtidos de diversas formas, como jornais, livros, artigos eletrônicos e diversos dados disponíveis na internet. Vale ressaltar que muitos desses documentos podem não estar em formato adequado (documentos não digitalizados, por exemplo) ou não possuírem rotulação, o que causaria inconvenientes para os algoritmos de mineração.

2.2.1.2 Pré-Processamento

O pré-processamento é uma das etapas mais importantes e demoradas em Mineração de Textos. Segundo Rezende [42], é comum que os documentos selecionados apresentem formatos inadequados para a extração dos dados. Além disso, há possibilidade de limitações computacionais ao processar grandes volumes de dados por meio de algoritmos de extração. Deste modo, utiliza-se de técnicas de limpeza, tratamento e redução dos dados para auxílio nesta fase. Moura [44] apresenta algumas ações que fazem parte deste processo, sendo elas:

- Verificação da qualidade da coleção de documentos.
- Remoção de termos repetidos.
- Em casos de classificação, realizar o balanceamento da coleção por reamostragem (dividir o conjunto de dados em subconjuntos, observe a Seção2.2.2.6).
- Quando possível, realizar a redução da coleção de documentos.

- Verificar na coleção, se há documentos com repartições (como seções em artigos, revistas e páginas da *web*), de tal forma que no final da extração esses dados sejam utilizados.
- Realizar a análise do tamanho dos documentos no conjunto, verificando assim a necessidade, ou não, de normalização dos pesos atribuídos aos termos em função do tamanho do texto.

Inicialmente, é realizada a padronização dos elementos e, para isso, duas etapas são fundamentais: 1) normalizar o formato dos documentos, tendo a possibilidade de se ter arquivos com diferentes formatos (exemplo: pdf, doc, xls) e alguns inelegíveis para a sua extração; 2) remoção de pontuações, símbolos matemáticos, caracteres especiais, etc [5, 45].

Figura 2.4: Exemplo de matriz atributo-valor.

	t_1	t_2	$\ldots$	t_M	Y
d_1	a_{11}	a_{12}	$\ldots$	a_{1M}	y_1
d_2	a_{21}	a_{22}	$\ldots$	a_{2M}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
d_N	a_{N1}	a_{N2}	$\ldots$	a_{NM}	y_N

Fonte: Moura [44]

Após a *normalização* e *limpeza dos dados*, a etapa seguinte é a estruturação dos dados, para facilitar a extração do conhecimento por algoritmos. A mais popular é chamada de *Espaço Vetorial*, em que cada documento é representado por um vetor e cada palavra possui uma posição dentro do vetor. A junção de todos os vetores, que representam o conjunto de documentos, formam a matriz de *Atributo-Valor*. Na Figura 2.4 é exposto uma representação genérica de N documentos (d_i, $1 \leq i \leq N$) e seus atributos (a_{ij}, $1 \leq i \leq N, 1 \leq j \leq M$). Nos casos de modelos que necessitem de dados rotulados, a última coluna da matriz (y_i, $1 \leq i \leq N$) é considerada a classe a qual o documento pertence [44, 5, 46].

Nesta fase, há um esforço para encontrar palavras que não representam conhecimento útil ou são irrelevantes para a extração de dados, as quais são conhecidas como *Stopwords*. Estes termos geralmente são as preposições, pronomes, artigos, interjeições, entre outros elementos da linguagem. É comum conseguir lista de *stopwords* já prontas para diversos tipos de domínio [44].

Após a remoção das *stopwords*, se dá início a busca por similaridade das palavras. Em casos de variações morfológicas, se reduz uma palavra a seu termo de origem com técnicas como *Stemming and Lemmatization* [45]) e também substantivar as palavras mediante a *Substantivação* [44]. Para os sinônimos, há a possibilidade de utilizar dicionários [44]. Com essas técnicas se permite identificar *uni-gramas*, também conhecidas como *Termos Simples*. Ainda, para melhorar a aplicação, se tem a oportunidade de formar *n*-gramas ou *Termos Compostos*, que são compostos por dois ou mais elementos que aparecem com frequência no conjunto de documentos [5, 44, 45].

Na matriz atributo-valor, cada célula relaciona um documento e um termo. Para essa correlação, é necessário a utilização do *TF-IDF* ou *Termo de Frequência* e *Frequência Inversa de Documentos* [47, 48, 49] para determinar a importância de determinada palavra em um documento. *Termo de Frequência (TF)* corresponde à frequência de determinada palavra no texto e pode ser obtido por meio da Equação 2.1, na qual N_{tt} é o número de vezes que o termo aparece no texto, sendo divido por N_{td} que é o número de termos que há no conjunto de dados.

$$TF = N_{tt}/N_{td} \tag{2.1}$$

Já *Frequência Inversa de Documentos (IDF)* é a métrica da relevância de um termo, definida na Equação 2.2, na qual N_d é a quantidade de documentos e N_{dx} é a quantidade de documentos que possui o termo x.

$$IDF = \log\left(\frac{N_d}{N_{d_x}}\right) \tag{2.2}$$

A partir de TF e IDF, se calcula uma pontuação final chamada aqui de TF_IDF, como descrito na Equação 2.3:

$$TF\_IDF = TF * IDF \tag{2.3}$$

Os termos com TF_IDF alto são os mais relevantes e, consequentemente, as pontuações baixas apresentam menor relevância.

A matriz atributo-valor tende a ser grande, pois há enorme quantidade de termos na coleção dos dados, e esparsa, por haver uma gigantesca quanti-

dade de palavras que ocorrem em poucas partes dos documentos, de maneira que sua medida, em relação a todo o conjunto, seja próxima de 0.

Moura [44] expõe a importância de tornar o conjunto de termos mais compacto, porém sem perder, entretanto, a essência da versão original, em busca de reduzir o custo computacional. Para tal, há duas técnicas comumente utilizadas: *Extração de Atributos* e *Seleção de Atributos*.

A *Extração de atributos* envolve um processo de criar um novo conjunto de termos por via de uma função de mapeamento entre as suas representações. Geralmente o novo conjunto formado apresenta uma dimensão menor do que sua versão original e os novos atributos são gerados mediante a combinação da frequência dos originais. Porém, há uma desvantagem na utilização desta técnica: os novos atributos não mantém uma relação que evidencie com a configuração real do problema, deste modo os novos modelos gerados são difíceis de serem interpretados e alguns casos há a necessidade de uma pré-interpretação dos grupos gerados para o melhor entendimento (Moura [44]). Algumas técnicas se destacam nesse meio, relatadas em Nogueira [5] e Moura [44], como: PCA (*Principal Compenent Analysis*) relatado por Jolliffe [50]; LSA (*Latent Semantic Analysis* - Landauer et al. [51]) ou LSI (*Latent Semantic Indexing* - Deerwester et at. [52]); agrupamento de palavras e busca por grupos similares de palavras (Manning et al. [45]; Slonim e Tishby [53]).

Já a *Seleção de Atributos* está relacionada à separação de um subgrupo de atributos do conjunto original, a partir de critérios definidos. Deste modo, os atributos não sofrem modificações e continuam a ter correlação direta com o domínio do problema [5, 44].

2.2.1.3 Extração de Padrões

Após realizar a identificação do problema e pré-processar os dados com técnicas de limpeza, normalização, entre outras, a próxima etapa é a *Extração de Padrões*. A técnica a ser utilizada nesta etapa está relacionada ao tipo de aplicação que será utilizada e, para tal, há dois grupos de técnicas: as *Preditivas* e *Descritivas* [54].

Em tarefas *preditivas*, o foco é encontrar uma função, que pode ser uma hipótese ou modelo, por meio do conjunto de treinamento para que possa ser realizado a predição de um rótulo ou valor que caracterize um novo exemplo,

com base nos valores de seus atributos de entrada. Métodos de aprendizagem supervisionado representam essa tarefa.

Em tarefas *descritivas*, o objetivo é descrever ou explorar a coleção de documentos; para tal, algoritmos dessa classe não utilizam o atributo de saída, deste modo, pertencem ao ramo de aprendizagem não supervisionada. Uma explicação mais detalhada sobre aprendizado de máquina supervisionado e não supervisionado pode ser observado na Seção 2.2.2 deste capítulo.

2.2.1.4 Pós-Processamento e Uso do Conhecimento

Após a extração de padrões, o passo seguinte é a análise e interpretação dos resultados com auxílio de um especialista, a fim de avaliar as informações obtidas em termos de sua representatividade, novidades encontradas nos resultados e como o conhecimento poderá ser utilizado [44].

A averiguação dos resultados obtidos é necessário para conferir se estão condizentes com as configurações do domínio e são aplicáveis aos objetivos propostos na primeira etapa. Em tarefas preditivas, a avaliação ocorre a partir de métricas que expressem a acurácia das predições sobre novos dados, como por exemplo: *taxa de erro, precisão, cobertura* e *recall*. Essa verificação com métricas e grandezas comparáveis são conhecidas como *Avaliações Objetivas*. Já em tarefas descritivas, a validação é difícil, variando a cada problema e objetivo do projeto; deste modo, a avaliação de um especialista passa a ser necessária. Esse tipo de avaliação é dita *subjetiva* [5, 44].

2.2.2 Aprendizado de Máquina

Figura 2.5: Hierarquia de Aprendizagem de Máquina.

Fonte: Elaborado pelo autor (2022).[2]

[2]Baseado nos seguintes autores: Carvalho et al. [54], Russel e Norving [55] e Luger [56].

O AM é uma sub-área da *Inteligência Artificial* (IA) que visa ao desenvolvimento de algoritmos que permitem ao computador aprender mediante à experiência adquirida em dados passados.

A forma de conduzir um algoritmo de inteligência artificial ao aprendizado é baseado no raciocínio indutivo, em que busca-se extrair padrões a partir de grandes conjuntos de dados. Os algoritmos de AM são baseados no aprendizado indutivo e podem ser subdivididos nos seguintes tipos de aprendizagem, como apresentado na Figura 2.5: *supervisionado*, *não supervisionado*, *semissupervisionado* e *por reforço*.

2.2.2.1 Aprendizado Supervisionado

É realizado com base em um conjunto de dados totalmente rotulados, tornando necessária uma forte intervenção humana na preparação dos dados. Por exemplo, segundo Russel e Norving [55], classificando pessoas como "inadimplentes" e "de boa reputação". Essas duas categorias apresentam aspectos bem característicos: as inadimplentes podem ter contas em atraso, CPF no Serasa, *score* de cartão de crédito baixo e avaliação negativa do banco; as pessoas de boa reputação têm características como pagamentos em dia, *score* de cartão de crédito alto e boa avaliação do banco. Tendo as características definidas, o algoritmo analisará os atributos de cada pessoa e irá predizer qual a probabilidade de ser "de boa reputação" ou "inadimplente". Destacam-se neste aprendizado algoritmos como k-NN (k-*Nearest Neighbors* ou k-Vizinhos mais Próximos) [57], *Máquinas de Vetores de Suporte* (SVM) [55, 58] e *Naïve Bayes* [54, 55].

2.2.2.2 Aprendizado Semissupervisionado

Nos modelos de aprendizado de máquina semissupervisionados, uma parte de dados são rotulados (supervisão humana) e o restante dos dados de treinamento são não-rotulados. Para entender mais sobre ele, pode-se destacar o exemplo apresentado por Russel e Norving [55], no qual foram tiradas várias fotos de pessoas e a cada foto foi associada a idade da pessoa, quando disponível (rotulação característica de aprendizado supervisionado). Porém, algumas pessoas podem não ter informado uma idade ou informado idade falsa, causando ruídos aos dados, caracterizando entradas de um apren-

dizado não supervisionado. Para tal, esse ruído e a falta de rótulos criam um espaço entre esses dois modelos no qual o modelo semissupervisionado se encontra.

Alguns exemplos de algoritmos semissupervisionados podem ser citados, como: *Máquinas de Vetores de Suporte Transdutivas* (TSVM), *Self Training*, *Generative Models*, *Graph-Based Algorithms* e *Multiview Algorithms* [59].

2.2.2.3 Aprendizado Não Supervisionado

Os algoritmos de aprendizado de máquina não supervisionado utilizam um conjunto de dados sem nenhum tipo de rotulação e tem como objetivo descobrir semelhanças entre os objetos analisados, agrupando-os de acordo com suas similaridades. Como no exemplo anterior das pessoas "inadimplentes" e "de boa reputação", o algoritmo irá fazer um agrupamento de pessoas com as características semelhantes e irá classificá-la entre as duas categorias, sem que nenhum dos elementos tenham sido rotulados anteriormente (Luger [56]).

Este modelo de aprendizado, segundo Carvalho et al. [54], pode ser divido em subcategorias, sendo *Agrupamento* (*Clustering*), *Associação* e *Sumarização*. Os algoritmos de agrupamentos têm o foco em agrupar os dados por via de sua similaridade ou dissimilaridade; algoritmos de associação visam encontrar padrões de associações entre os elementos do conjunto de dados; por fim, os algoritmos de sumarização têm a meta de descobrir uma caracterização básica e compacta sobre um conjunto de dados.

Os modelos baseados em agrupamento têm grande utilização dentro do aprendizado não supervisionado, com destaque para os seguintes algoritmos: o modelo *k-Means* [60], *Agrupamento Hierárquicos* [61], *Agrupamento Mean-Shift* [62].

2.2.2.4 Aprendizado por Reforço

O objetivo dos modelos de aprendizado por esforço é aprender a partir de uma série de bonificações ou punições que auxiliem no aprendizado futuro (Luger [56]). Este tipo de aprendizado é muito utilizado em jogos e na área

de Robótica. Um exemplo de aplicação é o jogo AlphaGo [63], o qual tem como objetivo aprender sobre o alvo em questão e consequentemente atingir a meta. A situação enfrentada pelo agente é tentativa e erro para se encontrar uma solução em determinado problema, para tal ele recebe recompensas ou penalidades em suas ações e seu foco é maximizar a recompensa total. Tendo em vista que o programador delimitada as regras do jogo, ele não dá pistas de como o agente deve se comportar durante a execução, para tal, o algoritmo deve descobrir como realizar a tarefa por meio de testes inicias aleatórios, até se obter táticas sofisticadas.

Dayan e Niv [64] destacam a divisão desta área em dois conjuntos, sendo: *Model-free RL* (RL sem modelo), o agente depende da experiencia entre erro e tentativa para definir suas regras; *Model-based RL* (RL baseada em modelo), o agente vislumbra tentar modelar o ambiente e em seguida escolher uma diretriz para se orientar, se baseando no ambiente estudado. Alguns algoritmos são destaque, como: *A2C/A3C* [65], *DQN* [66], *DDPG* [67], *MBMF* [68], *AlphaZero* [69].

2.2.2.5 Algoritmos de Aprendizado de Máquina para Análise de Documentos

Para o presente projeto, serão aplicadas técnicas de mineração de texto juntamente com aprendizado de máquina, a fim de se alcançar os objetivos delimitados anteriormente. Em relação ao aprendizado de máquina, serão realizadas tanto tarefas preditivas quanto descritivas, pois utilizará algoritmos supervisionados para predição de notas do Enade e não supervisionados para agrupamento de termos com similaridades presentes nos documentos. A seguir serão apresentados algoritmos utilizados neste projeto, juntamente com as métricas de avaliação utilizadas para validação destes modelos.

2.2.2.5.1 k-NN ou k-Vizinhos mais Próximos

Popularmente conhecido como *k-NN* (do inglês *k-Nearest Neighbors*) [57], este é um algoritmo de aprendizado supervisionado, que realiza classificação ou regressão de um determinado exemplo do conjunto de dados com base nas características dos seus k vizinhos mais próximos. Em modelos de classificação o intuito é analisar os k vizinhos mais próximos da entrada sobre a qual você está tentando fazer uma predição. Em seguida, ele pro-

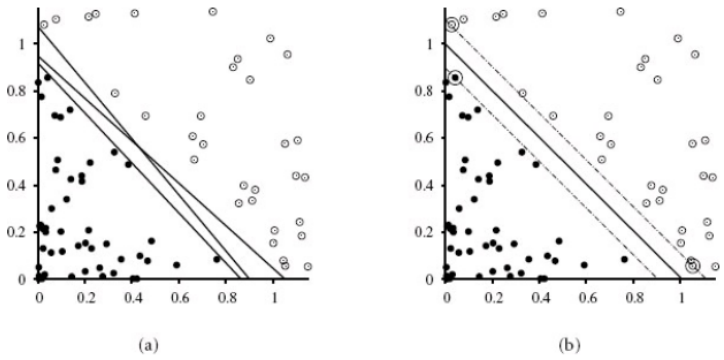
duzirá o rótulo mais frequente entre esses k exemplos (Cover e Hart [57]). Já no modelo de regressão, o algoritmo coletaria os valores associados aos k exemplos mais próximos da amostra para o qual se deseja realizar alguma predição e os combinaria adequadamente para gerar o valor associado a este exemplo [70].

A escolha do melhor k pode ser realizada pelo usuário empiricamente após sucessivos testes iniciais. Carvalho et al. [54] apresenta duas estratégias, baseado na literatura, para obter o melhor k: a primeira é utilizando *Validação Cruzada* e outra é atribuir pesos aos vizinhos por meio de sua contribuição.

2.2.2.5.2 Máquinas de Vetores de Suporte

As *Máquinas de Vetores de Suporte* (do inglês, Support Vector Machines - SVMs) são sistemas de aprendizado supervisionado, geralmente aplicados a problemas de classificação, que analisam os dados e reconhecem padrões. Esta estratégia de aprendizado obteve, em poucos anos de sua introdução, resultados mais satisfatórios do que a maioria dos sistemas em uma ampla variedade de aplicações. Seu objetivo é encontrar em um hiperplano, pontos de dados distintos e classificá-los [58].

Figura 2.6: Exemplo SVM com duas classes.



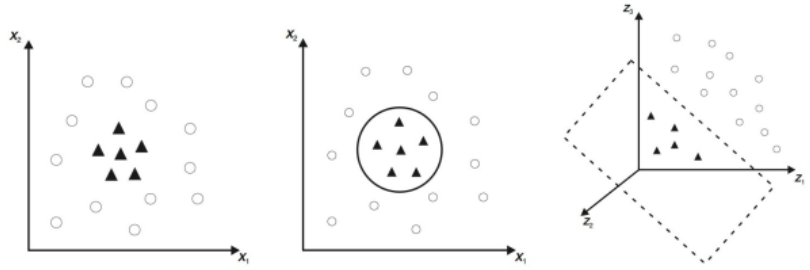
Fonte: Russel e Norving [55].

Na Figura 2.6 (a) é exposto a forma simplificada de um conjunto de dados com duas classes e três candidatos a limiares de decisão que separam linearmente as duas classes. Na Figura 2.6 (b) contêm uma solução obtida pela SVM, na qual houve a procura pelo ponto médio entre as duas classes. A linha preta no centro é conhecida como *Hiperplano* ou *separador de margem*

máxima, a distância entre as duas linhas tracejadas é a *Margem* e o pontos próximo a cada linha tracejada se chamam *Vetor de Suporte*.

Segundo Cristianini e Shawe-Taylor [58], nem todo conjunto de dados é linearmente separável. Logo, as SVMs são extensíveis a modelos não-lineares (duas ou mais dimensões, Figura 2.7). Para isso, são utilizados distintos *Kernels*, os quais são funções para mapear dados de um espaço multidimensional para outro. De maneira mais intuitiva, é uma função que transforma os dados para que se sejam mais facilmente entendidos por um regressor ou classificador. Podendo se utilizar SVMs com diversos *kernels*, dos quais destacam-se os seguintes: *Linear, Polinomial, Sigmoid, RBF* (do inglês, *Radial Basis Function*).

Figura 2.7: Exemplo de SVM em duas ou mais dimensões.



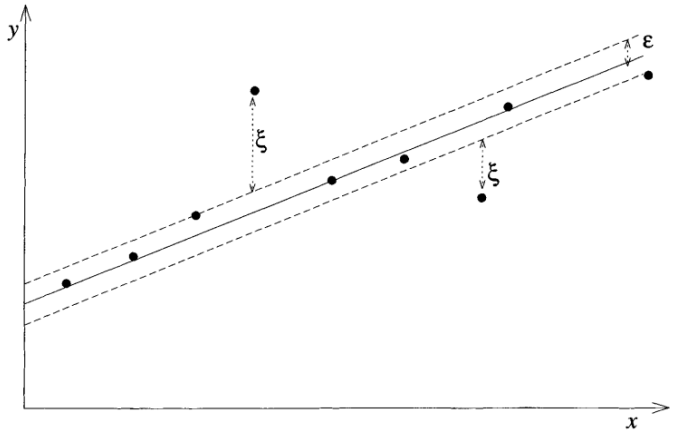
Fonte: Carvalho et al. [54].

2.2.2.5.3 Máquinas de Vetores de Suporte para Regressão

As *Máquinas de Vetores de Suporte Aplicadas à Regressão* (do inglês, Support Vector Regression - SVRs), são modelos de regressão baseadas em SVMs, que apresentam características bastante similares a sua precursora [58]). A diferença está na utilização do parâmetro ajustável, ϵ (*Epsilon*), que determina a largura do tubo em torno do hiperplano. Os atributos que se localizam dentro do tubo são considerados previsões corretas e não serão penalizações. Já aqueles pontos fora do plano são considerados vetores de suporte, diferentemente do modelo SVM no qual os pontos se encontram na margem. Então, como ilustrado na Figura 2.8, o ξ (*slack - folga*) é a medida para a se encontrar distância entre os pontos fora do tubo, podendo ser ajustado pelo parâmetro C do regressor (Cristianini e Shawe-Taylor [58]).

Na Figura 2.8 é detalhado um hiperplano representado pela linha cheia, e os limites do tubo ilustrados por linhas tracejadas (delimitadas por ϵ). O objetivo do modelo é minimizar o erro, identificando uma função que aumente o número de pontos dentro do tubo, consequentemente, reduzindo a ξ.

Figura 2.8: Exemplo de SVR.



Fonte: Cristianini e Shawe-Taylor [58].

2.2.2.5.4 Perceptron Multicamadas

Popularmente conhecida como *Multilayer Perceptron* (MLP), é uma rede neural *Feedforward* em camada, o que significa que o processo flui unidirecionalmente da camada de entrada até a saída. Vale ressaltar que entre essas duas etapas há uma ou mais camadas ocultas e que para se ter uma MLP é necessário ter pelo menos três camadas: camada de entrada, camada oculta e camada de saída (Bishop [71]). Cada neurônio (módulo ou nó, responsável por processar a informação e repassar sequencialmente) presente na rede possui seu próprio peso (ou *Bias*) e *perceptrons* da mesma camada apresentam a mesma função de ativação. A função de ativação pode variar dependendo do objetivo da utilização da MLP juntamente com os dados que serão processados. Uma das funções de ativação mais popular é a *Sigmoid* [54, 71, 72].

Um algoritmo presente na MLP é a *retropropagação* (*backpropagation*), que visa corrigir os pesos das camadas de forma que este processo se inicie na camada de saída e venha alterando os valores até as camadas iniciais. Isto é, o algoritmo realiza a implementação de um gradiente de descida no espaço de parâmetros para diminuir o erro de saída [55, 72].

Taud e Mas [72] explicam que o desempenho desta rede neural não depende somente da escolha correta do número de camadas ocultas, de nós e dos dados de treinamento, mas também da calibragem de parâmetros como, por exemplo, a taxa de aprendizagem, o controle adequado da mudança de peso e um número ideal de iterações.

O *Regressor Perceptron Multicamadas* (simplificado neste trabalho pela sigla MLPR) se refere à variante da MLP para problemas de regressão, utilizando o erro quadrado como a função de perda e tendo como saída um conjunto de valores contínuos.

2.2.2.5.5 k-Means

O algoritmo de aprendizado não supervisionado denominado *k-Means* foi proposto em 1967 por MacQueen [60] e visa particionar um conjunto de dados em k *clusters*, no qual k é determinado pelo usuário, de acordo com a similaridade entre seus atributos e dissimilaridade entres os grupos. Para tal, o *k-Means* utiliza uma técnica interativa para realizar a alocação correta dos centroides (ponto cujas coordenadas são as médias das coordenadas dos pontos que formam o *cluster* que, nesse modelo, terá formato esférico).

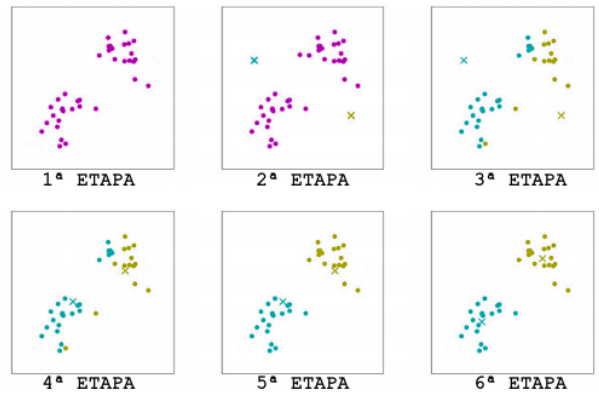
A essência do *k-Means* é minimizar o *Erro Quadrático* (E), descrito na Equação 2.4 [54], e que corresponde ao somatório de variação dentro de um *cluster*, no qual $\bar{\mathbf{x}}^{(j)}$ representa o centroide do *cluster* $\mathbf{C}_j$ e $d(\mathbf{x}_i,\bar{\mathbf{x}}^{(j)})$ é a distância euclidiana entre o ponto $\mathbf{x}_i$ e o centroide $\bar{\mathbf{x}}^{(j)}$.

$$E = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathbf{C}_j} d(\mathbf{x}_i, \bar{\mathbf{x}}^{(j)})^2 \tag{2.4}$$

O objetivo é encontrar uma divisão que contenha k *clusters*, a fim de minimizar o erro quadrático E. A inicialização de k pode ser feita de várias formas, inclusive aleatoriamente. Logo após, cada alvo do conjunto é agregado ao *cluster* no qual há um centroide mais próximo. Em seguida, os centroides são calculados novamente e este processo sofre iterações até que não haja mais alterações na associação dos objetos aos *clusters* [54, 73], como exibido na Figura 2.9. O algoritmo é sensível à primeira escolha de *cluster*, podendo reagir adequadamente e com ótimos resultados ou de forma decepcionante, causando *clusters* desbalanceados.

O método mais comumente utilizado para validar os resultados obtidos pelo *k-Means* é chamado de *Silhueta*, que foi proposto por Kaufman e Rousseeuw [74] em 1990.

Figura 2.9: Exemplo do *k-Means*.



Fonte: Elaborado pelo autor.[3]

A silhueta é definida na Equação 2.5 [61], e terá um valor no intervalo [-1,1], sendo quanto mais próximo de 1, melhor o agrupamento. Na Equação 2.5, $b(\mathbf{x}_i)$ é a distância média entre o elemento i e todos o elementos do *cluster* mais próximo, $a(x_i)$ é a distância média do ponto i aos demais pontos do *cluster* e *max* é o valor máximo mediano dos elementos $a(x_i)$ e $b(x_i)$.

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max(a(\mathbf{x}_i), b(\mathbf{x}_i))} \tag{2.5}$$

2.2.2.6 Amostragem dos Dados para Estimativas de Desempenho dos Algoritmos de Aprendizado de Máquina

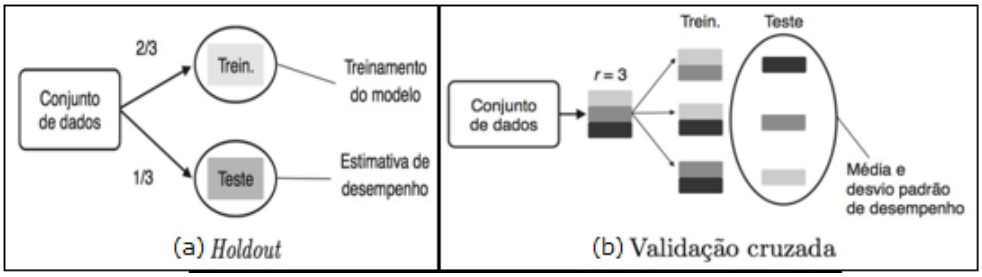
Para Carvalho et al. [54], calcular o desempenho preditivo, como taxa de erro ou acerto, do algoritmo de AM utilizando o mesmo conjunto de dados empregados no treino do modelo, produz estimativas não reais e pouco confiáveis. A capacidade de generalização de um modelo de AM está correlacionado a dados ainda não vistos. Esta avaliação, da capacidade de generalização, na prática é de suma importância, pois auxilia na escolha do modelo de AM e retorna uma medida de qualidade em relação ao modelo (Hastie et al. [75]; Tan et al. [76]).

[3]Baseado no seguinte autor: Flach [61].

Um fato corriqueiro, e correto, para a avaliação do modelo é analisar o desempenho dele a partir do subconjunto de teste, que ainda não foi utilizado em nenhuma etapa anterior. Neste caso, utiliza-se métodos de *amostragem* alternativos para obter resultados confiáveis, no qual se tem o conjunto completo D e reparticionando ele em subgrupos de treino ($D.treino$), que é utilizado para a seleção de modelo, e teste ($D.teste$), usado para calcular alguma medida de desempenho (Carvalho et al. [54]; Tan et al. [76]).

A seguir serão apresentado alguns métodos popularmente utilizados para particionarem o conjunto de dados e avaliarem o modelo, como ilustrado na Figura 2.10.

Figura 2.10: Exemplos de técnicas de amostragens.



Fonte: Carvalho et al. [54].

2.2.2.6.1 Holdout

É o método mais simples de separação do conjunto principal D, Figura 2.10 (a), em um subconjunto $D.treino$, voltado para treinar o algoritmo de AM, e outro subgrupo denominado $D.teste$, para validar o modelo com dados não visto antes, realizando assim, uma avaliação de generalização. Essa repartição é realizada de forma aleatória sendo, em muitos casos, 80% para treino e 20% para teste (Tan et al. [76]).

Contudo, Tan et al. [76] enfatizam que determinar a porcentagem de separação para cada subconjunto não é trivial. No caso de se ter um $D.treino$ pequeno, o modelo pode não aprender corretamente, pois utilizou um número insuficiente de amostras em seu treinamento. Já em relação ao $D.teste$, se for pequeno ele pode não ser muito confiável, pois ele estaria utilizando poucas amostras para o teste e sua generalização. Ao alterar a porcentagem para cada subconjunto, pode se obter uma alta variação dos resultados gerados pelo modelo.

2.2.2.6.2 Cross-Validation

A Validação Cruzada (do inglês, *Cross-Validation*) é o método mais utilizado para calcular o erro de predição. Este modelo reparticiona em subconjuntos de *D.treino* e *D.teste*, realizando várias iterações, na qual a cada repetição a posição (ou a repartição) dos subconjuntos se alteram, garantido diversas combinações e aumentando a confiabilidades dos dados. Ao final de cada iteração é calculada a medida de avaliação, deste modo, ao final do ciclo calcula-se a média da medida ao longo das iterações (Hastie et al. [75]; Tan et al. [76]).

Há diversas variações de *Cross-Validation*, sendo as mais populares: *k-Fold Cross-Validation* e *Stratified k-Fold Cross-Validation*.

2.2.2.6.2.1 k-Fold Cross-Validation

Neste modelo de validação cruzada, segundo Carvalho et al. [54], reparticiona o conjunto de dados em *k-Folds* (pasta, amostras), no qual se tem *D.treino* e *D.teste*. A cada iteração do processo, é utilizado uma partição diferente para *D.teste*. Na Figura 2.10 (b), se tem 3-*Folds* e a cada ciclo a amostra de teste é alternada, até que ela tenha utilizada cada posição disponível. A cada ciclo o conjunto de teste gera uma taxa de desempenho, que ao final do método calcula-se a média e desvio padrão dessas taxas.

Contudo, delimitar a quantidade de *Folds* adequado para o seu uso, é algo arbitrário. Kohavi [77] apresenta diversos testes com valores de k e diálogo sobre a importância do mesmo. Por exemplo, a utilização de valores como 10 ou 20 faz com que a variação de dados seja reduzida e o viés seja aumentado. Contudo, ao utilizar valores pequenos, como 2 ou 5, a variação tende a aumentar e o viés a diminuir. Para encontrar um valor correto de k, entretanto, é necessário realizar diversos testes. Pode se afirmar, portanto, que a escolha do parâmetro k é feita por meio da prática e uma técnica empírica, considerando que cada fold tenha entre 20% a 30% do total de exemplos. Com essa porcentagem, os *Folds* terão uma quantidade de amostras suficientes e variadas do conjunto de dados.

2.2.2.6.2.2 Stratified k-Fold Cross-Validation

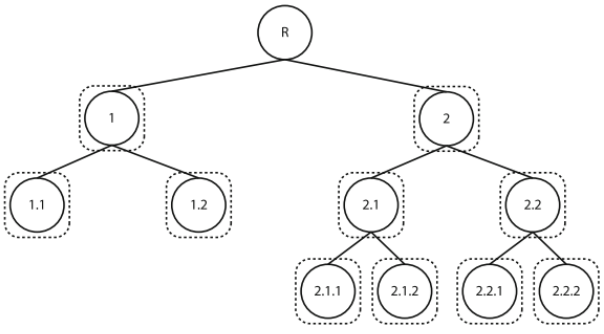
Carvalho et al. [54], salienta que é uma variação do *k-Fold Cross-Validation*, que visa manter uma proporção semelhante de cada classes, do conjunto original, em todas as partições (*Folds*) do processo. Como exemplo, um conjunto de dados no qual há 20% da classe c_1 e 80% da classe c_2, deste modo, ao aplicar o método estratificado ele buscará manter esta proporção em cada *Fold*.

2.2.2.7 Classificação Hierárquica

Métodos de classificação são popularmente utilizados em problemas nos quais os dados estão relacionados a poucas classes, geralmente 2, sendo esta metodologia conhecida como Classificação Simples ou Plana [78, 79], sendo um classificador para predizer todas as classes do conjunto de dados. Porém, há projeto no qual os dados estão atribuídos a multiclasses, tais métodos podem sofrer diversos problemas e, consequentemente, não classificarem corretamente os dados. Para tal, denomina-se Classificação Hierárquica (CH, ou *Hierarchical Classification* - HC) [79, 80, 81], a heurística de problemas multiclasses.

Um dos métodos mais popular de CH é a abordagem *Top-Down* (de Cima para Baixo, ou Classificadores Locais) [81], no qual cada nível, ou nó, da árvore (ou estrutura) de classificação possui seu classificador responsável por aquela seção.

Figura 2.11: Exemplo de árvore do método *Top-Down* em CH.



Fonte: Silla e Freitas [81]

Na Figura 2.11, mostra-se um exemplo da estrutura na heurística do *Top-Down*, sendo que os círculos representam as classes e os quadros tracejados são os classificadores locais para cada nó. R indica a raiz da árvore, no qual contém todas as classes, antes bifurcação da hierárquia. Nesta metodologia, cada nó filho é dependente de seu nó pai, assim, se a informação não vir dos níveis superiores os demais níveis subsequentes não poderão trabalhar.

Esta técnica tem como principal desvantagem a propagação de erros, ou seja, o encadeamento de decisões, que, consequentemente, irão propagar o erro para cada nível subsequente. Este fato ocorre pelo dependência que cada nó tem com seu antecessor. Como principal vantagem, tem-se a correção de predições inconsistentes das classes em diferentes níveis, durante a etapa de teste e não durante o treinamento.

Há outras variações de CH, como explanado em [79, 80, 81], sendo algumas delas: *One-Against-One and the One-Against-All, Binary Hierarchical Classifier (BHC), H-SVM (Hierarchical SVM)* e *The Big-Bang Approach*.

2.2.2.8 *Data Augmentation*

Ao se utilizar técnicas de IA com poucos dados, é costumeiro que os modelos sofram *Overfitting* e, consequentemente, não consigam generalizar corretamente o problema. Uma solução para este problema é aplicar *Data Augmentation* (DA, ou *Aumento de Dados*, em tradução) [82, 83] sobre os dados, a fim de aumentar a quantidade e a qualidade das amostras de dados. Esta técnica é popularmente utilizada em conjunto de dados formados por imagens, de modo que, ao empregar DA, uma imagem poderá ser replicada de ângulos diferentes ou de outras formas, posições, filtros, etc.

Contudo, quando o conjunto de dados é composto por documentos textuais, há a necessidade de aplicar mecanismos diferentes como: Substituição por Sinônimos [84], Tradução Reversa [85, 86], *Word Embeddings* [87], *Contextualized Word Embeddings* [88, 89], *Text Generation* [90], entre outras.

Neste projeto, utiliza-se a substituição por Sinônimos. Suponha que haja um conjunto pequeno de documentos textuais, causando a necessidade de mais amostras para melhorar o desempenho de um determinado modelo, com baixo custo computacional. Para solucionar essa questão, pode-se criar novas amostras de textos, substituindo-se termos originais por seus sinôni-

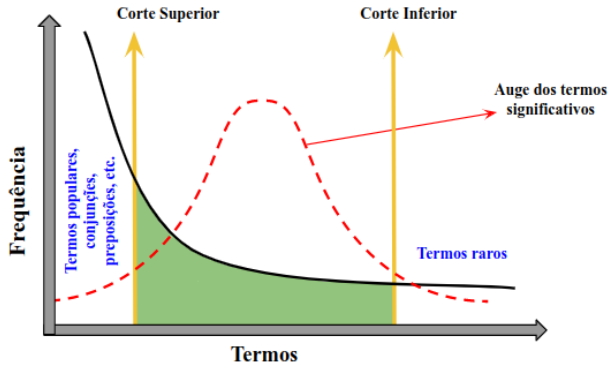
mos [84]. Isto irá impactar no tamanho do conjunto de dados, mantendo o texto com o mesmo sentido. É simples de ser usada: primeiro seleciona-se uma palavra do documento, em seguida se busca e troca este termo por um dos seus sinônimos disponíveis. Um exemplo seria a substituição da palavra "quarto" por sinônimos, como sinônimos "aposento" ou "sala".

2.2.2.9 Corte de Luhn

Ao trabalhar com dados textuais, é habitual que haja termos com altas pontuações TF-IDF e altas Frequência dos Termos (TF), ou termos com poucas ocorrências. Frente a isso, Luhn [91] propõe retirar termos com altas frequências, pois considera que sejam palavras comuns e pouco significativas; além disso, também sugere a retirada de termos com baixa frequência, que são consideradas palavras raras, que não trazem contribuições significativas ao conteúdo do texto. Este processo é chamado de *Corte de Luhn (CL)*.

Luhn se baseava na Lei de Zipf [92], cuja a frequência de um elemento em certo ambiente é inversamente proporcional a sua posição, ou seja, o elemento com mais ocorrências irá aparecer com frequência duas vezes maior do que o elemento com a segunda maior frequência, e assim sucessivamente.

Figura 2.12: Exemplo do Corte de Luhn.



Fonte: Elaborado pelo autor (2022). [4]

Na Figura 2.12, é exibido um exemplo do Corte de Luhn, no qual na parte anterior ao corte superior se tem as preposições, conjunções, entre outros elementos da língua utilizada, e após o corte inferior há os termos raros. A região entre os dois cortes é a que contém os termos relevantes,

[4]Baseado no seguinte autor: Luhn [91].

os quais trarão informações significativas no processamento do texto. Luhn ainda salienta que não há uma posição ideal para o corte, deve-se analisar os dados e fazer vários cortes, a fim de se encontrar a melhor opção.

2.2.2.10 *Grid Search* para Otimização de Hiperparâmetros

Algoritmos de aprendizado de máquinas possuem vários hiperparâmetros. Há diversas técnicas para automatizar a escolha dos melhores valores para esses parâmetros, em busca de melhorar o desempenho dos modelos de aprendizado de máquina, diminuindo o esforço manual do desenvolvedor (Hutter et al. [93]). Algumas destas técnicas são: *Grid Search*, *Random Search* (*Pesquisa Aleatória*, em tradução) e *Genetic Algorithm* (*Algoritmo Genético*, em tradução) (Hutter et al. [93]; Liashchynskyi e Liashchynskyi [94]).

O *Grid Search* (*Pesquisa em Grade*, em tradução) é o modelo de otimização de hiperparâmetros mais popular. O usuário deve fornecer ao modelo, um conjunto de valores finito para o(s) hiperparâmetro(s) para o(s) qual(is) deseja-se otimizar a escolha. Deste modo, a *pesquisa em grade* irá averiguar o produto cartesiano dos conjuntos e avaliar os resultados obtidos a partir destas combinações. Embora seja simples de aplicar, este modelo sofre com a dimensionalidade, devido ao número de avaliações/combinações que cresce exponencialmente, em relação ao conjunto de valores especificados (Hutter et al. [93]; Liashchynskyi e Liashchynskyi [94]).

2.2.2.11 Métricas para avaliação dos Modelos

A presente seção apresenta algumas métricas para avaliação de modelos de aprendizado de máquina, as quais serão utilizadas neste projeto. Para avaliar os modelos de regressão, é necessário utilizar métricas para prever valores escalares, como *Erro Médio Quadrático, Erro Quadrático Média da Raiz, Erro Médio Absoluto*, entre outras 2.2.2.11.1. Para avaliação dos modelos de classificação, utilizam-se métricas como a *accuracy*, *F1*, *precision*, entre outras (Seção 2.2.2.11.2). A métrica adotada para os modelos de agrupamento já foi apresentada na Seção 2.2.2.5.5.

2.2.2.11.1 Métricas para avaliação de modelos de regressão

2.2.2.11.1.1 Erro Médio Quadrático

Ou *Mean Squared Error* (MSE), tem como objetivo informa a proximidade da linha de regressão em relação ao conjunto de dados. Para tal, é capturado a distância (conhecida como erros) entre os pontos até a linha que regressão e as elevando ao quadrado, para evitar erros negativos e dando mais peso a diferenças maiores. Essa métrica encontra a média dos erros dos dados, deste modo vem o seu nome (Wang e Bovik [95]). Na Equação 2.6 é apresentado a MSE, por Wang e Bovik [95]:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 \tag{2.6}$$

2.2.2.11.1.2 Erro Médio Absoluto

Do inglês *Mean Absolute Error* (MAE), oferece a média da diferença absoluta entre a previsão do modelo e o atributo alvo, ela mede a magnitude média dos erros em um conjunto de previsões, sem considerar sua direção. Em outras palavras, MAE mede a distância média absoluta entre os dados reais e os dados previstos (Willmott e Matsuura [96]; Chai e Draxler [97]). Nas Equações 2.7 e 2.8, são representações da MAE, segundo Willmott e Matsuura [96] ou Chai e Draxler [97]:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i| \tag{2.7}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |e_t| \tag{2.8}$$

2.2.2.11.1.3 Erro Percentual Absoluto Médio

Se utiliza métricas de erro curtas e úteis para avaliar a qualidade dos regressores ou, em outras palavras, para determinar o quão bem suas previsões correspondem aos valores reais. O *Erro Médio Absoluto (MAE)*, descrito na Equação 2.7, é a mais intuitiva das métricas, pois expõe a diferença absoluta entre os dados e as previsões do modelo. No entanto, o MAE não indica desempenho inferior ou superior do modelo (se o modelo está abaixo ou acima dos dados reais). Então, se utiliza o MAPE (*Erro Percentual Absoluto Médio ou Mean Absolute Percentage Error*), expresso na Equação 2.9, que é igual ao MAE, mas com ajustes para converter tudo em porcentagens [96, 97, 98].

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} |\frac{y_i - x_i}{x_i}| \tag{2.9}$$

2.2.2.11.2 Métricas para Modelos de Classificação

Para avaliar modelos de classificação, é realizada a comparação entre o alvo original e o alvo predito por ele, para tal, as métricas dessa seção visam medir o quão distante o classificador está em relação ao alvo original ou a classificação perfeita.

As métricas utilizadas neste projeto foram a *Accuracy*, *Microaveraging Precision* e *Microaveraging Recall*. Para utilizá-las, deve-se antes elaborar a *Matriz de Confusão*, que auxilia na avaliação do desempenho do classificador em relação ao dados de teste. A Matriz de Confusão [99, 100, 101] é uma matriz bidimensional, na qual em uma dimensão encontram-se as classes verdadeiras e na outra as classes preditas pelo modelo. Também pode haver matrizes binárias, de duas classes ou multiclasses.

Um exemplo de matriz de confusão é ilustrado na Tabela 2.2. Na diagonal da esquerda para direita decrescente encontra-se "Verdadeiro Positivo (VP ou *True Positive - TP*)" e o "Verdadeiro Negativo (VN ou *True Negative - TN*)", TP significa que o atributo alvo que estava sendo procurado foi previsto corretamente, já TN decorre quando o atributo alvo que não estava sendo procurado foi previsto corretamente. Para "Falso Negativo (FN ou *False Negative - FN*)" significa que o atributo alvo que estava sendo pro-

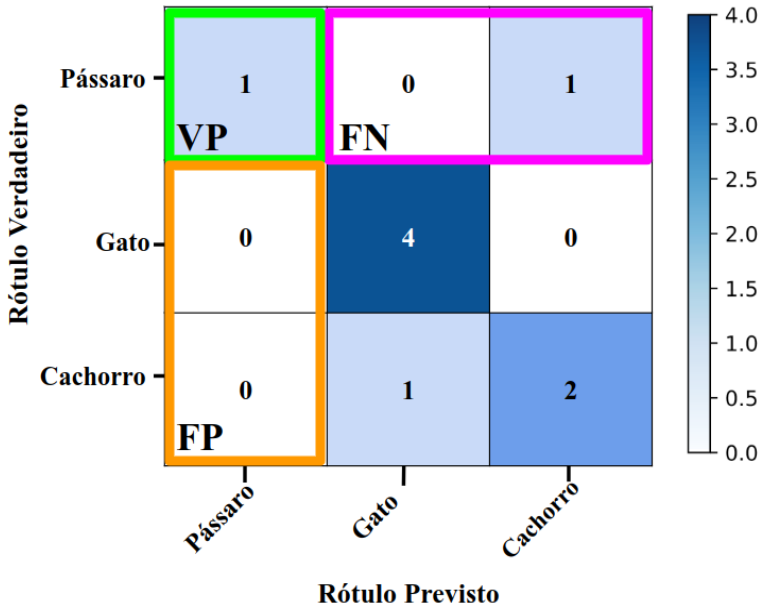
curado foi previsto incorretamente e o "Falso Positivo (FP ou *False Positive - FP*)" decorre de quando o atributo alvo que não estava sendo procurado foi previsto incorretamente.

Tabela 2.2: Exemplo de Matriz de Confusão.

		Valor Predito	
		Positivo	**Negativo**
Valor Real	**Positivo**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Elaborado pelo autor (2022)[5].

Figura 2.13: Exemplo Matriz de Confusão Multiclasses.



Fonte: Elaborado pelo autor (2022)[6].

Na Figura 2.13 é apresentando um exemplo de uma matriz de confusão multiclasses. A análise sobre ela é um pouco diferente, pois os acertos diagonal (na cor verde) são considerados como VP. Para obter os FN, deve-se olhar a linha (ou a horizontal, representado pela cor rosa) de cada VP e para os FP atentar para a coluna no qual se encontra o VP, delimitado pela cor laranja.

[5]Baseado nos seguintes autores: Sammut e Webb [99]
[6]Baseado nos seguintes autores: Sammut e Webb [99]

2.2.2.11.2.1 Acurácia

A *Acurácia*, ou *Accuracy* [102, 103], é uma das métricas mais populares e utilizada no meio da classificação, por ser simples e direta em relação aos resultados finais. Sua heurística é bastante simples: número total de predições correta, divido pelo número total de amostras utilizadas. Ou se baseando na Matriz de Confusão 2.2, Número de Predições Correta seria VP e VN, os acertos, divido pela soma total de amostras utilizada, ou seja, $VP+VN+FP+FN$. Ambas versões podem ser analisadas nas Equação 2.10 e Equação 2.11.

$$Acuracia = \frac{NumeroPredicoesCorreta}{TotalAmostras} \tag{2.10}$$

$$Acuracia = \frac{VP + VN}{VP + VN + FP + TN} \tag{2.11}$$

2.2.2.11.2.2 Micro-Avereged

Ao trabalhar com classificação multiclasses, há a necessidade de ponderar igualmente cada predição gerada pelo modelo, deste modo, a *Micro-Avereged* [76] possui duas variações, sendo a *Precision* e a *Recall*, indicadas para quando houver um certo desbalanceamentos das classes. Ambas podem ser vislumbradas nas Equações 2.12 e 2.13.

A *Precision* (ou *Precisão*, em tradução) pode ser definida como o somatório do conjunto de todas as VP dividas pelo somatório do conjunto de todas as VP mais o somatório das FP. Já a *Recall* (ou *Recuperar*, em tradução) é bem similar, mudando a penas de FP para FN na soma final.

$$Precision = \frac{\sum_{i=1}^{|c|} VP_i}{\sum_{i=1}^{|c|} (VP_i + FP_i)} \tag{2.12}$$

$$Recall = \frac{\sum_{i=1}^{|c|} VP_i}{\sum_{i=1}^{|c|} (VP_i + FN_i)} \tag{2.13}$$

2.2.3 Considerações Finais

A presente seção apresentou as subáreas de IA (Inteligencia Artificial), Mineração de Textos e Aprendizado de Máquina, apresentando algoritmos para classificação, regressão e clusterização, técnicas para amostragem, otimização da escolha de parâmetros e métricas para avaliação dos modelos.

3 TRABALHOS RELACIONADOS

Os trabalhos anteriormente publicados, relacionados a este projeto, foram divididos em dois tópicos: projetos que contêm análises críticas de documentos de gestão educacional, de forma não automatizada, são abordados na Seção 3.1; e trabalhos com aplicação de algoritmos para automatizar o processo de análise de documentos textuais no mesmo domínio deste projeto, são descritos na Seção 3.2.

3.1 ANÁLISES DE DOCUMENTOS DE GESTÃO E AVALIAÇÃO EDUCACIONAL

Esta seção aborda algumas análises críticas, não automatizadas, sobre documentos de gestão educacional e instrumentos de avaliação.

Brito [104] realiza uma análise crítica sobre o Enade e os Projetos Pedagógicos de Cursos de graduação. A autora se baseia na bibliografia original do Enade para elaborar uma análise sobre os objetivos do SINAES que, por meio dos resultados do Enade, passou a avaliar cada IES. O Enade avalia a habilidade acadêmica, em que o aluno deve demonstrar sua capacidade de realizar determinado trabalho, solucionar problemas e demonstrar a conhecimentos de sua área; e a competência profissional, que é a capacidade de mobilizar, articular e aplicar na prática os conhecimentos, habilidades, atitudes e valores necessários para o desempenho eficiente e eficaz em atividades solicitas por determinado trabalho e do desenvolvimento tecnológico. A ação de comparação de resultados entre cursos é muito popular, porém é uma ação incorreta, pois a única comparação na prova seria em torno das questões do componente de formação geral, no qual graduandos recebem uma qualificação padrão a todos os cursos. Outro ponto critico é em relação aos rankings gerados a partir dos relatórios finais do Enade, sendo utilizados de forma propagandísticas e manipuladora. Os resultados do Enade não são elaborados com intuito de comparação, elaboração de rankings e decidir qual a melhor universidade, mas sim auxiliarem as instituições a avaliarem o desempenho de seus alunos e verificarem se o ensino está sendo bem repassado, se o aluno desenvolveu um carácter humano e profissional adequado e muitos outros aspectos. A sociedade necessita compreender que o Enade é voltado para a

melhora das instituições e não averiguar qual a melhor entre elas, esse fato pode ocasionar aos alunos escolherem por universidades com conceito alto e prejudicando as universidades com conceitos baixos e, consequentemente, a dificuldades dessas instituições.

Para a autora, um PPC tem como principal objetivo a adequação coletiva que envolva os conceitos presentes nas DCNs e o cenário ao qual está presente; deste modo, a construção de um projeto pedagógico do curso deve contemplar algumas etapas, como: grupo de diretrizes operacionais e organizacionais, que manifestem e auxiliem a prática pedagógica do curso; estrutura curricular, ementas e bibliografia; o perfil do graduando concluinte; e tudo que envolva ou se refere ao curso estabelecidas pelo MEC. Brito [104] destaca também que muitos PPCs apresentam uma descrição cansativa sobre as disciplinas e o conteúdo presente nelas, deixando em segundo plano o perfil do formando, as habilidades que ele poderá adquirir durante o curso e, por fim, as características profissionais do acadêmico ao final de sua formação. A autora também ressalta que as IESs necessitam ter clareza no tipo de formação que desejam, implementando PPCs que formem profissionais úteis à sociedade, alguns com perfis mais práticos e aplicados, já outros com perfis acadêmico e teórico. Toda essa caracterização do profissional e o âmbito da sociedade em torno da instituição devem estar presente no PPC, auxiliando os docentes a um aprendizado com o foco pré-estabelecido.

O estudo de Canan e Eloy [105], analisa diversas questões relativas ao Enade e suas contribuições para gestores de universidades. O Enade além de ser um dos indicadores de qualidade, no qual o governo consegue avaliar o ensino superior no país, ele também analisa o grau de conhecimento dos alunos coletivamente e verifica como as práticas de ensino afetam positivamente ou negativamente os acadêmicos. Os autores também argumentam que com o passar dos anos, muitas universidades passaram a preparar seus alunos para o exame nos meses que o antecedem, o que pode comprometer as conclusões sobre as instituições e cursos avaliados. Além disso, outra argumentação é que muitas universidades não sabem como processar os resultados do Enade, para promover evoluções que possibilitem melhores resultados futuramente. Neste contexto, um trabalho, como esta presente dissertação, poderia favorecer a análise dos gestores na concepção de mudanças nos projetos pedagógicos visando a busca de melhores avaliações futuras.

Canan e Eloy [105] apresentaram dois aspectos negativos a respeito do Enade, primeiro, que o exame perdeu sua real função de avaliação e tem se assemelhado com o antigo "Provão", com o objetivo de ranquear as institui-

ções por meios de notas e as penalidades sofridas por elas em caso de baixa pontuação. De forma geral, o Enade é aplicado em nível nacional, no qual os autores consideram como um erro, pois cada estado apresenta sua diversidade cultural e econômica, onde cada instituição se encontra, de tal forma a universidade tem de adequar a sua própria realidade, algo que o Enade não considera na hora de realizar as pontuações. Segundo, o Enade deveria auxiliar as universidades a encontrarem soluções para melhorarem seus níveis de aprendizagem e não em criar um novo problema decorrente das notas baixas, as quais lhes causam penalidades.

Quais fatores podem influenciar no resultado, em relação a notas altas e baixas, no Enade? Está pergunta norteou os autores Lemos e Miranda [106] em uma pesquisa exploratória com foco em entender o que se passa nas instituições para gerarem bons ou más resultados. Nos anos de 2009 e 2012, foram analisados 383 e 464 cursos, respectivamente, atribuindo-se notas para infraestrutura, escolaridade da família, nota para organização didático-pedagógica, desempenho dos acadêmicos na prova de entrada na universidade (Enem), regime de trabalho dos profissionais, quantidades de mestres ou doutores na instituição e dentre todas essas variáveis, a única a não influenciar no resultado do Enade, foi o grau de escolaridade da família. Os resultados obtidos por eles, indicaram que a qualificação dos profissionais, o melhoramento da infraestrutura e a contratação de mais professores para atendimento extraclasse contribuem mais significativamente para melhores pontuações no Enade. Outro ponto interessante, é que alunos com uma ótima formação no ensino médio garantem melhor aprendizado na graduação e, consequentemente, contribuem para melhores notas no Enade. Deve-se observar que as universidades públicas obtiveram melhores resultados quando comparadas às instituições privadas de ensino. A metodologia das instituições públicas de ensino, com foco em produção científica, contra um foco maior no mercado de trabalho, adotado pelas faculdades privadas, pode ser a causa deste resultado. Além disso, os autores também mencionam que o rigor aplicado aos processos seletivos das instituições públicas também garantem a captação dos melhores alunos.

3.2 APRENDIZADO DE MÁQUINA APLICADO À GESTÃO DO ENSINO DE GRADUAÇÃO

Esta seção apresenta trabalhos de automatização de análise sobre documentos de gestão educacional com alguma similaridade com o que se pretende adotar durante o desenvolvimento deste projeto de mestrado.

Neto [107] apresenta uma avaliação da estrutura curricular do curso de Ciência da Computação por meio do aprendizado de máquina. Algumas motivações para a pesquisa são: a grande maioria dos graduandos não se forma no tempo estimado de 8 semestres, com um significante número de alunos que precisam de 10 a 12 semestres para a conclusão; outro ponto, é o alto nível de evasão, mesmo sendo considerado um dos melhores cursos de Ciência da Computação do país.

Os autores propuseram uma regressão linear para criar uma unidade sintética a partir de dados reais a fim de se comparar com outras unidades reais (as unidades são as grandes áreas de estudos, como Matemática, Programação, Teoria da Computação, Sistemas de Informação e Ciência da Computação). O conjunto de dados é formado por notas dos alunos que se matricularam de 2005 a 2016 no curso de Ciência da Computação da UFC (Universidade Federal do Ceará). Foi utilizado o Método de Controle Sintético, um regressor linear, para prever as notas dos graduandos em certa disciplina mediante as notas já obtidas por ele em outra matérias. A partir dos resultados obtidos, que giravam em torno de um erro médio de 0,13, foi possível deduzir o elo de dependência entre as matérias e planejar mais adequadamente a grade curricular, forçando ou relaxando pré-requisitos de disciplinas.

A autora Silva [108] discorre sobre a evasão dos alunos de cursos de graduação, que é um assunto muito comum na educação superior brasileira. Realizou-se um levantamento de que até 2018, apenas 36,68% dos 3.445.935 ingressantes na educação superior conseguiram concluir o curso. Este elevado número de evasão reflete principalmente no mercado de trabalho, que a cada ano necessita de mais mão de obra qualificada e se depara com um mercado escasso. Outro ponto que é afetado é a IES, pois a evasão de alunos afeta os indicadores de qualidade e, consequentemente, os investimentos que a instituição recebe. Nesse trabalho, foram utilizados dados do Inep para tentar compreender, antecipadamente, quais fatores podem influenciar o aluno a evadir. Nos dados do Inep, os alunos são classificados em: aluno formado,

cursando, com matricula trancada, desvinculado do curso, transferido para outro curso da instituição e finalizado ele, ou ocorreu o falecimento do aluno. Deste modo, aplicando métodos supervisionados, para predizer o desempenho final dos alunos, que era sobre 46 característica referente ao relatório do Inep, que eram classificados em formado, com matricula trancada, desvinculado do curso, transferido para outro curso da instituição e finalizado ele. As técnicas utilizadas foram Árvore de Decisão, Naïve Bayes, Regressão Logística e Redes Neurais e calculando a acurácia, sensibilidade e especificidade sobre eles. O melhor modelo foi a árvore de decisão com melhores resultados, tendo acurácia de 73%, sensibilidade de 60% e 89% de especificidade. Estes resultados expõe que há uma possibilidade de predizer antecipadamente a possível evasão de um aluno. As dificuldade citadas pelos autores foram: a qualidade dos dados, a diversidade das regiões para as quais cada IES tem que se adequar e os recursos computacionais.

Priyambada, Mahendrawathi e Yahya [109] realizaram a avaliação da grade curricular do curso de Sistemas de Informação (SI), utilizando técnicas de agrupamento sobre os dados dos alunos, formando, assim,*clusters* de acordo com a semelhança em seus perfis escolares. A motivação da pesquisa é dar condições para a instituição garantir que a grade curricular esteja dentro dos padrões educacionais estipulados pelos órgãos superiores e a missão da universidade. Cientes do motivo, basearam-se em um área que vem ganhando destaque no meio educacional, sendo ela: *Educational Process Mining* (EPM - Mineração de Processos Educacionais), contendo em seus princípios a exploração e a melhora dos processos relacionados com a educação, por meio dos próprios documentos escolares das instituições, dos alunos, dos órgãos governamentais e muitos outros.

O desenvolvimento do projeto seguiu as etapas de preparação dos dados, avaliação curricular e verificação. Na primeira fase, o objetivo é a mineração de dados por via do *plugin* ProM que ocasiona recuperação de dados, filtragem e a conversão destes dados recuperados. Na fase de avaliação é realizado o agrupamento por meio do algoritmo *k-Means*, detalhado anteriormente na Seção 2.2.2.5.5. Após testes preliminares, foi definido $k = 3$, pois o agrupamento dos dados em 3 *clusters* apresentou melhores grupos e resultados. Na última etapa, os *clusters* formados foram confrontados com a estrutura curricular, a fim de encontrar correlação entre ambos. Os resultados obtidos na análise dos agrupamentos mostraram que o *cluster* 2 agrupou alunos com um padrão de notas que permite a conclusão de curso em apenas 8 semestres. O *cluster* 3 apresentou a média mais baixa de notas, em relação aos demais,

necessitando de até 10 semestre para a conclusão do curso, em contrapartida o *cluster* 1 contém a melhor média de notas entre os agrupamentos, contudo há uma grande quantidade de alunos que concluíram o curso em até 10 semestres.

3.3 CONSIDERAÇÕES FINAIS

Este capítulo apresentou alguns trabalhos que realizaram análises de documentos de gestão educacional ou de dados do desempenho estudantil de algumas universidades com o objetivo de auxiliar no processo de concepção e gestão dos cursos de graduação, melhorar a grade curricular e propor alternativas para as instituições de ensino.

Neste contexto, propõem-se o uso de técnicas de aprendizado de máquina para a análise dos projetos pedagógicos dos cursos (PPCs), confrontando-os e correlacionando-os com outros documentos educacionais que viabilizam sua própria elaboração e o melhoramento. Está análise possibilitará descobrir quais aspectos presentes nos PPCs podem influenciar, por exemplo, resultados do Enade e, em contrapartida, como resultados do Enade ou outras avaliações podem induzir falhas nos PPCs e posteriormente ocasionar melhoras no mesmo.

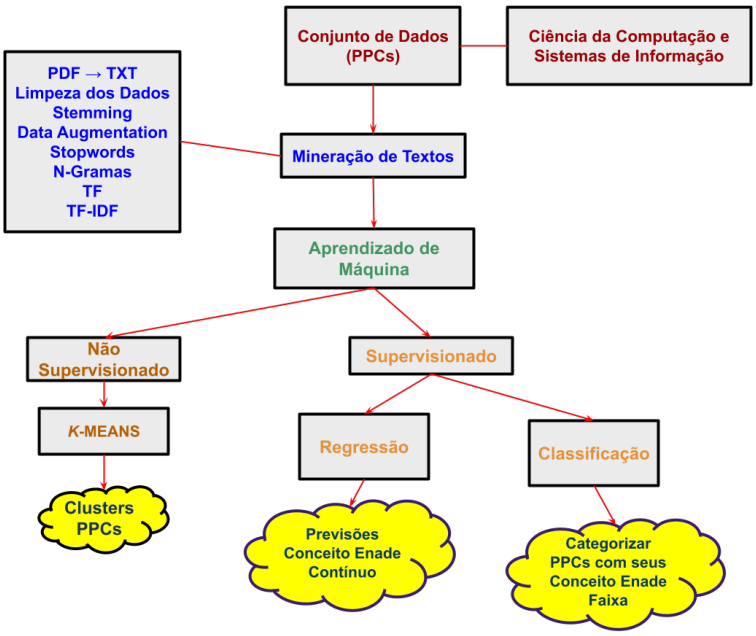
Durante a pesquisa, não foram encontrados trabalhos que utilizem projetos pedagógicos como base de dados para relacioná-los ou compará-los de forma automatizada às diretrizes educacionais vigentes ou ao desempenho estudantil no Enade ou em outros instrumentos de avaliação, mostrando que a pesquisa é promissora e pode colaborar com a gestão dos cursos de graduação no país.

4 AVALIAÇÃO EXPERIMENTAL

Neste trabalho, a avaliação experimental tem por objetivo utilizar dados textuais contidos nos projetos pedagógicos de cursos do ensino superior brasileiro para predizer os índices de avaliação dos mesmos. Neste sentido, os experimentos estão divididos em dois grupos.

Primeiro, foram empregados algoritmos de aprendizado de máquina não-supervisionado com o objetivo de agrupar os projetos pedagógicos, com base em suas similaridades. A segunda parte tem o foco na utilização do aprendizado de máquina supervisionado, em especial técnicas de Regressão, para predizer o Conceito Enade Contínuo (CEC) dos cursos, a partir dos seus respectivos PPCs, e a Classificação, com intuito de categorizar cada Conceito Enade Faixa (CEF) utilizando seu respectivo PPC no processo.

Figura 4.1: Mapa estrutural da formação do projeto prático.



Fonte: Elaborado pelo autor (2022).

Na Figura 4.1, tem-se uma síntese das etapas principais destes experimentos, que serão melhor descritos nas demais seções deste projeto. Os experimentos foram realizados utilizando a linguagem de programação *Python* [110] aliada à diversas bibliotecas, sendo algumas principais

como *Scikit-Learn* [111], *Pandas* [112], *Matplotlib* [113], NLTK (*Natural Language Toolkit*) [114].

4.1 O CONJUNTO DE DADOS

Para os experimentos, foram coletados projetos pedagógicos de cursos de Ciência da Computação (CC) e Sistemas de Informação (SI), de universidades públicas e privadas brasileiras. A escolha por estes cursos deu-se pela proximidade dos pesquisadores participantes deste projeto com as áreas destes cursos, facilitando o processo de validação e avaliação dos resultados. Foram utilizados dados de Universidades Federais, Estaduais e Municipais, IFs, CEFETs e instituições de ensino particulares. Embora haja obrigatoriedade de divulgação deste documento nos portais de todas as instituições públicas de ensino (Lei do acesso a informação pública [38]), nem todas as universidades apresentam o PPC em seus portais, ou o documento é difícil de ser encontrado, ou não foi disponibilizado para *download*, ou possuem diversas versões sem apontar qual é a vigente e em muitos casos datados com anos anterior à 2017. No caso das instituições particulares, poucas disponibilizam este documento mesmo tendo a obrigatoriedade, como as públicas, por via do Artigo 32 da Portaria Normativa nº 40, de 12 de dezembro de 2007 [115].

A quantidade de instituições de ensino superior, destacada na Introdução (Capítulo 1) e apresentada com mais detalhes no Anexo A, pode comparada com a quantidade de instituições a partir das quais foi possível coletar projetos pedagógicos para utilizar nesse trabalho. No conjunto de dados utilizado neste projeto, há documentos dos cursos de CC e SI de 123 instituições de ensino, sendo elas distribuídas da seguinte forma: 68 universidades públicas (50 federais, 16 estaduais e 2 municipais) e 8 universidades privadas; 1 centro universitário público estadual e 11 centros universitários privados; 1 faculdade pública municipal e 9 faculdades privadas; 24 IFs e 1 CEFET. Contudo, os dados presentes no Anexo A são gerais, deste modo, pode haver alguns instituições que não possuem oferta dos cursos de CC e SI.

A partir do portal E-Mec [37], a distribuição das 531 instituições de ensino que contém cursos de CC e/ou SI no modelo presencial é a seguinte: 91 universidades públicas (59 Federais, 27 Estaduais e 5 Municipais) e 73 privadas; 3 centros universitários públicos (1 Estadual e 2 Municipais) e 170 privados; 5 faculdades públicas municipais e 158 privadas; 24 IFs e 1 CEFET; 5 instituições da categoria "Especial", criadas por meio da Lei Estadual ou

Municipal, que recebem verbas públicas para seu funcionamento, mas não são gratuitas.

Das 531 IESs reportadas pelo E-Mec, tem-se que apenas 125 são públicas ($\approx 24\%$) e 406 privadas ($\approx 76\%$). Das 123 IESs que se utilizou para a coleta de PPCs, 95 são públicas (que equivalem a $\approx 77\%$ e 76% perante as 125 do total de IESs públicas), e apenas 28 são privadas (que equivalem a $\approx 23\%$ e $\approx 7\%$ perante as 406 totais).

As 123 IESs a partir das quais foi possível coletar PPCs correspondentes a $\approx 23\%$ das 531 IESs reportadas pelo E-Mec. Das 408 IESs não cobertas, cerca de $\approx 7\%$ são públicas (30 instituições) e $\approx 93\%$ são particulares (378 instituições). Ou seja, embora o percentual de instituições não cobertas na pesquisa seja alto ($\approx 77\%$), a maioria absoluta, $\approx 93\%$, são particulares. Isso se justifica, principalmente, pela falta de divulgação dos projetos pedagógicos de muitas destas instituições, demonstrando maior descaso das instituições privadas no cumprimento da Lei de Acesso à Informação. Além disso, há também instituições de ensino que passaram a ofertar os cursos de CC e SI recentemente e não puderam participar do Enade 2020, devido a pandemia, e que não colaboraram com dados desta pesquisa.

No total, foram separados 223 projetos pedagógicos, a partir das 123 IESs mencionadas acima, os quais estão quantitativamente distribuídos na Tabela 4.1, de acordo com a categoria da instituição de ensino e o Conceito Enade Faixa do curso (que possui valores 1, 2, 3, 4, 5 ou SC-Sem Conceito).

Na Figura 4.2 apresenta-se a distribuição das 223 amostras do conjunto de dados. Os documentos ilustrados como SC (Sem Conceito) são de cursos que foram iniciados nos últimos anos e ainda não passaram pelo exame Enade e, consequentemente, ainda não foram avaliados. De forma mais detalhada, na Figura 4.3 é ilustrado a distribuição das 176 amostras de PPCs que contém avaliação Enade, organizadas em relação ao Conceito Enade Faixa (de 1 a 5) e com base no Conceito Enade Contínuo, de forma crescente.

Para que os projetos pedagógicos pudessem serem utilizados como conjunto de dados, foi realizada a conversão dos arquivos em seus formatos originais (PDF - *Portable Document Format*) para um formato de texto puro, utilizando a ferramenta *CERMINE* (Tkaczyk et al. [116]). Em seguida, foi realizar o tratamento adequado para que algoritmos de aprendizado de máquina conseguissem atuar sobre os dados. Para tal, foi executada uma limpeza em cada documento, de modo a retirar os links de páginas *web*, pontuações,

caracteres especiais, e converter todo o texto em letras minúsculas.

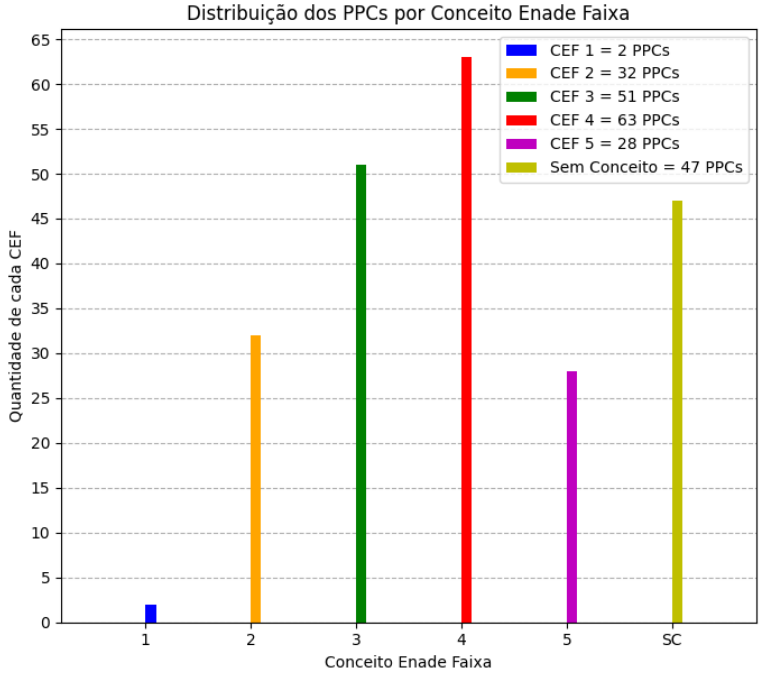
Tabela 4.1: Distribuição dos PPCs dos cursos de Ciência da Computação e Sistemas de Informação de acordo com a categoria da instituição de ensino e do Conceito Enade Faixa do curso.

Curso: Ciência da Computação						
Categoria da Instituição de Ensino	**Quantidade de PPCs por Conceito Enade Faixa**					
	1	**2**	**3**	**4**	**5**	**SC**
Universidades Públicas Federais	0	3	12	23	17	5
Universidades Públicas Estaduais	1	3	4	6	2	3
Universidade Pública Municipal	0	0	0	1	0	0
Centros Federais de Educ. Tecnológica	0	0	0	1	0	0
Institutos Federais de Educação	0	1	5	2	0	14
Instituições Privadas	1	5	3	1	1	2
Curso: Sistemas de Informação						
Categoria da Instituição de Ensino	**Quantidade de PPCs por Conceito Enade Faixa**					
	1	**2**	**3**	**4**	**5**	**SC**
Universidades Públicas Federais	0	7	12	14	4	4
Universidades Públicas Estaduais	0	4	3	4	1	4
Universidade Pública Municipal	0	1	0	1	0	0
Centros Federais de Educ. Tecnológica	0	0	0	0	1	0
Institutos Federais de Educação	0	1	5	9	2	14
Instituições Privadas	0	7	7	1	0	1
Total de PPCs	2	32	51	63	28	47

Fonte: Elaborado pelo autor (2022).

Após experimentos iniciais com os documentos na íntegra, verificou-se a necessidade de retirar os conteúdos iniciais dos PPCs, como apresentação, histórico do curso, dados organizacionais das instituições de ensino. Foram mantidas as seções referentes ao próprio curso, como estrutura curricular, disciplinas obrigatórias e optativas, ementário, bibliografia.

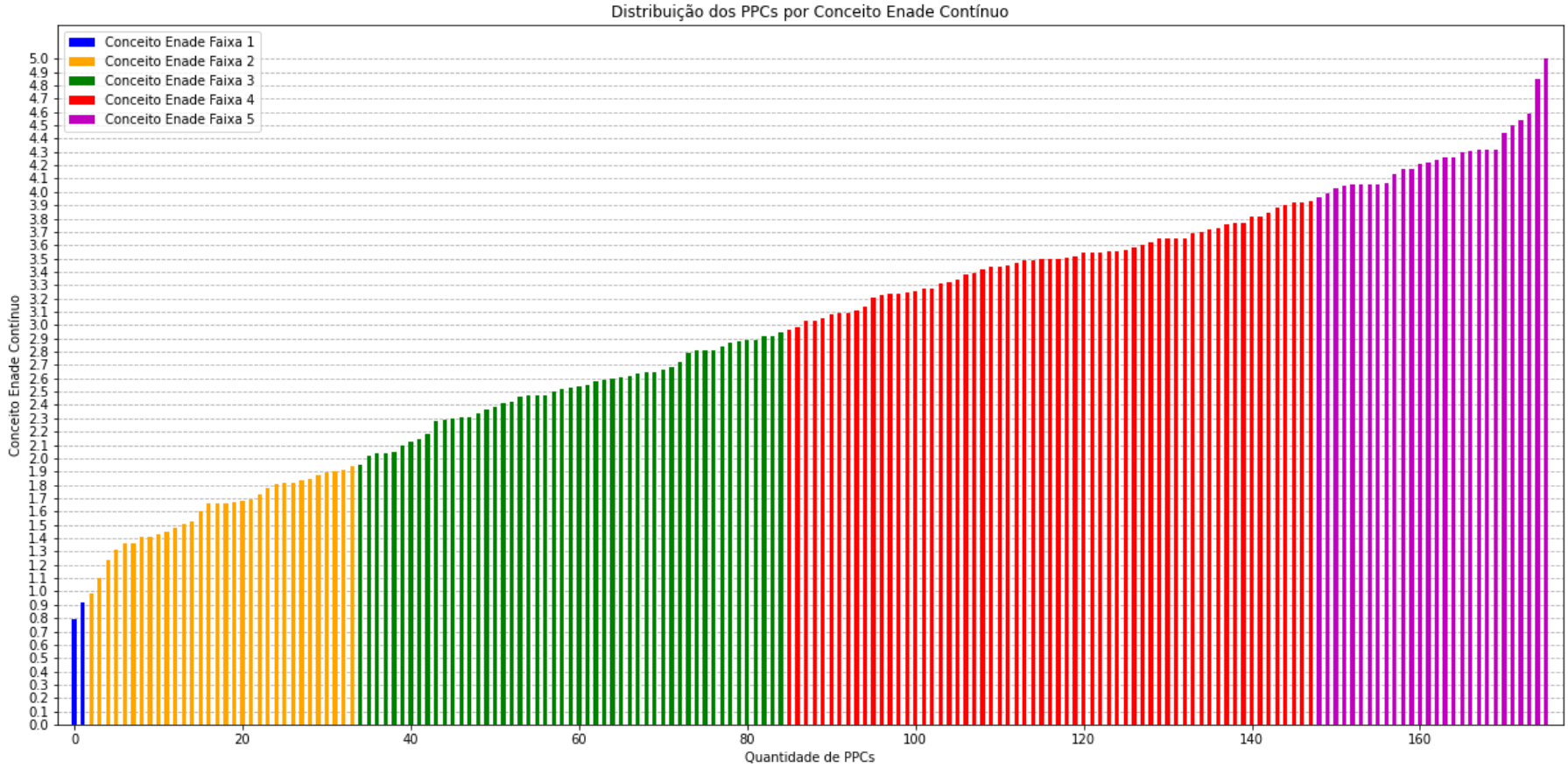
Figura 4.2: Distribuição dos PPCs de acordo com o Conceito Enade Faixa.



Fonte: Elaborado pelo autor (2022).

Ademais, foram utilizadas técnicas como: *Stemming*, levando as palavras, presentes nos PPCs, para seus radicais, reduzindo assim a quantidade de termos repetidos e restando apenas o necessário, ou o verdadeiro significado; e *Data Augmentation*, com a substituição das palavras por seus sinônimos, possibilitando o aumento do conjunto de dados na etapa de Classificação. Todos estes processos serão descritos, como foram utilizadas adequadamente, nas Seções seguintes.

Figura 4.3: Distribuição dos PPCs de acordo com o Conceito Enade Contínuo.



Fonte: Elaborado pelo autor (2022).

4.2 METODOLOGIA DE PRÉ-PROCESSAMENTO, PARAMETRIZAÇÃO E AVALIAÇÃO PARA AGRUPAMENTO

Esta seção descreve a metodologia de pré-processamento, parametrização e avaliação aplicada ao modelo de agrupamento.

4.2.1 Pré-processamento para Agrupamento

O conjunto de dados destinado à Agrupamento utilizava os 223 documentos, sobre os quais foi aplicada a técnica de *Stemming*, a fim de levar as palavras para seu radical, mantendo seu significado e retirando eventuais erros de escrita.

Em seguida, foram aplicadas as técnicas de *TF* e *TF-IDF*, para remoção das *stopwords* e encontrar a frequência de unigramas e bigramas, considerando-se uma frequência miníma em documentos (DF) igual a três. Ressalta-se que a utilização de TF, foi utilizada apenas para possibilitar o Corte de Luhn de maneira mais precisa; os demais experimentos utilizam os dados oriundos de *TF-IDF*.

Inicialmente, haviam 146.927 termos, dos quais 15.899 eram unigramas (10,82% do total) e 131.028 eram bigramas (89,18% do total). Ao aplicar o Corte de Luhn, o montante foi o seguinte: 5.442 unigramas (34,22% do conjunto com CL e 3,70% em relação ao conjunto todo) e 19.938 bigramas (15,21% do conjunto com CL e 13,57% em relação ao conjunto todo), totalizando 25.380 termos (17% em relação ao conjunto completo).

Finalizada essa etapa, foram criadas duas versões do conjunto de dados para agrupamento, uma abrangendo unigramas/bigramas e outra contendo apenas unigramas. Salienta-se que houve outras metodologias anteriores a essa, contudo nenhuma delas retornou bons resultados e foram dispensadas.

4.2.2 Agrupamento dos PPCs

O algoritmo de agrupamento utilizado foi o *k-Means*, com diferentes combinações de parâmetros, para ambas versões do conjunto de dados (completo ou somente com unigramas):

- `n_clusters = range(2, 20)` variando de 2 a 20;
- `n_init = 10`, número de vezes que o algoritmo era executado com diferentes sementes de centróide;
- `init = 'k-means++'`, método de inicialização do modelo;
- `max_iter = 100`, número de iterações máximas que o modelo poderá ser executado.

Para cada combinação de parâmetros testadas, a métrica *Silhueta* era responsável por avaliar o desempenho do algoritmo.

4.3 METODOLOGIA EXPERIMENTAL PARA CLASSIFICAÇÃO

Nesta seção, descreve-se a metodologia de pré-processamento, parametrização e avaliação utilizadas no processo supervisionado de classificação.

4.3.1 Pré-processamentos para Classificação dos Dados

Para que o conjunto de dados seja utilizado como entrada para os algoritmos de classificação ou regressão, todos os dados precisam estar rotulados adequadamente, exigindo, assim, que todos os PPCs do conjunto sejam referentes à cursos que já tenham seu conceito Enade Contínuo Faixa disponibilizado. Devido a isso, todos os PPCs sem Conceito Enade Faixa foram retirados do conjunto de dados. Ademais, também foram removidos outros dois PPCs (ambos do CEF 1), que em execuções anteriores apresentaram *outliers*, restando, portanto, 174 documentos no *dataset*.

Em seguida foi aplicada a técnica de *Data Augmentation*, para os CEFs 2 e 5, que possuíam menores quantidades de amostras, utilizando a técnica de Substituição por Sinônimos. Na prática, um documento de tais conceitos era escolhido, e sendo percorrido palavra por palavra, de modo que, a cada termo era selecionado a primeira opção de sinônimo para a mesma, quando disponível. Ao final do processo tinha-se, então, dois documentos, o original e a sua nova versão contendo os sinônimos do arquivo original. A seguir é

apresentado um pequeno trecho inicial de um documento, com sua versão original e com a substituição por sinônimos:

- **Documento Original:** 'organização', 'curricular', 'preparar', 'profissional', 'formação', 'conceitual', 'teórica', 'áreas', 'formação', 'básica', [...].
- **Documento Sinônimo:** 'gestão', 'curricular', 'planejar', 'especialista', 'composição', 'teórico', 'zonas', 'composição', 'fundamentais', [...].

Sendo assim, possibilitou-se realizar o balanceamento entre as classes. A partir dos dados apresentados na Seção 4.1, nota-se que os CEFs 2 e 5 possuem poucas amostras em relação às demais e que os CEF 3 e 4 são as majoritárias. Este desbalanceamento dificulta que os classificadores façam a distinção correta das classes. A partir da aplicação de DA, o CEF 2 passou de 32 amostras para 64, e o CEF 5 de 28 amostras para 56, totalizando 234 amostras. Ao final, foi aplicado *Stemming* sobre todo o *dataset*.

Após preparar os dados, foi utilizado Corte de Luhn para gerar matrizes de unigramas e bigramas. A matriz original possuía 148.940 termos, sendo 17.619 unigramas (11,83% do total) e 131.321 bigramas (89,17% do total). Ao aplicar o Corte de Luhn (CL), o montante foi o seguinte: 5.435 unigramas (15,60% do conjunto com CL e 3,67% em relação ao conjunto todo) e 29.410 bigramas (84,40% do conjunto com CL e 19,74% em relação ao conjunto todo), totalizando 34.845 termos (23,41% em relação ao conjunto completo).

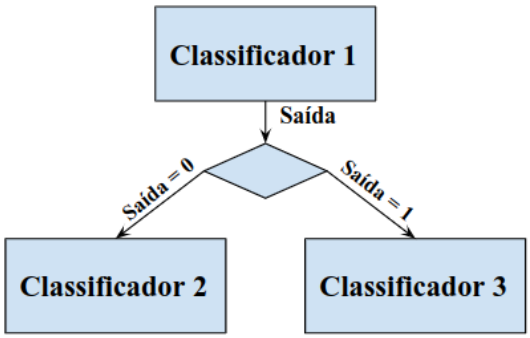
4.3.2 Classificação dos PPCs de acordo com o Conceito Enade (Faixa)

Neste trabalho, adotou-se uma abordagem hierárquica para classificação dos documentos, tendo cinco execuções, no qual cada execução possui três repetições. A motivação para tal decisão baseia-se na estruturação ordinal dos Conceitos Enade Faixa (aqui, variando de 2 a 5), utilizados como saída para os 45 classificadores. Com isso, a definição dos três classificadores presentes nesta hierarquia é a seguinte:

- **Classificador 1:** para diferenciar classe 0 (CEF 2 e 3) e classe 1 (CEF 4 e 5).

- **Classificador 2:** para diferenciar classe 0 (CEF 2) e classe 1 (CEF 3).
- **Classificador 3:** para diferenciar classe 0 (CEF 4) e classe 1 (CEF 5).

Figura 4.4: Exemplo estrutura de relação e dependência dos classificadores.



Fonte: Elaborado pelo autor (2022).

Para melhor compreensão, na Figura 4.4 é retratada a relação de dependência dos classificadores, sendo que a saída do Classificador 1 são os dados de entrada para os Classificadores 2 ou 3. Portanto, o Classificador 1 distingue as classe 0 e 1, em seguida encaminha ao modelo subsequente, se acaso a predição foi 0 irá para o Classificador 2 ou se a predição for 1, os dados serão destinados ao Classificador 3.

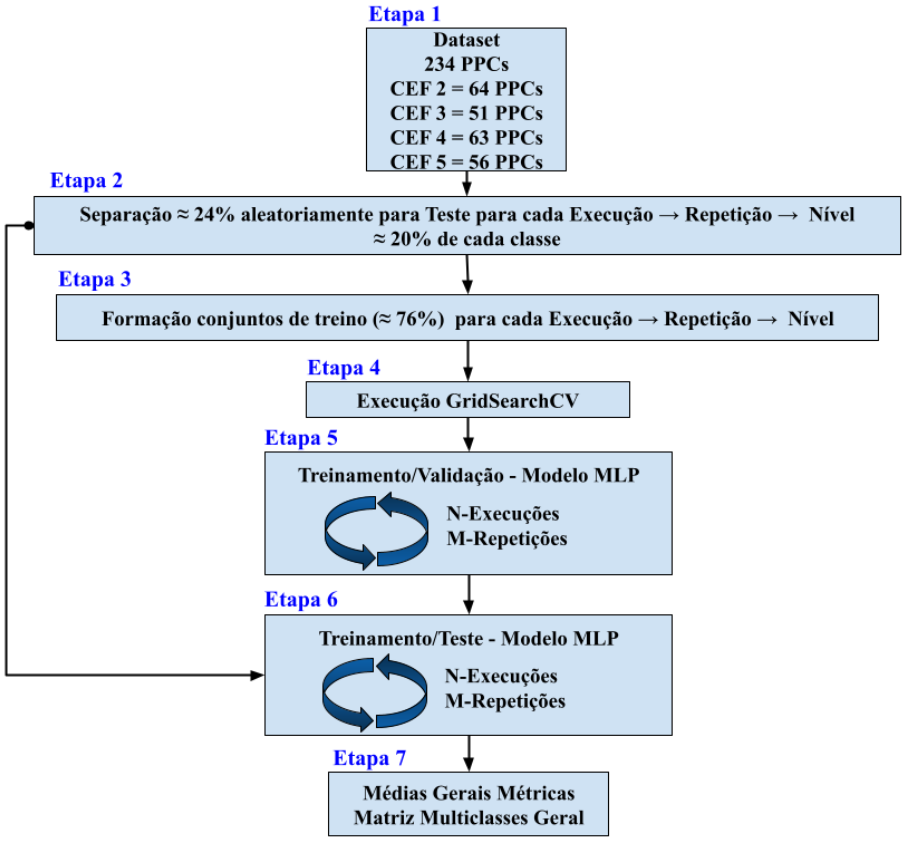
Eram destinados ao treino $\approx 76\%$ dos dados e $\approx 24\%$ para teste (sendo $\approx 20\%$ de cada Conceito Enade Faixa e escolhidos de forma aleatória) e os conjuntos para em cada situação eram os seguintes:

- **Classificador 1:**
  - Treino: 189 exemplos, sendo 93 amostras para o classe 0 (CEF 2 = 52 e CEF 3 = 41) e 96 para o classe 1 (CEF 4 = 51 e CEF 5 = 45).
  - Teste: 45 exemplos, sendo 22 amostras classe 0 (CEF 2 = 12 e CEF 3 = 10) e 23 para o classe 1 (CEF 4 = 12 e CEF 5 = 11).
- **Classificador 2:**
  - Treino: 93 exemplos para treino, sendo 52 amostras para o classe 0 (CEF 2 = 52) e 41 para o classe 1 (CEF 3 = 41).

– Teste: 22 exemplos, sendo 12 amostras classe 0 (CEF 2 = 12) e 10 para o classe 1 (CEF 3 = 10).

- **Classificador 3:**
  - Treino: 96 exemplos, sendo 51 amostras para o classe 0 (CEF 4 = 51) e 45 para o classe 1 (CEF 5 = 45).
  - Teste: 23 exemplos, sendo 12 amostras classe 0 (CEF 4 = 12) e 11 para o classe 1 (CEF 5 = 11).

Figura 4.5: Estrutura Simplificada da Classificação.



Fonte: Elaborado pelo autor (2022).

Salienta-se que que houve a utilização de outros modelos de classificação, como SVC (*Support Vector Classification*), a qual refere-se a uma variação do SVM destinado à classificação, e *k-NN*. Entretanto, ambos retornavam baixa acurácia. Deste modo, optou-se por manter apenas o modelo MLP nas seguintes etapas.

Na Figura 4.5 ilustra-se a estrutura simplificada da metodologia para o desenvolvimento e avaliação dos classificadores. Cada uma das etapas mencionadas nesta ilustração serão descritas nos próximos parágrafos.

Etapas 1, 2 e 3: As três primeiras etapas fazem referência ao *dataset* e a formação dos conjunto de treino e teste (explicados anteriormente).

Etapa 4: A Tabela 4.2 apresenta os valores utilizados como candidatos para o refinamento dos parâmetros do modelo MLP, bem como os valores de k utilizados validação cruzada *Stratified K-Fold* (métodos anteriormente explicados na Seção 2.2.2.6.2). Como há 45 classificadores nesta árvore de classificação, optou-se por direcionar todos os melhores parâmetros ao Apêndice A.

Tabela 4.2: Hiperparâmetros do modelo MLP na execução de *GridSearchCV*, juntamente com a variação da *Stratified K-Fold*.

Parâmetro	**Função**	**Valores**
`hidden_layer_size`	Quantidade de camadas ocultas	(10,), (25,), (50,), (100,), (200,), (10,-10), (25,25), (50,50), (100,100), (200,200)
`Activation`	Função de ativação para a camada oculta	'relu', 'logistic'
`Solver`	Otimização de peso	'sgd', 'adam', 'lbfgs'
`Alpha`	Regularização de peso	0.1, 0.01, 0.001, 0.0001
`learning_rate`	Taxa de aprendizado para atualizações de peso	'constant','adaptative'
`early_stopping`	Interrupção antecipada usada para encerrar o treinamento quando a pontuação de validação não estiver melhorando	True
`k` do Classificador 1	Quantidade de *folds* da *Stratified K-Fold*	k=3
`k` do Classificador 2 e 3	Quantidade de *folds* da *Stratified K-Fold*	k=2

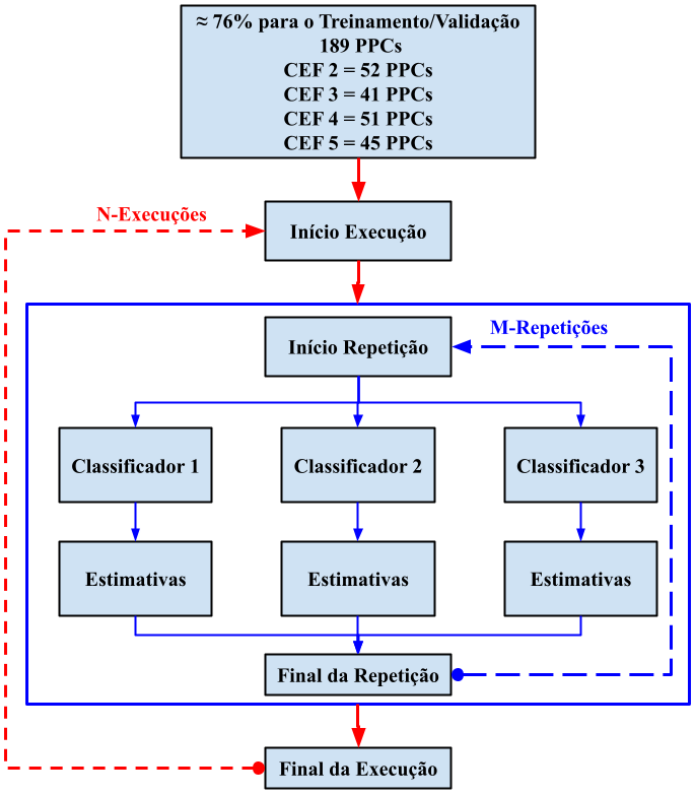
Fonte: Elaborado pelo autor (2022).

Etapas 5 e 6: A Figura 4.6 apresenta a estrutura de treino/validação, considerando que a MLP é treinada e gera estimativas para a etapa seguinte. Já para a Etapa 6, sendo a fase real de treino/teste, os modelos eram treinados. Para ambas etapas há "N-Execuções" e "M-Repetições"; nos experimentos,

foram utilizados $N = 5$ e $M = 3$, ou seja, as etapas eram executadas cinco vezes, no qual possuíam três repetições com três níveis de classificação.

Etapa 7: Foi calculada a acurácia (*accuracy*) média de cada nível separadamente, também foi elaborada da matriz multiclasses geral e foram deliberadas as métricas *Micro-Avereged Precision e Recall* e *accuracy* geral sobre ela.

Figura 4.6: Síntese da Estrutura de Treino/Validação da Classificação.



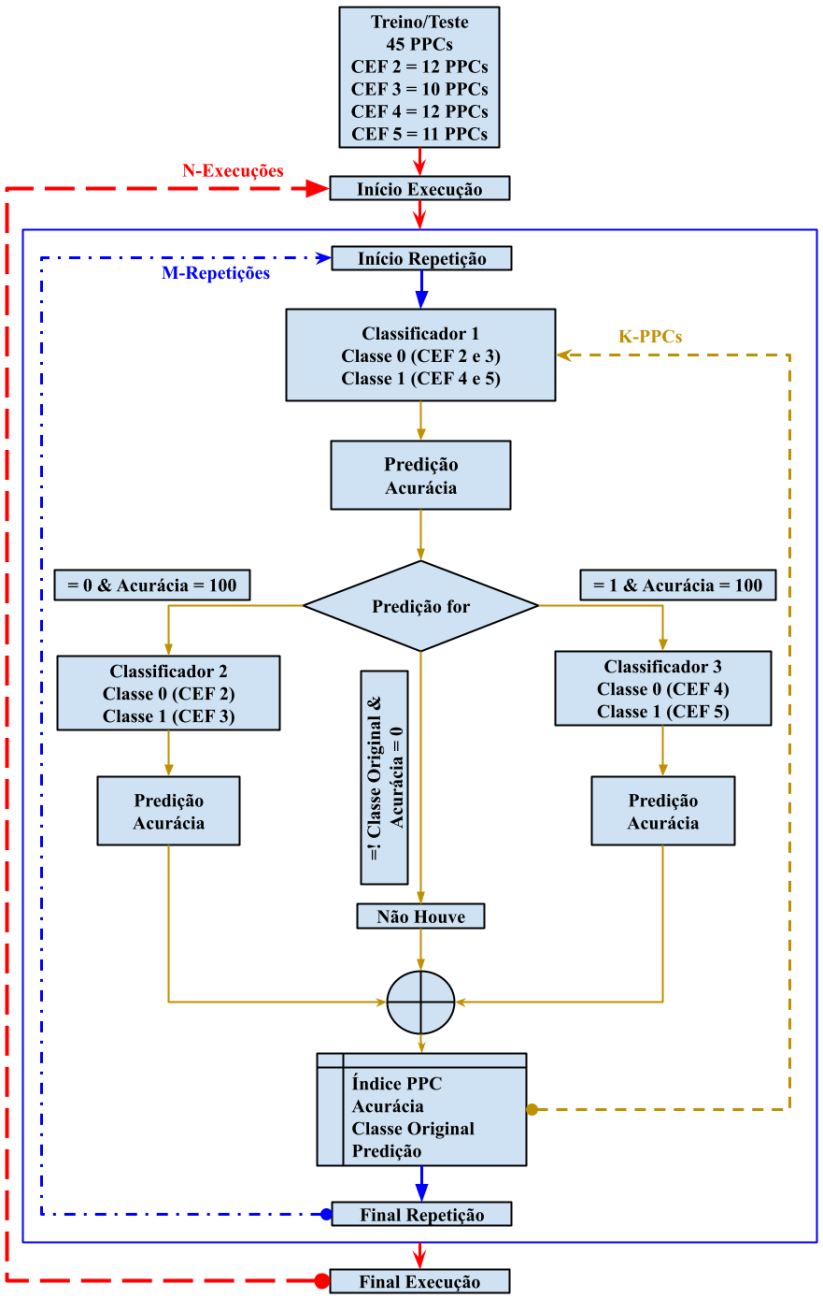
Fonte: Elaborado pelo autor (2022).

A estrutura da árvore (Figura 4.7) é a seguinte: o primeiro classificador é responsável por classificar entre classe 0 (CEF 2 e CEF 3) e classe 1 (CEF 4 e CEF 5); o classificador dois, classificar entre classe 0 (CEF 2) e classe 1 (CEF 3); e classificador três, classificar entre classe 0 (CEF 4) e classe 1 (CEF 5).

Os dados de teste eram passados por essa estrutura um por vez. Deste modo, se o classificador 1 classificasse corretamente o documento em questão, ele era encaminhado para o classificador 2 ou 3 (dependendo de qual era sua classe). Porém, ao errar no primeiro nível, o documento não era repassado

para as camadas seguintes.

Figura 4.7: Estrutura da Árvore de Classificação.



Fonte: Elaborado pelo autor (2022).

Ao final das repetições de cada execução, foi calculada a acurácia média de cada classificador, bem como a matriz multiclasses unificada (junção das

três matrizes das repetição da execução) e as métricas *Averaged Precision e Recall* sobre ela. Ao finalizar as cinco execuções, uma matriz multiclasses unificada (utilizando a soma de todas as matrizes unificadas de cada execução) e o cálculo das métricas, mencionadas acima, sobre ela.

Tabela 4.3: Exemplo sintético da Estrutura gerada ao final da hierarquia de classificação.

	PPC	**Predição**	**Classe**	**Acurácia**
Classificador 1	16	0	0	100
	227	0	0	100
	97	1	0	0
	50	1	1	100
	2	0	1	0
	100	1	1	100
	PPC	**Predição**	**Classe**	**Acurácia**
Classificador 2	16	1	0	0
	227	0	0	100
	97	Não Houve	Não Houve	Não Houve
	PPC	**Predição**	**Classe**	**Acurácia**
Classificador 3	50	0	0	100
	2	Não Houve	Não Houve	Não Houve
	100	1	0	0

Fonte: Elaborado pelo autor (2022).

A Tabela 4.3 ilustra o resultado do processo de classificação de um subconjunto de seis PPCs, dos quais quatro deles, identificados por 16, 227, 50 e 100, tiveram acerto na classificação do classificador 1 e os outros dois, identificados por 97 e 2, tiveram erro de classificação de classificador 1. Os PPCs classificados incorretamente no classificador 1 não são classificados nos classificadores 2 ou 3 e, portanto, as linhas correspondentes a estes PPCs nos classificadores 2 e 3 foram preenchidos com "Não Houve". Como outro exemplo, o PPC identificado com 16 teve acerto na classificação de classificador 1, por isso a acurácia no PPC 16 neste classificador recebeu o valor 100. No classificador 2, esse PPC não foi classificado corretamente e, então, recebeu valor 0 de acurácia no classificador 2.

4.4 METODOLOGIA DE PRÉ-PROCESSAMENTO, PARAMETRIZAÇÃO E AVALIAÇÃO PARA REGRESSÃO

Nesta seção será apresentada a metodologia de pré-processamento, parametrização e avaliação utilizada durante o processo de aplicação de regressão para predição da nota do Enade (contínuo) a partir do PPC de um curso.

4.4.1 Pré-Processamentos para Regressão

Assim como para a aplicação dos classificadores, também foram retirados os PPCs "sem conceito" do conjunto de dados, a fim de utilizá-lo para desenvolvimento dos regressores. Ademais, também foram retiradas as amostras que pertenciam ao conceito Enade Faixa 1, resultando em 174 amostras para esta etapa, no qual sofreram limpeza dos dados e a aplicação da *Stemming*.

Com os dados preparados, o passo seguinte é a elaboração das matrizes esparsas com TF-IDF e TF, removendo as *stopwords*, gerando unigramas e bigramas com uma frequência miníma em documentos (DF) igual a três, possibilitando aplicar o Corte de Luhn. Ao final do processo, tem-se dois conjunto de dados ou matrizes: um com unigramas e bigramas, e outro formada unicamente de unigramas.

As matrizes originais dispunham de 116.558 termos, sendo 13.843 unigramas (11,87% do total) e 102.715 bigramas (88,13% do total). Ao aplicar o Corte de Luhn o montante foi o seguinte: 5.458 unigramas (15,41% do conjunto com CL e 4,68% em relação ao conjunto todo) e 29.952 bigramas (84,59% do conjunto com CL e 25,69% em relação ao conjunto todo), totalizando 35.410 termos (30,37% em relação ao conjunto completo).

4.4.2 Modelos de Regressão para a Predição do Conceito Enade Contínuo a partir dos PPCs

Após o pré-processamento descrito na Seção 4.4.1, o passo seguinte foi utilizar este conjunto contendo apenas unigramas para a aplicação do método de regressão, sendo uma árvore de decisão de três elementos, a saber:

- **Classificador 1:** modelo para diferenciar classe 0 (CEF 2 e 3) e atributo 1 (CEF 4 e 5).
- **Regressor 2:** modelo para predizer o Conceito Enade Contínuo. Os valores de 0,95 à 1,94 correspondem à CEF 2 e os valores de 1,95 à 2,94 correspondem à CEF 3.
- **Regressor 3:** modelo para predizer o Conceito Enade Contínuo. Os valores de 2,95 à 3,94 correspondem à CEF 4 e os valores de 3,95 à 5 correspondem à CEF 5.

Apenas uma execução foi realizada, a qual continha em si três repetições, tendo um classificador e dois regressores cada. Assim, foram preparados preparados conjuntos de treino e teste para cada nível, totalizando 18 conjuntos, sendo nove para treino e nove para teste.

O conjunto de dados foi separado entre treino/validação e teste, de modo que, $\approx 76\%$ para treino/validação e $\approx 24\%$ para teste ($\approx 20\%$ de cada CEF, sendo selecionados de forma aleatória), selecionando documentos com Conceito Enade Faixa para classificação e Conceito Enade Contínuo para regressão.

A distribuição dos dados em cada *dataset* para as três repetições foram as seguintes:

- **Classificador 1:**
  - Treino: 140 exemplos, sendo 67 amostras para o atributo 0 (CEF 2 = 26 e CEF 3 = 41) e 73 para o atributo 1 (CEF 4 = 51 e CEF 5 = 22).
  - Teste: 34 exemplos, sendo 16 amostras atributo 0 (CEF 2 = 6 e CEF 3 = 10) e 18 para o atributo 1 (CEF 4 = 12 e CEF 5 = 6).
- **Regressor 2:**
  - Treino: 67 exemplos, sendo 26 amostras de valores contínuos (CEF 2) e 41 de valores contínuos (CEF 3).
  - Teste: 16 exemplos, sendo 6 amostras de valores contínuos (CEF 2) e 10 de valores contínuos (CEF 3).
- **Regressor 3:**

- 73 exemplos, sendo 51 amostras de valores contínuos (CEF 4) e 22 para valores contínuos (CEF 5).
- Teste: 18 exemplos, sendo 12 amostras de valores contínuos (CEF 4) e 6 para valores contínuos (CEF 5).

Foram realizados ajustes dos parâmetros de cada modelo mediante ao *GridSearchCV*; os hiperparâmetros para cada modelo apresentados no Apêndice B.1, os parâmetros candidatos da MLP foram apresentados na Seção 4.3.2 e os melhores parâmetro de cada modelo se encontram disponível no Apêndice B.2.

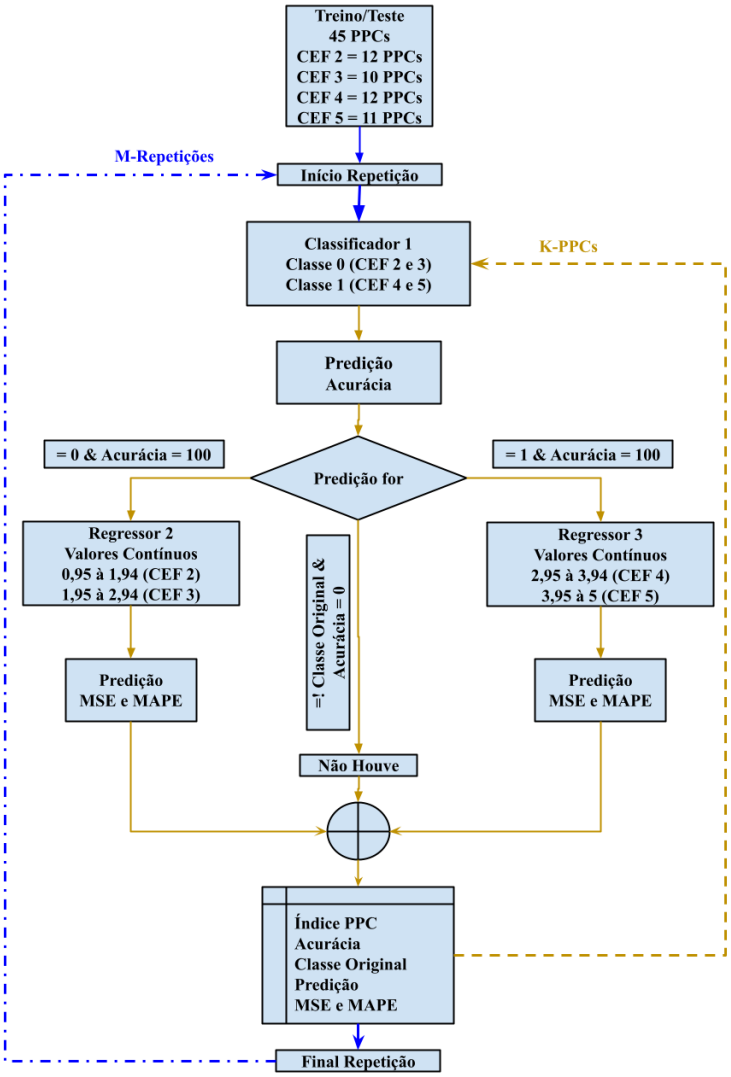
Também foi aplicada a validação cruzada, na qual, para o classificador 1: *Stratified k-Fold Cross-Validation* com $k = 3$, $Shuffle = True$ e $random\_state = 1$. Para os regressores 2 e 3 utilizou-se a *k-Fold Cross-Validation*, com $k = 2$.

A estrutura de treino/validação é similar à Figura 4.6, tendo no classificador 1 o modelo MLP como único candidato e gerando suas estimativas médias aritméticas da métrica *accuracy* nas três repetições. Nos regressores 2 e 3, foram utilizados modelos SVR, o regressor *k-NN* e a MLPRegressor. Ao final de cada repetição, foram geradas os erros com bases nas métricas MSE e MAPE. Durante etapas preliminares dos experimentos de regressão, utilizava-se outras métricas como a MAE, RMSE e R^2 [96, 97, 117]. Contudo, optou-se por aperfeiçoar os modelos utilizando apenas as duas métricas MSE e MAPE.

Ao analisar as estimativas de cada classificador e regressores para cada repetição, é escolhido o melhor algoritmo para ser fixo durante o treino/teste. Primeiramente os modelos são treinados. Deste modo, cada modelo atua sobre todos os dados destinados ao treino e, em seguida, é realizado o teste, de forma que cada documento passe pela árvore ilustrada na Figura 4.8, um por vez.

Ao final, são calculadas as métricas *Accuracy* (classificador 1), MSE e MAPE (regressores 2 e 3), a matriz de confusão do primeiro classificador, gráficos comparativos entre o valor original e o predito pelo modelo (regressores 2 e 3). Após três repetições, é feita a média aritmética da acurácia (*accuracy*) do primeiro classificador e das métricas da regressão.

Figura 4.8: Estrutura da Árvore de Regressão.



Fonte: Elaborado pelo autor (2022).

Na Tabela 4.4, é apresentado uma versão sintética da planilha gerada ao final de cada repetição da árvore de regressão. Ao passar pelo classificador 1, a entrada é encaminhada para o regressor seguinte, de acordo com o resultado da classificação. Contudo, se houver erro na predição no classificador 1, não será realizado processo de regressão no nível seguinte, resultando, portanto, nas entradas do tipo "Não Houve" na Tabela, para esse documento.

Tabela 4.4: Exemplo sintético da estrutura gerada ao final da árvore de regressão.

Classificador 1	**PPC**	**Predição**	**Classe**	**Acurácia**	**-**
	15	0	0	100	-
	7	1	0	0	-
	45	1	1	100	-
	28	0	1	100	-
Regressor 2	**PPC**	**Predição**	**Valor Ori.**	**MSE**	**MAPE**
	15	2,1155	2,0212	0,008	4,66
	7	Não Houve	Não Houve	Não Houve	Não Houve
Regressor 3	**PPC**	**Predição**	**Valor Ori.**	**MSE**	**MAPE**
	45	3,5777	3,6947	0,013	3,16
	28	Não Houve	Não Houve	Não Houve	Não Houve

Fonte: Elaborado pelo autor (2022).

4.5 Considerações Finais

Neste capítulo, foram expostas informações referente à elaboração do conjunto de dados, composto por projetos pedagógicos de cursos de Ciência da Computação e Sistemas de Informação. Foram apresentadas as técnicas de pré-processamento adotadas e, por fim, a metodologia empregada para a aplicação do algoritmo k-Means para agrupamento dos PPCs, métodos para classificação para de PPCs de acordo com o Conceito Enade (Faixa) e regressão para predição do Conceito Enade (Contínuo).

5 ANÁLISE DOS RESULTADOS EXPERIMENTAIS

Nesta Seção, serão explanados os resultados dos experimentos de validação e avaliação dos modelos, cuja metodologia foi descrita no Capítulo 4. Além dos resultados, haverá explicações sobre as melhores soluções, detalhes que ocasionaram resultados insatisfatórios e pontos para potenciais melhorias nos resultados

5.1 AVALIAÇÃO DO MODELO DE AGRUPAMENTO DOS PPC'S

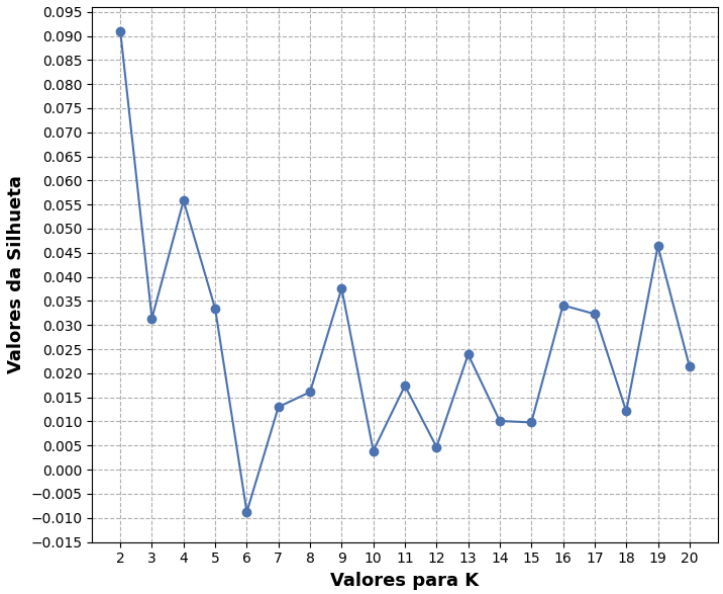
Tanto para a execução do *k-Means* com o conjunto completo (unigramas/brigramas), quanto utilizando somente unigramas, foi executado o modelo em um laço de repetição variando o valor de k de 2 a 20, o qual refere-se a quantidade de clusters do modelo. A métrica Silhueta foi calculada ao final de cada iteração. Deste modo, ao final do processo, com base no valor da Silhueta, foram escolhidos os melhores valores de k.

Após a execução com o conjunto completo (unigramas e bigramas), Os valores de k selecionados foram: 2, 4, 9, 13 e 19, com as respectivas silhuetas, 0,091, 0,055, 0,037, 0,023 e 0,046. Para execução com unigramas, os valores de k foram: 2, 4, 8, 12 e 17, com suas respectivas silhuetas, 0,01, 0,029, 0,0289, 0,034 e 0,023. Na Figura 5.1 são apresentadas as silhuetas para todos os valores de k, na execução completa 5.1(a) e apenas com unigramas 5.1(b).

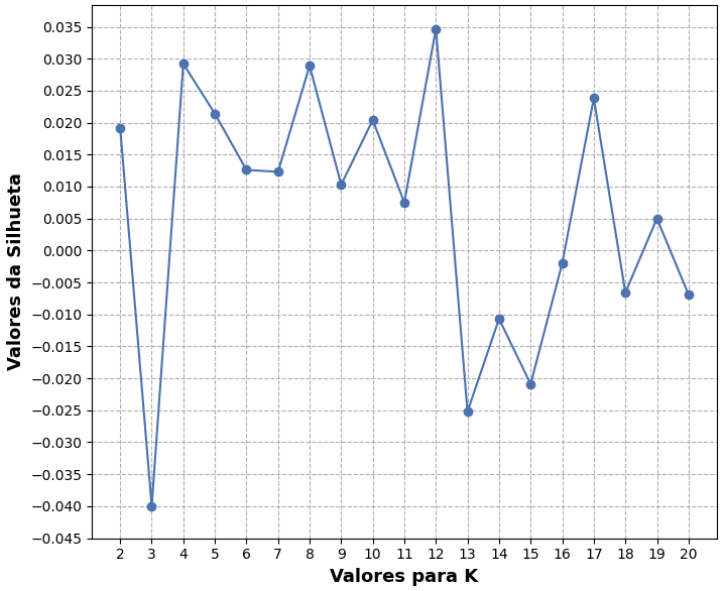
Ao elaborar a representação gráfica dessas execuções, para apresentar a formação dos agrupamentos, notou-se que a execução com apenas unigramas e valor fixo de 12 *clusters* foi a que melhor configuração na formação dos agrupamentos. Na Figura 5.2, é exposta a representação em 2D, que em sua maioria possui *clusters* com muitos documentos. Contudo, houve a formação de pequenos agrupamentos, como: dois *clusters* com 3 PPCs, um *cluster* com 2 PCCs e e cinco *clusters* contendo 1 PPC.

Figura 5.1: Resultado da Silhueta no laço de repetição para execução com conjunto completo e somente unigramas.

(a) *k-Means* Utilizando Conjunto Completo



(b) *k-Means* Utilizando Conjunto Unigramas

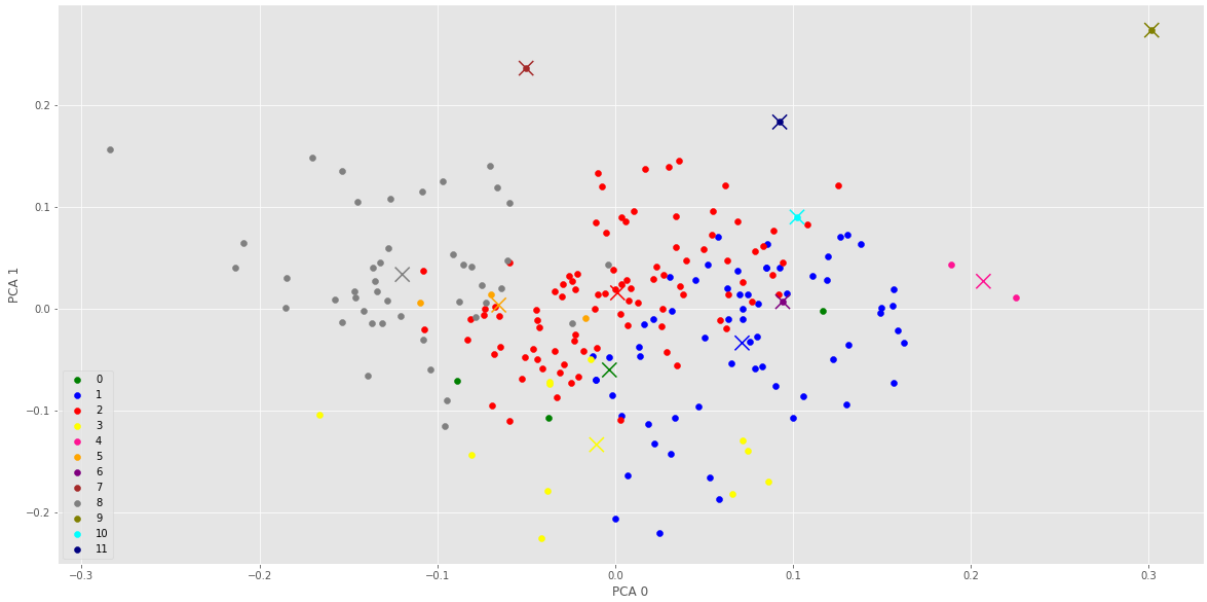


Fonte: Elaborado pelo autor (2022).

Com a ilustração gráfica pronta, o passo seguinte foi analisar os *clusters* formados, a fim de compreender porque ficaram com poucos documentos. Para tal análise, os termos presentes em seus agrupamentos, acrescido

de outras informações para caracterizar melhor cada agrupamento, como quantidade de documentos em cada *cluster*, tipo de IES que tais documentos pertencem e alguns outros presentes na Tabela 5.1. No Apêndice C encontra-se o relatório geral desta análise. Na Tabela 5.1, apresenta-se uma síntese desse relatório geral, contendo somente os agrupamento que tiveram poucos documentos em sua formação.

Figura 5.2: Agrupamentos gerados pelo *k-Means* utilizando Unigramas em uma representação 2D.



Fonte: Elaborado pelo autor (2022).

É importante ressaltar que a matriz gerada pelo TF e pelo TF-IDF são esparsas, na qual as palavras são representadas em um espaço dimensional de comprimento igual ao vocabulário. Para a redução da dimensão, foi utilizada a *Principal Component Analysis* (PCA) [50], com duas componentes possibilitando a plotagem dos dados do *k-Means*.

Analisando a Tabela 5.1 e a separação de documentos gerados pelo algoritmo, foi possível notar que os *clusters* 0, 4 e 5 são formados por documentos da mesma instituição, ou seja, são PPCs de cursos oferecidos em distintos campus, pela mesma universidade. Contudo, apenas o *cluster* 0 possui os três documentos avaliados pelo Enade e com a mesma nota. Deste modo, é possível afirmar que tais documentos possuem uma estrutura muito semelhante, embora sejam dois de SI e um de CC. Outras características que diferem os PPCs desse *cluster* são algumas adequações do curso à região a

qual pertence.

Tabela 5.1: Síntese da Tabela Características dos *clusters* gerados pelo *k-Means*.

	Clusters							
	0	**4**	**5**	**6**	**7**	**9**	**10**	**11**
Quantidade de Termos	2871	2191	2863	1326	1080	544	1658	430
Documentos	3	2	3	1	1	1	1	1
Inst. Pública	3	2	3	-	1	1	1	1
Pública Federal	3	-	3	-	1	1	1	1
Instituto Federal	-	2	-	-	-	-	-	-
Inst. Privada	-	-	-	1	-	-	-	-
Curso de CC	1	1	-	1	1	-	-	1
Curso de SI	2	1	3	-	-	1	1	-
PPCs Região Norte	-	-	3	-	-	-	-	-
PPCs Região Nordeste	3	2	-	-	-	-	-	1
PPCs Região Sudeste	-	-	-	1	1	1	-	-
PPCs Região Sul	-	-	-	-	-	-	1	-
Conceito 3	-	1	1	-	-	-	-	-
Conceito 4	3	-	-	-	1	-	1	-
Conceito 5	-	-	-	1	-	1	-	1
Sem Conceito	-	1	2	-	-	-	-	-

Fonte: Elaborado pelo autor (2022).

Já nos *clusters* 4 e 5, também com documentos de uma mesma instituição em sua formação, foi possível notar que há documentos "Sem Conceito". Ao elaborar estes *clusters*, o algoritmo indicou a similaridades entre eles, mas devido ao adiantamento (Pandemia da COVID-19) da aplicação da prova para o ano seguinte. Para o *cluster* 4, notou-se que, mesmo sendo documentos de cursos diferente, há uma grande semelhança entre eles. E para o *cluster* 5 é ainda mais evidente que houve uma replicação do documento em campus ou unidades distintas. A média de unigramas para o *cluster* 4 é de ≈ 1.095 termos por documentos e para o *cluster* 5 é de ≈ 954 termos.

Os *clusters* 6, 7, 9, 10 e 11 contém apenas 1 PPC; além de possuírem uma quantidade reduzida de termos em sua formação, são documentos únicos. Os *clusters* 7 e 10 contém PPCs de cursos diferentes, porém os documentos são diferentes entre si. Por exemplo, o *cluster* 7 contém um PPC de CC, para a IES deste documento há a opção do curso de SI. Contudo, ele foi agrupado em outro *cluster*, indicando que houve uma preocupação em criar um PPC específico para cada curso. Destacando também, os ótimos

conceito Enade (Faixa) que esses documentos dos *clusters*, 5, 6, 7, 9, 10 e 11, possuem, tendo somente notas 4 e 5.

Deve-se levar em conta que espera-se alta similaridade entre os documentos, já que são elaborados seguindo uma diretriz obrigatória. Deste modo, é interessante verificar a Tabela completa no Apêndice C e notar como os *clusters* 1, 2, 3 e 8 possuem grandes quantidades de documentos em suas formações, demonstrando como esta similaridade afetou a elaboração dos agrupamentos, e como somente os *clusters* destacados na Tabela 5.1 chamaram atenção por não se agruparem juntamente aos demais.

5.2 AVALIAÇÃO DOS MODELOS DE CLASSIFICAÇÃO DOS PPC'S

Utilizou-se a MLP como algoritmo principal nos três classificadores da hierarquia de classificação. Foram realizadas cinco execuções, com três repetições cada. Conforme visto anteriormente, o classificador 1 é responsável por separar entre a classe 0 (CEF 2 e 3) e classe 1 (CEF 4 e 5). Portando, ao predizer a classe 0, o documento é repassando ao classificador 2, que tem o foco em separar entre a classe 0 (CEF 2) e a classe 1 (CEF 3). E ao predizer a classe 1, o dado é repassado ao classificador 3, que tem o intuito de classificar entre as classe 0 (CEF 4) e 1 (CEF 5). É possível afirmar que o classificador 1 é o modelo mais importante nesta hierarquia, pois é a partir de suas predições corretas os dados são encaminhados para o classificador seguinte responsável. Ainda, os resultados do classificador 1 são úteis na separação dos cursos mais qualificados (CEF 4 e 5) para os curso menos qualificados (CEF 2 e 3). No Apêndice D, são apresentadas as médias das acurácias para cada classificador, em suas respectivas repetições e execuções, para o treino/validação, gerando estimativas que variam de 60% à 82%.

Como mostrado na Tabela 5.2 a melhor situação se encontra na quinta execução, com acurácias iguais ou acima de 80%. Como salientando anteriormente, ao final de cada repetição, foi elaborada a matriz multiclasses, bem como calculadas as métricas de *Micro-Avereged Precision e Recall* sobre essa matriz. A relação destas métricas pode ser notada na Tabela 5.3.

Na Tabela 5.4 são apresentadas as médias gerais das métricas já mencionadas. Para a *Micro-Avereged Precision e Recall*, o valor de 0,6444 é obtido a partir da média aritmética de cada métrica, a partir dos dados de cada

execução (Tabela 5.3).

Tabela 5.2: Acurácias referente ao treino/teste da classificação.

1º Execução				
Modelos	**Acurácia**			
	Repetição 1	**Repetição 2**	**Repetição 3**	**Média**
Classificador 1 - MLP	86,66%	73,33%	82,22%	80,74%
Classificador 2 - MLP	72,22%	87,50%	68,42%	76,05%
Classificador 3 - MLP	90,47%	82,35%	72,22%	81,68%
2º Execução				
Modelos	**Acurácia**			
	Repetição 1	**Repetição 2**	**Repetição 3**	**Média**
Classificador 1 - MLP	80,00%	80,00%	80,00%	80,00%
Classificador 2 - MLP	66,66%	83,33%	83,33%	77,77%
Classificador 3 - MLP	83,33%	77,77%	83,33%	81,48%
3º Execução				
Modelos	**Acurácia**			
	Repetição 1	**Repetição 2**	**Repetição 3**	**Média**
Classificador 1 - MLP	80,00%	75,55%	80,00%	78,52%
Classificador 2 - MLP	77,77%	82,35%	76,47%	78,86%
Classificador 3 - MLP	83,33%	70,58%	84,21%	79,37%
4º Execução				
Modelos	**Acurácia**			
	Repetição 1	**Repetição 2**	**Repetição 3**	**Média**
Classificador 1 - MLP	77,77%	80,00%	80,00%	79,26%
Classificador 2 - MLP	88,23%	82,35%	83,33%	84,64%
Classificador 3 - MLP	83,33%	84,21%	83,33%	83,62%
5º Execução				
Modelos	**Acurácia**			
	Repetição 1	**Repetição 2**	**Repetição 3**	**Média**
Classificador 1 - MLP	80,00%	82,22%	80,00%	80,74%
Classificador 2 - MLP	82,35%	81,25%	81,25%	81,62%
Classificador 3 - MLP	84,21%	80,95%	80,00%	81,72%

Fonte: Elaborado pelo autor (2022).

E para finalizar, foi elaborada a matriz multiclasses final, como ilustrado na Figura 5.3, que unifica todas as matrizes de cada execução em apenas uma, possibilitando averiguar quais CEFs estavam se sobressaindo das demais durante a classificação. Nessa matriz, é possível observar que CEF 2 e 5 possuem os maiores acertos. A maioria dos erros relacionados à

CEF 4 refere-se à predição incorreta dos seus elementos como do CEF 2 ou 3, e poucos erros predizendo seus elementos como do CEF 5. O CEF 3 é a mais deficiente nos classificadores, tendo mais erros do que acertos. A matriz multiclasse final, apresenta uma acurácia de 64,44%.

Tabela 5.3: *Micro-Avereged Precision e Recall* referente ao treino/teste da classificação.

1º Execução				
Micro-Avereged	**Repetição 1**	**Repetição 2**	**Repetição 3**	**Média**
Precision	0,7111	0,6222	0,5777	0,6370
Recall	0,7111	0,6222	0,5777	0,6370
2º Execução				
Micro-Avereged	**Repetição 1**	**Repetição 2**	**Repetição 3**	**Média**
Precision	0,6000	0,6444	0,6666	0,6370
Recall	0,6000	0,6444	0,6666	0,6370
3º Execução				
Micro-Avereged	**Repetição 1**	**Repetição 2**	**Repetição 3**	**Média**
Precision	0,6444	0,5777	0,6444	0,6222
Recall	0,6444	0,5777	0,6444	0,6222
4º Execução				
Micro-Avereged	**Repetição 1**	**Repetição 2**	**Repetição 3**	**Média**
Precision	0,6666	0,6666	0,6666	0,6666
Recall	0,6666	0,6666	0,6666	0,6666
5º Execução				
Micro-Avereged	**Repetição 1**	**Repetição 2**	**Repetição 3**	**Média**
Precision	0,6666	0,6666	0,6444	0,6592
Recall	0,6666	0,6666	0,6444	0,6592

Fonte: Elaborado pelo autor (2022).

Para detalhar a distribuição de acertos e erros da matriz multiclasses, na Tabela 5.5 apresenta-se a síntese geral das características dos documentos utilizados na classificação. A partir dessa Tabela, é possível observar alguns traços interessantes sobres os documentos, separando-os pelos seus conceitos Enade Faixa, como médias de inscritos e participantes do Enade, média do Conceito Enade Contínuo, quantidade média de páginas por documento, entre outros, e a distribuição de cada PPCs (perante seu CEF) para cada região brasileira.

Analisando a primeira parte da Tabela 5.5, referente às médias, os valores para cada característica pontuada são muito próximos, com exceção da média do "Conceito Enade (Contínuo)", na qual, para cada classe, há um

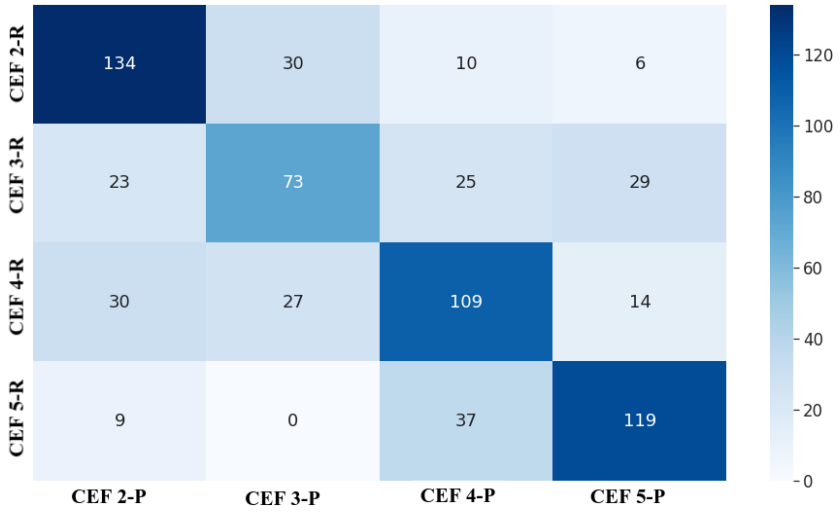
intervalo que delimita a mesma (os intervalos de cada classe foram apresentados na Seção 4.4.1).

Tabela 5.4: Médias Finais das métricas Acurácia, *Micro-Avereged Precision e Recall* referente ao treino/teste da classificação.

	Média Final
Micro-Avereged Precision	0,6444
Micro-Avereged Recall	0,6444
	Média Acurácia Geral
Classificador 1 - MLP	79,85%
Classificador 2 - MLP	79,79%
Classificador 3 - MLP	81,57%

Fonte: Elaborado pelo autor (2022).

Figura 5.3: Matriz Multiclasses Final, referente ao treino/teste da classificação.



Fonte: Elaborado pelo autor (2022).

Ao analisar as CEFs 2 e 5, verifica-se que ambas tiveram destaque na matriz multiclasses. É possível notar a similaridade entre elas, como na Quantidade de Termos presentes nos documentos originais ("Quant. Termos PPCs Ori."), nos documentos após o Corte de Luhn (CL) ("Quant. Termos PPCs após CL") e documentos após passarem pela limpeza e tratamento dos dados ("Quant. Termos PPCs Proc."). A linha "Quant. Termos PPCs Proc." é quantidade final de termos utilizado na Quinta Versão da classificação, e a diferença entre a CEF 2 e 5 gira em torno de 700 termos.

Tabela 5.5: Características dos dados utilizados na Classificação.

	CEF 2	**CEF 3**	**CEF 4**	**CEF 5**
		Médias		
Nº de Concluintes Inscritos	28	33	40	34
Nº de Concluintes Participantes	24	28	36	30
Conceito Enade (Contínuo)	1,6051	2,5083	3,4735	4,2664
Quantidade de Pág. dos PPCs	121,35	140,23	95,37	112,8
Ano do Arquivo	2015	2015	2015	2015
Quant. Termos PPCs Ori.	33.157,50	36.691,68	27.382,46	31.172,92
Quant. Termos PPCs após CL	24.370,44	29.367,74	21423,58	23.539,46
Quant. Termos PPCs Proc.	9.419,69	11.279,13	7.976,19	8.678,21
		Somatório		
Região - Norte	6	8	3	0
Região - Nordeste	4	9	13	6
Região - Centro-Oeste	8	9	9	2
Região - Sudeste	9	14	27	14
Região - Sul	5	11	11	6

Fonte: Elaborado pelo autor (2022).

As duas primeiras linhas da tabela indicam o número de concluintes inscritos e os que participaram, respectivamente, do Enade 2017. Note que o CEF 4 possui a maior quantidade de alunos para ambas situações e CEF 2 possuindo o menor índice entre elas, e a variação de alunos inscritos aos que participaram é de 4, exceto para o CEF 3, que houve variação de 5. A terceira linha indica a média do Conceito Enade Contínuo, que são os valores que formam o conceito final, o CEF 2 é a que possui um valor contínuo mais próximo da classe seguinte (os intervalos de cada conceito foram apresentados na Seção 4.4.1), que seria 1,95, as demais classes dispõem de valores medianos, nem muito alto e nem baixos.

As linhas quatro e cinco, dispõem das quantidade de página dos PPC e o ano do arquivo, ao analisar é possível destacar as medias próximas ou acima de 100 páginas por documento e note o ano da maioria dos documentos, com uma média de 2015, indicando uma defasagem em relação a atualizações de tais documentos. Já as linhas 6, 7 e 8 apresentam, respectivamente, a quantidade de termos originais (ou seja, termos presentes nos documentos originais, sem sofrerem nenhum tipo de tratamento), a quantidade de termos após o Corte de Luhn e a quantidade de termos processados, que é aqueles documentos que passaram pela limpeza de dados até a aplicação de *Stemming*. Ao

sofrer algumas etapas de processados dos dados, os termos foram diminuindo e estabilizando com médias próximas a 10.000 termos, mas quantidade não é qualidade. Note que os CEF 4 e 5, possuem menos termos, bem como quantidade de páginas, e dispõem de ótimas notas no Enade, indicando que os termos utilizados, possivelmente, estão focados em seguir as diretrizes e criar um perfil de estudante balanceado entre teoria e prática.

E as últimas linhas apresentam uma distribuição das classes por meio das regiões brasileiras. Em destaque as regiões sul, sudeste e nordeste possuem a maior quantidade de IES sendo do CEF 5, um aspecto para tal informação é que diversas dessas instituições são antigas, ou até as primeiras no Brasil, deste modo, dispõem de uma metodologia de ensino bem estruturada e uma reputação a zelar.

Ao analisar a segunda parte da Tabela 5.5, é possível notar como cada classe está distribuída pelas regiões brasileiras. A região Sudeste é a que mais possui documentos, independente da classe, bem como a maioria das classes estão nela. Analisando o CEF 2 é possível atentar como há um distribuição semelhante para cada região, com as região Centro-Oeste e Sudeste sendo as principais.

Na Tabela 5.6 é possível averiguar que grande maioria das IES são públicas (um total de 20) e apenas 12 são privadas, mas a região sudeste que possui ao todo 9 instituições, 6 delas são privadas e apenas 3 públicas.

Os resultados a respeito do CEF 5 destacam-se, pois das 28 IES apenas 1 é privada, as outras 27 são públicas. Além disso, destaca-se novamente a região sudeste, tendo 13 cursos avaliados com nova máxima. Note que para o CEF 5, as regiões dominantes são a Sudeste, Sul e Nordeste. Para os CEF 3 e 4, a maioria das instituições são públicas, 41 e 61 respectivamente, sendo para o CEF 3 apenas 10 privadas e apenas 2 para o CEF 4.

Tabela 5.6: Distribuição das classes de IES Privadas e Públicas por região.

	Sobre as IES Privadas			
	CEF 2	**CEF 3**	**CEF 4**	**CEF 5**
Quant. por CEF	12	10	2	1
Distribuição por Região				
Região - Norte	1	1	-	-
Região - Nordeste	2	1	-	-
Região - Centro-Oeste	1	2	-	-
Região - Sudeste	6	3	1	1
Região - Sul	2	3	1	-
	Sobre as IES Públicas			
	CEF 2	**CEF 3**	**CEF 4**	**CEF 5**
Quant. por CEF	20	41	61	27
Distribuição por Região				
Região - Norte	5	7	3	-
Região - Nordeste	2	8	13	6
Região - Centro-Oeste	7	7	9	2
Região - Sudeste	3	11	26	13
Região - Sul	3	8	10	6

Fonte: Elaborado pelo autor (2022).

5.3 AVALIAÇÃO DOS MODELOS DE PREDIÇÃO DO CONCEITO ENADE CONTÍNUO

A aplicação do modelo de árvore de decisão sobre o conjunto de unigramas gerou os seguintes resultados: Durante o treino/validação, com apenas unigramas, as estimativas foram:

- Classificador 1: acurácias médias de 62% à 67%.
- Regressor 2: MSE de 0,48 a 0,74 e MAPE de 32% a 42%.
- Regressor 3: MSE de 0,34 a 0,48 e MAPE de 12% a 15%.

No Apêndice E está disponível a versão detalhada com todas a métricas para cada modelo.

Em continuação, realizou-se os experimentos utilizando a MLP como algoritmo no classificador 1. Para o regressor 2, o modelo SVR e para o regressor 3 o modelo *k-NN Regressor*. Na Tabela 5.7 são apresentados os resultados finais, utilizando o conjunto contendo apenas unigramas.

Tabela 5.7: Médias finais referente ao treino/teste da regressão, contendo unigramas.

Classificador 1				
Modelo	**Repetição 1**	**Repetição 2**	**Repetição 3**	**Média Final**
		Acurácia		
MLP	79,41%	79,41%	73,52%	77,45%
Regressor 2				
Modelo	**Repetição 1**	**Repetição 2**	**Repetição 3**	**Média Final**
		Média Métrica MSE		
SVR	0,2965	0,3406	0,1508	0,2626
		Média Métrica MAPE		
SVR	28,08%	31,30%	14,75%	24,71%
Regressor 3				
Modelo	**Repetição 1**	**Repetição 2**	**Repetição 3**	**Média Final**
		Média Métrica MSE		
***k*-NNR**	0,3899	0,2264	0,1412	0,2525
		Média Métrica MAPE		
***k*-NNR**	15,95%	10,56%	8,44%	11,65%

Fonte: Elaborado pelo autor (2022).

Os resultados da execução da MLP no classificador 1 alcançou acurácia mínima de 73,52% e uma média de 77,45%, errando em 7 amostras das 34 destinadas ao teste. Porém, na Repetição 3, houve dificuldades do modelo em manter essa proporção, gerando 9 erros. O modelo SVR, no regressor 2, resultou MSE de, aproximadamente, 15% a 34%, com média de 26,26%; e MAPE variando de, aproximadamente, 14% a 31%, e média de 24,71%. É possível notar que, neste regressor, há uma certa dificuldade do modelo predizer corretamente, pois muitas de suas predições eram valores contínuos do CEF 3, que possuía mais amostras em relação ao CEF 2. Em continuação, no regressor 3, empregando o *k-NN Regressor*, os resultados foram similares ao regressor 2, com MSE de 14% a 38%, e média de 25,25%, e MAPE de 8% a 15%, com média de 11,65%. Desta forma, o modelo expôs a mesma deficiência que o anterior para a classe majoritária, que no caso era o CEF 4. Porém, com a segunda métrica (MAPE) mostrando que o modelo previu melhor os seus resultados em relação a classe real.

Vale ressaltar que o primeiro classificador, com acurácia média de 77,45%, sofre levemente pelo problema das classes majoritárias, dificultando a diferenciação dos elementos do CEF 3 (atributo 0) e CEF 4 (atributo 1). Analogamente, os regressores também sofrem influência do desbalanceamento, e também realizam predições com base nas classes majoritárias. Por exemplo, o segundo regressor prediz valores próximos a 2,00 (CEF 3) e terceiro regressor repetidamente prediz valores próximos de 3,00 (CEF 4). Portanto, conclui-se que há espaço para potenciais melhorias dos modelos e que uma alternativa promissora é a aplicação de DA para balanceamento dos conjuntos de dados.

5.4 CONSIDERAÇÕES FINAIS

O atual Capítulo apresentou resultados de clusterização dos PPCs na Seção 5.1. Em seguida, a Seção 5.2 apresentou os resultados dos métodos de classificação dos PPCs com relação aos CEFs, por meio da classificação em níveis. A aplicação dessa técnica gerou melhores resultados para o modelo MLP, o qual alcançou acurácia média de 80% e $\approx 0,65$ para a *Micro-Avereged Precision e Recall* sobre a matriz multiclasses final. Na Seção 5.3 foram apresentados os resultados de métodos de regressão para predizer o Conceito Enade do curso, tendo como base o seu PPC. Os resultados mostraram que o Classificador 1 conseguiu realizar a classificação com qualidade para as classe 0 e 1, possibilitando que os Regressores 2 e 3 recebessem amostras significativas para suas predições. Dentre eles, o Regressor 3 conseguiu predizer com mais precisão o CEC dos PPCs.

6 CONSIDERAÇÕES FINAIS

Compreender como um projeto pedagógico de curso afeta o aprendizado do estudante e, consequentemente, o conceito do curso, não é trivial e exige uma análise criteriosa sobre o documento e sobre os instrumentos de avaliação. Na tentativa de aumentar a abrangência da análise e de reduzir o tempo e esforço, pode-se utilizar técnicas de Inteligencia Artificial (IA) para automatizar parte do processo. Assim, essa dissertação de mestrado apresentou técnicas de mineração de dados e aprendizado de máquina sobre 223 projetos pedagógicos de curso de Ciência da Computação e Sistemas de Informação, com o objetivo de agrupá-los de acordo com similaridades e, posteriormente, classificá-los de acordo com conceitos (faixa e contínuo) do Enade.

Primeiramente, houve a necessidade de se elaborar uma base de dados de projetos pedagógicos, visto que não foram encontrados trabalhos anteriores com base de dados de acordo com as necessidades deste projeto. Os portais do MEC e Inep não oferecem a possibilidade de adquirir os PPCs dos cursos superiores em seus repositórios. Desta forma, formação do conjunto de dados foi realizada mediante o acesso aos portais online das IES em busca de tais documentos, sendo um processo demorado, minucioso e exploratório.

Com relação à Mineração de Dados, foram estudadas e implementadas as etapas de pré-processamento dos dados de entrada, por meio da limpeza e normalização dos dados, criação de matriz atributo-valor por via de técnicas TF e TF-IDF, entre outros; a extração de padrões, na qual se escolhe qual o tipo de algoritmo irá ser utilizado para se obter resultados, sendo descritivos ou preditivas; e o pós-processamento, última etapa que tem o objetivo de analisar os resultados e compreender o que foi obtido. Já com relação às técnicas de Aprendizado de Máquina, explicou-se a teoria relacionada aos modelos supervisionados, tanto para classificação, quanto para regressão.

Uma vez preparado o conjunto de treinamento composto por 223 projetos pedagógicos dos cursos de Ciência da Computação e Sistemas de Informação, foram realizados experimentos utilizando modelos de clusterização para criar agrupamentos dos projetos pedagógicos do conjunto de treinamento, com relação a suas similaridades. O objetivo desta etapa é direcionar os esforços em compreender o que cada agrupamento continha, qual a seria ligação entre eles. Foi implementado o modelo *k-Means*, o qual demandou um

alto esforço para avaliação de seus resultados. Foi possível encontrar métrica silhueta igual variando de 0,020 à 0,091, utilizando o conjunto completo ou apenas as unigramas. Notou-se que os *clusters* têm muita similaridade, levando assim, a agrupamentos próximos e dificultando a compreensão. Ainda assim, foi possível identificar o valor de $k = 12$ retornando um agrupamento interessante, aplicando apenas as unigramas, deste modo, realizou-se um levantamento dos principais termos de cada *cluster* para se obter mais entendimento sobre o conjunto de dados de entrada e sobre o agrupamento retornado. A análise dos agrupamentos permitiu identificar documentos destoantes dos demais, como os PPCs ainda sem conceito Enade e PPCs idênticos para uma mesma IES. Ainda, foi possível identificar grupos de PPCs com alta avaliação, reforçando a hipótese de que o conteúdo do PPC tem impacto direto na avaliação dos cursos.

Para a utilização dos métodos de classificação, houve a retiradas dos documentos que não possuíam Conceito Enade e também a aplicação de *Data Augmentation* sobre as classe 2 e 5, a fim de balanceá-las perante as outras duas. Deste modo, o intuito de se utilizar a Classificação era a possibilidade de se categorizar novos documentos mediante a um modelo treinado utilizado PPCs que já possuem Conceito Enade. Aplicando uma árvore de decisão, sendo que para cada classificador a MLP era o modelo fixo, foi possível se obter resultados que mostram as classes 2 e 5 tendo os maiores acertos, expondo os extremos das notas, pois tais classes indiciam IES abaixo da média e acima da média, respectivamente. Com uma acurácia média de $\approx 79\%$ nos classificadores 1 e 2, $\approx 81\%$ no classificador 3. A classificação proporcionou predizer o Conceito Enade Faixa com uma acurácia média, nos três classificadores, de $\approx 80\%$ e $\approx 64\%$ na matriz de confusão multiclasses final.

Além disso, também foram avaliados modelos de regressão com o objetivo de prever qual seria o Conceito Enade Contínuo de cada curso, tendo por base o seu projeto pedagógico. Inicialmente retirou-se os documentos que não possuíam Conceito Enade e que eram da classe 1, restando apenas 174 PPCs. Em seguida, ao utilizar a árvore de decisão, foram avaliados os modelos MLP para o classificador 1 e para os regressores 2 e 3, respectivamente, os modelos SVR e k-NNR, por meio das métricas de acurácia, para classificação, e MSE e MAPE, para regressão. O modelo MLP obteve acurácia média de $\approx 79\%$, já SVR (regressor 2) com MSE de 0,2626 e MAPE de $24,71\%$ e o modelo k-NNR (regressor 3) com MSE de 0,2525 e MAPE de $11,65\%$. Em resumo, a estrutura da regressão possibilita predizer o Conceito

Enade Contínuo com erro percentual absoluto médio de $\approx 11\%$, no melhor caso.

Um grande desafio neste projeto foi a busca por PPCs nas instituições de ensino, mesmo considerando a obrigatoriedade de disponibilização desse documento nos portais dos cursos. Mesmo quando divulgado, o acesso ao PPC, em muitos casos, não é intuitivo. Devido a isso, pode-se concluir que a Lei de Acesso a Informação [38] e o Artigo 32 da Portaria Normativa nº 40 [115] não estão sendo cumpridos por algumas instituições públicas e privadas. Outra situação interessante é que, muitas vezes, os projetos pedagógicos, embora disponíveis, estão desatualizados, datados anteriormente à 2016. Como o último Enade ocorreu em 2017, muitas instituições já poderiam ter versões de PPCs adequados para a melhoria dos seus índices de qualidade.

Em contrapartida, deve-se elogiar algumas intuições que, além de disponibilizarem a versão atual de seus PPCs, ainda informam que há novas versões em trâmites em órgãos colegiados para serem aprovadas. Em especial, os Institutos Federais apresentam essa preocupação em informar adequadamente a população, além de possuírem uma forma bem simples para que o usuário encontre os documentos em seus portais.

Uma sugestão ao MEC e ao Inep, seria a unificação de todas as informação dos cursos de graduação em apenas um portal, no qual o usuário poderia encontrar informações sobre o curso, a instituição, o projeto pedagógico do curso, os projetos de desenvolvimento institucionais (PDI, PPI), a Diretriz Curricular, os Indicadores de Qualidade da instituição e seu curso, entre outros. Pois em tempos atuais, para buscar tais informações, deve-se acessar diversos *websites*, tornando o processo descentralizado, cansativo e confuso.

Este projeto é o passo inicial para diversos outros, pois como mencionado na Seção 3, não há trabalhos específicos como este, utilizando técnicas de Mineração de Textos alinhadas com Aprendizado de Máquina sobre os PPCs, utilizando ainda os Conceitos Enade Faixa e Contínuo para classificação e regressão, respectivamente. Deste forma, há diversas possibilidades para trabalhos futuros, sendo alguma delas:

- **Aumento do conjunto de dados**: como apresentado na Seção 4.1, utilizou-se apenas um subconjunto do total disponível de PPC, dificultando diversas etapas deste projeto, portanto, a aquisição de novos

dados é de suma importância.

- **Projeto voltado à Aprendizagem Não Supervisionada**: um trabalho exclusivo sobre técnicas de agrupamentos ou outros modelos não supervisionados, a fim de aplicar análise com base em outros instrumentos e indicadores de qualidade.
- **Estudo sobre PPCs que estejam no limite entre faixas:** uma análise aprofundada sobre PPCs que estão em mudança de faixa, a fim de fornecer as características mais significativas para que o curso suba para faixa superior ou evitar que baixe para faixa inferior em uma próxima avaliação.
- **Projeto voltado à Classificação e/ou Regressão**: aplicação de novos conjuntos de dados, novas heurísticas, novos modelos de classificação e/ou regressão.

Toda a documentação, como os conjunto de dados utilizados, os códigos que implementaram todos os modelos apresentados neste projeto, gráficos, planilhas, desenhos dinâmicos, entre outras informações, estão disponíveis em `https://github.com/ChPro97/masters-project-education.git`

REFERÊNCIAS BIBLIOGRÁFICAS

[1] MEC. *Conselho Nacional de Educação – CNE*. Disponível em: http://portal.mec.gov.br/conselho-nacional-de-educacao/apresentacao. Acesso em 2 de dezembro de 2020.

[2] MEC. *CES0583*. Disponível em: http://portal.mec.gov.br/cne/arquivos/pdf/CES0583.pdf. Acesso em 2 de dezembro de 2020.

[3] MEC. *Diretrizes Curriculares - Cursos de Graduação*. Disponível em: http://portal.mec.gov.br/component/content/article?id=12991. Acesso em 2 de dezembro de 2020.

[4] INEP. *Exame Nacional de Desempenho dos Estudantes (Enade)*. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade. Acesso em 2 de dezembro de 2020.

[5] NOGUEIRA, B. M. "Hierarchical semi-supervised confidence-based active clustering and its application to the extraction of topic hierarchies from document collections". Tese de dout. Universidade de São Paulo, 2013.

[6] BRASIL. *Lei Nº 9.131, de 24 de novembro de 1995*. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9131.htm. Acesso em 21 de junho de 2021.

[7] BRASIL. *Legislação Informatizada - LEI Nº 9.394, de 20 de dezembro de 1996 - Publicação Original*. Disponível em: https://www2.camara.leg.br/legin/fed/lei/1996/lei-9394-20-dezembro-1996-362578-publicacaooriginal-1-pl.html. Acesso em 21 de junho de 2021.

[8] JUNIOR, P. R. T. "Diretrizes Curriculares Nacionais para o Ensino Superior: A Lógica das Competências em Foco". Em: *Crítica Educativa* 6.1 (dezembro de 2020), 1–18.

[9] FRAUCHES, C. C. *Diretrizes Curriculares para os Cursos de Graduação*. Brasília, DF: ABMES Editora, 2008. ISBN: 9788589597043.

[10] MEC. *Parecer CNE Nº 776/97*. Disponível em: http://portal.mec.gov.br/setec/arquivos/pdf/PCNE776_97.pdf. Acesso em 5 de julho de 2021.

[11] MEC. *Diretrizes Curriculares - Cursos de Graduação*. Disponível em: http://portal.mec.gov.br/escola-de-gestores-da-educacao-basica/323-secretarias-112877938/orgaos-vinculados-82187207/12991-diretrizes-curriculares-cursos-de-graduacao. Acesso em 5 de julho de 2021.

[12] INEP. *Instrumento Único de Avaliação de Cursos de Graduação*. Disponível em: http://inep.gov.br/documents/186968/484109/Instrumento+de+avalia%C3%A7%C3%A3o+de+cursos+de+gradua%C3%A7%C3%A3o/599968fa-b28e-4ce9-9bd8-4ef92fda88f7?version=1.2. Acesso em 21 de junho de 2021.

[13] UDESC. *Saiba mais sobre PDI, PPI e PPC e CURRÍCULO*. Disponível em: http://biblioteca.udesc.br/arquivos/id_submenu/75/texto_saiba_mais_sobre_pdi_ppi_ppc_e_curriculo.pdf. Acesso em 21 de junho de 2021.

[14] PROGRAD. *Projeto Pedagógico de Curso - PPC*. Rel. técn. Universidade Federal de Minas Gerais, 2007.

[15] INEP. *Instrumento de avaliação de Cursos 2017*. Disponível em: https://download.inep.gov.br/educacao_superior/avaliacao_cursos_graduacao/instrumentos/2017/curso_reconhecimento.pdf. Acesso em 5 de julho de 2021.

[16] UFMS. *Orientações para a construção do Projeto Pedagógico de Curso*. Disponível em: http://preg.sites.ufms.br/files/2015/09/Guia-para-elabora%C3%A7%C3%A3o-de-PPC-Primeira-vers%C3%A3o-publicada.pdf. Acesso em 5 de julho de 2021.

[17] BRASIL. *Lei Nº 10.861, de 14 de abril de 2004*. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.861.htm. Acesso em 21 de junho de 2021.

[18] MEC. *Conheça a CONAES*. Disponível em: http://portal.mec.gov.br/conaes-comissao-nacional-de-avaliacao-da-educacao-superior/conheca-a-conaes#:~:text=A%20Comiss%C3%A3o%20Nacional%20de%20Avalia%C3%A7%C3%A3o,14%20de%20Abril%20de%202004. Acesso em 21 de junho de 2021.

[19] INEP. *SINAES*. Disponível em: http://inep.gov.br/sinaes. Acesso em 21 de junho de 2021.

[20] INEP. *Resultados*. Disponível em: http://portal.inep.gov.br/educacao-superior/indicadores-de-qualidade/resultados. Acesso em 21 de junho de 2021.

[21] SEIXAS, P. S. et al. “Projeto Pedagógico de Curso e formação do psicólogo: uma proposta de análise”. Em: *Psicologia Escolar e Educacional* 17 (2013), 113–122.

[22] INEP. *Indicadores de Qualidade da Educação Superior*. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior. Acesso em 6 de julho de 2021.

[23] GRIBOSKI, C. M. “O Enade como Indutor da Qualidade da Educação Superior”. Em: *Estudos em Avaliação Educacional* 23.53 (2012), pp. 178–195.

[24] INEP. *Nota Técnica Nº 5/2020/CGCQES/DAES*. Disponível em: https://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2019/NOTA_TECNICA_N_5-2020_CGCQES-DAES_Metodologia_de_calculo_do_Conceito_Enade_2019.pdf. Acesso em 16 de maio de 2022.

[25] INEP. *Conceito Enade*. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior/conceito-enade. Acesso em 6 de julho de 2021.

[26] INEP. *Exame Nacional de Desempenho dos Estudantes (Enade)*. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade. Acesso em 6 de julho de 2021.

[27] INEP. *Nota Técnica CGCQES/DAES n.º 34/2020 – Descrição da metodologia do IDD 2019*. Disponível em: http://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2019/NOTA_TECNICA_N_34-2020_CGCQES-DAES_Metodologia_de_calculo_do_IDD_2019.pdf. Acesso em 6 de julho de 2021.

[28] INEP. *Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD)*. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior/indicador-de-diferenca-entre-os-desempenhos-observado-e-esperado-idd. Acesso em 6 de julho de 2021.

[29] INEP. *Nota Técnica CGCQES/DAES n.º 58/2020 – Descrição da metodologia do CPC 2019*. Disponível em: http://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2019/NOTA_TECNICA_N_58-2020_CGCQES-DAES_Metodologia_de_calculo_do_CPC_2019.pdf. Acesso em 6 de julho de 2021.

[30] INEP. *Conceito Preliminar de Curso (CPC)*. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior/conceito-preliminar-de-curso-cpc. Acesso em 6 de julho de 2021.

[31] INEP. *Índice Geral de Cursos (IGC)*. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior/indice-geral-de-cursos-igc. Acesso em 6 de julho de 2021.

[32] INEP. *Nota Técnica CGCQES/DAES n.º 59/2020 – Descrição da metodologia do IGC 2019*. Disponível em: http://download.inep.gov.br/educacao_superior/enade/notas_tecnicas/2019/NOTA_TECNICA_N_59-2020_CGCQES-DAES_Metodologia_de_calculo_do_IGC_2019.pdf. Acesso em 6 de julho de 2021.

[33] UFMS/DIAVI. *Conceito Institucional - CI*. Disponível em: https://diavi.ufms.br/ci/. Acesso em 7 de julho de 2021.

[34] C. S. SILVA. "Comunicação nas Plataformas Digitais : Um Estudo Sobre Universidades Brasileiras com Conceito Institucional Cinco". Dissertação (Mestrado). Porto Alegre: Pontifícia Universidade Católica do Rio Grande do Sul, 2017.

[35] INEP. *Censo da Educação Superior*. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior. Acesso em 5 de julho de 2021.

[36] INEP. *Tabelas de Divulgação*. Disponível em: https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2020/Tabelas_de_divulgacao_Censo_da_Educacao_Superior_2019.xls. Acesso em 7 de julho de 2021.

[37] MEC. *Cadastro Nacional de Cursos e Instituições de Educação Superior Cadastro e-MEC*. Disponível em: https://emec.mec.gov.br/. Acesso em 16 de maio de 2022.

[38] BRASIL. *Lei Nº 12.527, de 18 de novembro de 2011*. Disponível em: `http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm`. Acesso em 8 de fevereiro de 2021.

[39] IBM. "Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations". Em: *IBM Marketing Cloud* (2017), pp. 1–18.

[40] HOTHO, A.; NÜRNBERGER, A.; PAAß, G. "A Brief Survey of Text Mining". Em: *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology* 20.1 (maio de 2005), pp. 19–62. ISSN: 0175-1336.

[41] MORAIS, E. A. M.; AMBRÓSIO, A. P. L. "Mineração de Textos". Em: *Relatório Técnico–Instituto de Informática (UFG)* (dezembro de 2007), pp. 1–30.

[42] REZENDE, Solange O. *Sistemas Inteligentes: Fundamentos e Aplicações*. Barueri, SP: Editora Manole Ltda, 2003. ISBN: 8520416837.

[43] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. "Knowledge Discovery and Data Mining: Towards a Unifying Framework." Em: *KDD'96*. Vol. 96. Portland, Oregon: AAAI Press, 1996, pp. 82–88.

[44] MOURA, Maria F. "Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico". Tese de dout. Universidade de São Paulo, 2009.

[45] MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[46] BEKKERMAN, R.; ALLAN, J. *Using Bigrams in Text Categorization*. Rel. técn. Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst, 2004.

[47] QAISER, S.; ALI, R. "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents". Em: *International Journal of Computer Applications* 181.1 (julho de 2018), pp. 25–29. ISSN: 0975-8887.

[48] JOACHIMS, T. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. Rel. técn. School of Computer Science Carnegie Mellon University Science, Pittsburgh, PA, 1996.

[49] ROBERTSON, S. "Understanding inverse document frequency: on theoretical arguments for IDF". Em: *Journal of documentation* 60.5 (2004), pp. 503–520.

[50] JOLLIFFE, I.T. *Principal Component Analysis*. 2ª ed. Springer-Verlag New York, 2002. ISBN: 9780387954424.

[51] LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. "An introduction to Latent Semantic Analysis". Em: *Discourse Processes* 25.2-3 (1998), pp. 259–284.

[52] DEERWESTER, S. et al. "Indexing by latent semantic analysis". Em: *Journal of the American Society for Information Science* 41.6 (1990), pp. 391–407.

[53] SLONIM, N.; TISHBY, N. "Document Clustering Using Word Clusters via the Information Bottleneck Method". Em: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '00. New York, NY, USA: Association for Computing Machinery, 2000, 208–215. ISBN: 1581132263.

[54] CARVALHO, A.C.P.L.F. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. Rio de Janeiro: Grupo GEN-LTC, 2011. ISBN: 9788521618805.

[55] RUSSEL, S. J.; NORVIG, P. *Inteligência Artificial*. 3ª ed. Rio de Janeiro: Elsevier, 2013. ISBN: 9788535237016.

[56] LUGER, G. F. *Inteligência Artificial*. 6ª ed. Pearson Education do Brasil, 2013.

[57] COVER, T.; HART, P. "Nearest Neighbor Pattern Classification". Em: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27.

[58] CRISTIANINI, N.; SHAWE-TAYLOR, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[59] ZHU, X. "Semi-Supervised Learning Tutorial". Em: *International Conference on Machine Learning (ICML)*. 2007, pp. 1–135.

[60] MACQUEEN, J. "Some Methods for Classification and Analysis of Multivariate Observations". Em: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967, pp. 281–297.

[61] FLACH, P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.

[62] CHENG, Y. “Mean shift, mode seeking, and clustering”. Em: *IEEE transactions on pattern analysis and machine intelligence* 17.8 (1995), pp. 790–799.

[63] SILVER, D. et al. “Mastering the game of Go with deep neural networks and tree search”. Em: *Nature* 529.7587 (2016), pp. 484–489.

[64] DAYAN, P.; NIV, Y. “Reinforcement Learning: The Good, The Bad and The Ugly”. Em: *Current Opinion in Neurobiology* 18.2 (2008). Cognitive neuroscience, pp. 185 –196. ISSN: 0959-4388.

[65] MNIH, V. et al. “Asynchronous Methods for Deep Reinforcement Learning”. Em: *International conference on machine learning*. PMLR. 2016, pp. 1928–1937.

[66] MNIH, V. et al. “Playing atari with deep reinforcement learning”. Em: *arXiv e-prints* (2013).

[67] LILLICRAP, T. P. et al. “Continuous control with deep reinforcement learning”. Em: *arXiv e-prints* (julho de 2016).

[68] NAGABANDI, A. et al. “Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning”. Em: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 7559–7566.

[69] SILVER, D. et al. “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm”. Em: *ArXiv* abs/1712.01815 (2017).

[70] BUZA, K.; NANOPOULOS, A.; NAGY, G. “Nearest Neighbor Regression in the Presence of Bad Hubs”. Em: *Knowledge-Based Systems* 86 (2015), pp. 250–260. ISSN: 0950-7051.

[71] BISHOP, C. M. *Neural Networks for Pattern Recognition*. USA: Oxford University Press, Inc., 1995. ISBN: 0198538642.

[72] TAUD, H.; MAS, J. F. “Multilayer Perceptron (MLP)”. Em: *Geomatic Approaches for Modeling Land Change Scenarios*. Cham: Springer International Publishing, 2018, pp. 451–455. ISBN: 978-3-319-60801-3.

[73] ROGERS, S.; GIROLAMI, M. *A First Course in Machine Learning*. 2ª ed. Chapman e Hall/CRC, 2016. ISBN: 9781498738484.

[74] KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc, 1990.

[75] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2ª ed. Springer Series in Statistics. Springer-Verlag New York, 2009. ISBN: 9780387848846.

[76] TAN, P. et al. *Introduction to Data Mining*. 2nd. New York, NY: Pearson, 2018. ISBN: 0133128903.

[77] KOHAVI, R. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". Em: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, 1137–1143. ISBN: 1558603638.

[78] MITCHELL, T. M. *Machine Learning*. Vol. 1. New York: McGraw-hill New York, 1997.

[79] SUN, A.; LIM, E. "Hierarchical Text Classification and Evaluation". Em: *Proceedings 2001 IEEE International Conference on Data Mining*. 2001, pp. 521–528.

[80] GORDON, A. D. "A Review of Hierarchical Classification". Em: *Journal of the Royal Statistical Society: Series A (General)* 150.2 (1987), pp. 119–137.

[81] SILLA, C. N.; FREITAS, A. A. "A survey of Hierarchical Classification Across Different Application Domains". Em: *Data Mining and Knowledge Discovery* 22.1 (2011), pp. 31–72.

[82] SHORTEN, C.; KHOSHGOFTAAR, T. M. "A Survey On Image Data Augmentation for Deep Learning". Em: *Journal of big data* 6.1 (2019), pp. 1–48.

[83] VAN DYK, D. A.; MENG, X. "The Art of Data Augmentation". Em: *Journal of Computational and Graphical Statistics* 10.1 (2001), pp. 1–50.

[84] ZHANG, X.; ZHAO, J.; LECUN, Y. "Character-level Convolutional Networks for Text Classification". Em: *arXiv e-prints* (abril de 2016).

[85] SUGIYAMA, A.; YOSHINAGA, N. "Data augmentation using back-translation for context-aware neural machine translation". Em: *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*. Hong Kong, China: Association for Computational Linguistics, novembro de 2019, pp. 35–44.

[86] HAYASHI, T. et al. “Back-Translation-Style Data Augmentation for end-to-end ASR”. Em: *2018 IEEE Spoken Language Technology Workshop (SLT)*. 2018, pp. 426–433.

[87] WANG, W. Y.; YANG, D. “That’s So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets”. Em: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, setembro de 2015, pp. 2557–2563.

[88] FADAEE, M.; BISAZZA, A.; MONZ, C. “Data Augmentation for Low-Resource Neural Machine Translation”. Em: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2017.

[89] KOBAYASHI, S. “Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations”. Em: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, junho de 2018, pp. 452–457.

[90] KAFLE, K.; YOUSEFHUSSIEN, M.; KANAN, C. “Data Augmentation for Visual Question Answering”. Em: *Proceedings of the 10th International Conference on Natural Language Generation*. Santiago de Compostela, Spain: Association for Computational Linguistics, setembro de 2017, pp. 198–202.

[91] LUHN, H. P. “The Automatic Creation of Literature Abstracts”. Em: *IBM Journal of Research and Development* 2.2 (1958), pp. 159–165.

[92] ZIPF, G. K. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[93] HUTTER, F.; KOTTOFF, L.; VANSCHOREN, J. *Automated Machine Learning: Methods, Systems, Challenges*. 1st. The Springer Series on Challenges in Machine Learning. Springer Nature, 2019. ISBN: 3-030-05318-0.

[94] LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. “Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS”. Em: *arXiv e-prints* (dezembro de 2019).

[95] WANG, Z.; BOVIK, A. C. "Mean Squared Error: Love it or Leave it? A New Look at Signal Fidelity Measures". Em: *IEEE Signal Processing Magazine* 26.1 (2009), pp. 98–117.

[96] WILLMOTT, C. J.; MATSUURA, K. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". Em: *Climate Research* 30.1 (2005), pp. 79–82.

[97] CHAI, T.; DRAXLER, R. R. "Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature". Em: *Geoscientific Model Development* 7.3 (2014), pp. 1247–1250.

[98] BOWERMAN, B. L.; O'CONNELL, R. T.; KOEHLER, A. B. *Forecasting, time series, and regression: an applied approach*. Vol. 4. South-Western Pub, 2005.

[99] SAMMUT, C.; WEBB, G. I. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[100] HART, P. E.; STORK, D. G.; DUDA, R. O. *Pattern Classification*. Wiley Hoboken, 2000.

[101] STEHMAN, S. V. "Selecting and interpreting measures of thematic classification accuracy". Em: *Remote Sensing of Environment* 62.1 (1997), pp. 77–89. ISSN: 0034-4257.

[102] BATARSEH, F. A.; YANG, R. *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*. Academic Press, 2020.

[103] BISHOP, C. M.; NASRABADI, N. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[104] BRITO, M. R. F. "O SINAES e o ENADE: da Concepção a Implantação". pt. Em: *Avaliação: Revista da Avaliação da Educação Superior (Campinas)* 13.3 (novembro de 2008), pp. 841–850. ISSN: 1414-4077.

[105] CANAN, S. R.; ELOY, V. T. "Políticas de Avaliação em Larga Escala: o ENADE Interfere na Gestão dos Cursos?" Em: *Práxis Educativa* 11.3 (2016), pp. 621–640.

[106] LEMOS, K. C. S.; MIRANDA, G. J. "Alto e Baixo Desempenho no ENADE: que Variáveis Explicam?" Em: *Revista Ambiente Contábil - Universidade Federal do Rio Grande do Norte - ISSN 2176-9036* 7.2 (junho de 2015), pp. 101–118.

[107] NETO, A. N. A. "Avaliação da estrutura do currículo do ensino superior com aprendizado de máquina". Dissertação (Mestrado). Fortaleza: Universidade Federal do Ceará, 2018.

[108] SILVA, J. J. "Uma Comparação de Técnicas de Aprendizado de Máquina para Predição de Evasão de Estudantes no Ensino Público Superior". Dissertação (Mestrado em Sistemas de Informação). São Paulo: Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2022.

[109] PRIYAMBADA, S. A.; MAHENDRAWATHI, E. R.; YAHYA, B. N. "Curriculum Assessment of Higher Educational Institution Using Aggregate Profile Clustering". Em: *Procedia Computer Science* 124 (2017). 4th Information Systems International Conference 2017, ISICO 2017, 6-8 November 2017, Bali, Indonesia, pp. 264–273. ISSN: 1877-0509.

[110] VAN ROSSUM, G.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[111] PEDREGOSA, F. et al. "Scikit-learn: Machine Learning in Python". Em: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[112] The pandas development team. *pandas-dev/pandas: Pandas*. Versão latest. Fevereiro de 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[113] HUNTER, J. D. "Matplotlib: A 2D graphics environment". Em: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95.

[114] BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing With Python: Analyzing Text With The Natural Language Toolkit*. "O'Reilly Media, Inc.", 2009.

[115] Inep. *Portaria Normativa nº 40, de 12 de dezembro de 2007*. Disponível em: https://download.inep.gov.br/educacao_superior/censo_superior/legislacao/2007/portaria_40_12122007.pdf. Acesso em 22 de agosto de 2022.

[116] TKACZYK, D. et al. "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature". Em: *International Journal on Document Analysis and Recognition (IJDAR)* 18.4 (2015), pp. 317–335. ISSN: 1433-2833.

[117] TELLINGHUISEN, J.; BOLSTER, C. H. "Using R2 to compare least-squares fit models: When it must fail". Em: *Chemometrics and Intelligent Laboratory Systems* 105.2 (2011), pp. 220–222. ISSN: 0169-7439.

APÊNDICES

APÊNDICE A - MODELOS E SEUS PARÂMETROS DE CLASSIFICAÇÃO

Tabela A.1: Parâmetros da Classificação com Unigramas, execuções 1 e 2.

1º Execução - Parâmetros						
Modelo	**activation**	**alpha**	**early_stopping**	**hidden_layer_sizes**	**learning_rate**	**solver**
			1º Repetição			
Classificador 1 - MLP	relu	0.001	True	(200, 200)	constant	adam
Classificador 2 - MLP	relu	0.0001	True	(100,)	constant	adam
Classificador 3 - MLP	relu	0.0001	True	(25,)	constant	adam
			2º Repetição			
Classificador 1 - MLP	relu	0.0001	True	(100, 200, 100)	constant	adam
Classificador 2 - MLP	relu	0.01	True	(50,)	adaptive	adam
Classificador 3 - MLP	relu	0.0001	True	(10, 10)	adaptive	lbfgs
			3º Repetição			
Classificador 1 - MLP	relu	0.001	True	(200, 200)	constant	lbfgs
Classificador 2 - MLP	logistic	0.001	True	(200, 200)	adaptive	adam
Classificador 3 - MLP	logistic	0.0001	True	(50,)	adaptive	adam
2º Execução - Parâmetros						
Modelo	**activation**	**alpha**	**early_stopping**	**hidden_layer_sizes**	**learning_rate**	**solver**
			1º Repetição			
Classificador 1 - MLP	relu	0.001	True	(200, 200)	constant	adam
Classificador 2 - MLP	relu	0.0001	True	(200, 200)	adaptive	lbfgs
Classificador 3 - MLP	relu	0.0001	True	(25,)	constant	adam
			2º Repetição			
Classificador 1 - MLP	relu	0.1	True	(10, 10)	constant	adam
Classificador 2 - MLP	relu	0.0001	True	(200, 200)	adaptive	adam
Classificador 3 - MLP	logistic	0.001	True	(100,)	constant	lbfgs
			3º Repetição			
Classificador 1 - MLP	relu	0.01	True	(100, 10)	adaptive	adam
Classificador 2 - MLP	relu	0.0001	True	(200, 200)	adaptive	lbfgs
Classificador 3 - MLP	relu	0.001	True	(200, 10)	adaptive	adam

Fonte: Elaborado pelo autor (2022).

Tabela A.2: Parâmetros da Classificação com Unigramas, execuções 3 e 4.

3º Execução - Parâmetros						
Modelo	**activation**	**alpha**	**early_stopping**	**hidden_layer_sizes**	**learning_rate**	**solver**
			1º Repetição			
Classificador 1 - MLP	relu	0.0001	True	(200, 200)	constant	adam
Classificador 2 - MLP	relu	0.001	True	(100, 100)	adaptive	adam
Classificador 3 - MLP	relu	0.001	True	(100, 100)	adaptive	adam
			2º Repetição			
Classificador 1 - MLP	relu	0.0001	True	(200,)	adaptive	adam
Classificador 2 - MLP	relu	0.01	True	(100,)	adaptive	adam
Classificador 3 - MLP	relu	0.0001	True	(50, 50)	constant	lbfgs
			3º Repetição			
Classificador 1 - MLP	relu	0.01	True	(25,)	constant	adam
Classificador 2 - MLP	relu	0.01	True	(200, 200)	adaptive	adam
Classificador 3 - MLP	logistic	0.0001	True	(25, 25)	adaptive	lbfgs
4º Execução - Parâmetros						
Modelo	**activation**	**alpha**	**early_stopping**	**hidden_layer_sizes**	**learning_rate**	**solver**
			1º Repetição			
Classificador 1 - MLP	relu	0.0001	True	(100, 100)	constant	adam
Classificador 2 - MLP	relu	0.0001	True	(100,)	adaptive	adam
Classificador 3 - MLP	relu	0.1	True	(100, 100)	adaptive	adam
			2º Repetição			
Classificador 1 - MLP	relu	0.01	True	(10, 10)	adaptive	lbfgs
Classificador 2 - MLP	relu	0.001	True	(200, 200)	adaptive	lbfgs
Classificador 3 - MLP	relu	0.001	True	(200, 50, 30)	adaptive	lbfgs
			3º Repetição			
Classificador 1 - MLP	relu	0.01	True	(200, 200)	adaptive	adam
Classificador 2 - MLP	relu	0.0001	True	(200, 200)	adaptive	lbfgs
Classificador 3 - MLP	relu	0.1	True	(100, 100)	adaptive	adam

Fonte: Elaborado pelo autor (2022).

Tabela A.3: Parâmetros da Classificação com Unigramas, execução 5.

5º Execução - Parâmetros						
Modelo	**activation**	**alpha**	**early_stopping**	**hidden_layer_sizes**	**learning_rate**	**solver**
			1º Repetição			
Classificador 1 - MLP	relu	0.001	True	(200,)	constant	lbfgs
Classificador 2 - MLP	relu	0.001	True	(100,)	adaptive	adam
Classificador 3 - MLP	relu	0.001	True	(100,)	adaptive	adam
			2º Repetição			
Classificador 1 - MLP	relu	0.0001	True	(100, 100)	constant	adam
Classificador 2 - MLP	relu	0.01	True	(50, 50)	adaptive	adam
Classificador 3 - MLP	relu	0.01	True	(50, 50)	adaptive	adam
			3º Repetição			
Classificador 1 - MLP	relu	0.001	True	(150, 30)	adaptive	adam
Classificador 2 - MLP	relu	0.001	True	(200,)	adaptive	adam
Classificador 3 - MLP	relu	0.001	True	(50, 50)	adaptive	adam

Fonte: Elaborado pelo autor (2022).

APÊNDICE B - HIPERPARÂMETROS GRIDSEARCH E MELHORES PARÂMETROS DA REGRESSÃO

APÊNDICE B.1 - Parâmetros Candidatos dos Modelos na Execução do *GridSearch*

Tabela B.1: Modelos e seus Hiperparâmetros na execução de *GridSearchCV*, juntamente com a variação da *k-Fold Cross-Validation*, para a Regressão.

	Parâmetro	Função	Valores
Modelo SVR	`Kernel`	Função usada para ajudar a resolver problemas	'linear', 'poly', 'rbf', 'sigmoid'
	`C`	Parâmetro de regularização	0.1, 1, 10, 100, 1000
	`Degree`	Grau do Kernel 'poly'	0, 1, 2, 3, 4, 5, 6
	`Coef0`	Termo independente na função Kernel	0.01, 10, 0.5
	`Gamma`	Coeficiente de Kernel	'alto', 'scale'
	Parâmetro	**Função**	**Valores**
Modelo *k-NN Regressor*	`n_neighbors`	Número de vizinhos a serem usados	range(1,21)
	`weights`	Função de peso usada na previsão	'uniform', 'distance'
	`algorithm`	Algoritmo usado para calcular os vizinhos mais próximos	'auto', 'ball_tree', 'kd_tree', 'brute'
	`metric`	A métrica de distância	'euclidean','chebyshev','minkowski'
	Parâmetro	**Função**	**Valores**
Modelo MLPRegressor	`hidden_layer_sizes`	Quantidade de camadas ocultas	(10,), (25,), (50,), (100,), (200,), (10,10), (25,25), (50,50), (100,-100), (200,200)
	`Activation`	Função de ativação para a camada oculta	'relu', 'logistic'
	`Solver`	Otimização de peso	'sgd', 'adam', 'lbfgs'
	`Alpha`	Regularização de peso	0.1, 0.01, 0.001
	`learning_rate`	Taxa de aprendizado para atualizações de peso	'constant','adaptative'
	`max_iter`	Número máximo de iterações	2000
k-Fold Cross-Validation regressor 2 e 3		Amostragem/Validação dos dados	k=2

Fonte: Elaborado pelo autor (2022).

APÊNDICE B.2 - Modelos e seus Melhores Parâmetros Regressão

Tabela B.2: Parâmetros da Regressão utilizando Unigramas, repetições 1 e 2.

Modelo	**1º Repetição - Parâmetros**					
	activation	**alpha**	**early_stopping**	**hidden_layer_sizes**	**learning_rate**	**solver**
Classificador 1 - MLP	relu	0.001	True	(100, 100)	adaptive	adam
	C	**coef0**	**degree**	**gamma**	**kernel**	**-**
Regressor 2 - SVR	5	0.01	3	auto	linear	-
Regressor 3 - SVR	1	0.01	3	scale	poly	-
	n_neighbors	**algorithm**	**metric**	**weights**	**-**	**-**
Regressor 2 - k-NNR	15	auto	chebyshev	uniform	-	-
Regressor 3 - k-NNR	6	ball_tree	chebyshev	uniform	-	-
	activation	**alpha**	**max_iter**	**hidden_layer_sizes**	**learning_rate**	**solver**
Regressor 2 - MLPR	logistic	0.01	2000	(50, 100, 50)	adaptive	adam
Regressor 3 - MLPR	logistic	0.1	2000	(50, 100, 50)	adaptive	adam
Modelo	**2º Repetição - Parâmetros**					
	activation	**alpha**	**early_stopping**	**hidden_layer_sizes**	**learning_rate**	**solver**
Classificador 1 - MLP	relu	0.0001	True	(10, 10)	adaptive	lbfgs
	C	**coef0**	**degree**	**gamma**	**kernel**	**-**
Regressor 2 - SVR	1	0.01	8	scale	poly	-
Regressor 3 - SVR	1	0.01	3	scale	poly	-
	n_neighbors	**algorithm**	**metric**	**weights**	**-**	**-**
Regressor 2 - k-NNR	12	auto	chebyshev	uniform	-	-
Regressor 3 - k-NNR	14	ball_tree	chebyshev	uniform	-	-
	activation	**alpha**	**max_iter**	**hidden_layer_sizes**	**learning_rate**	**solver**
Regressor 2 - MLPR	logistic	0.001	2000	(100,)	adaptive	adam
Regressor 3 - MLPR	logistic	0.001	2000	(50, 100, 50)	adaptive	adam

Fonte: Elaborado pelo autor (2022).

Tabela B.3: Parâmetros da Regressão utilizando Unigramas, repetição 3.

Modelo	3º Repetição - Parâmetros					
	activation	**alpha**	**early_stopping**	**hidden_layer_sizes**	**learning_rate**	**solver**
Classificador 1 - MLP	relu	0.0001	True	(10, 10)	adaptive	lbfgs
	C	**coef0**	**degree**	**gamma**	**kernel**	**-**
Regressor 2 - SVR	5	0.01	8	scale	poly	-
Regressor 3 - SVR	1	0.01	3	scale	poly	-
	n_neighbors	**algorithm**	**metric**	**weights**	**-**	**-**
Regressor 2 - k-NNR	10	auto	chebyshev	uniform	-	-
Regressor 3 - k-NNR	2	auto	chebyshev	uniform	-	-
	activation	**alpha**	**max_iter**	**hidden_layer_sizes**	**learning_rate**	**solver**
Regressor 2 - MLPR	logistic	0.01	2000	(50, 100, 50)	adaptive	adam
Regressor 3 - MLPR	logistic	0.01	2000	(50, 100, 50)	constant	adam

Fonte: Elaborado pelo autor (2022).

APÊNDICE C - CARACTERÍSTICAS DOS AGRUPAMENTOS GERADAS PELO K-MEANS

Tabela C.1: Características dos *clusters* gerados pelo *k-Means* utilizando Unigramas.

	Clusters											
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Termos	2871	5346	5386	4522	2191	2863	1326	1080	5326	544	1658	430
Documentos	3	63	92	11	2	3	1	1	44	1	1	1
Inst. Pública	3	62	83	10	2	3	-	1	27	1	1	1
Pública Federal	3	40	29	7	-	3	-	1	15	1	1	1
Pública Estadual	-	16	16	-	-	-	-	-	3	-	-	-
Pública Municipal	-	-	3	-	-	-	-	-	-	-	-	-
Instituto Federal	-	6	35	3	2	-	-	-	7	-	-	-
CEFET	-	-	-	-	-	-	-	-	2	-	-	-
Inst. Privada	-	1	9	1	-	-	1	-	17	-	-	-
Curso de C.C.	1	52	33	6	1	-	1	1	20	-	-	1
Curso de S.I.	2	11	59	5	1	3	-	-	24	1	1	-
Docs Região Norte	-	3	8	-	-	3	-	-	10	-	-	-
Docs Região Nordeste	3	18	10	2	2	-	-	-	7	-	-	1
Docs Região Centro-Oeste	-	15	13	6	-	-	1	-	4	-	-	-
Docs Região Sudeste	-	21	35	1	-	-	-	1	18	1	-	-
Docs Região Sul	-	6	26	2	-	-	-	-	5	-	1	-
Conceito 1	-	-	1	-	-	-	-	-	1	-	-	-
Conceito 2	-	8	13	-	-	-	-	-	11	-	-	-
Conceito 3	-	13	19	3	1	1	-	-	14	-	-	-
Conceito 4	3	22	28	2	-	-	-	1	6	-	1	-
Conceito 5	-	12	8	3	-	-	1	-	2	1	-	1
Sem Conceito	-	8	23	3	1	2	-	-	10	-	-	-

Fonte: Elaborado pelo autor (2022).

APÊNDICE D - ESTIMATIVAS GERADAS PELOS MODELOS DURANTE TREINO/VALIDAÇÃO NA CLASSIFICAÇÃO

Tabela D.1: Estimativas médias do modelo MLP em cada execução.

1º Execução			
Modelos	**Acurácia**		
	Repetição 1	**Repetição 2**	**Repetição 3**
Classificador 1 - MLP	69,31%	76,19%	74,60%
Classificador 2 - MLP	73,12%	64,54%	70,99%
Classificador 3 - MLP	65,62%	79,16%	82,29%
2º Execução			
Modelos	**Acurácia**		
	Repetição 1	**Repetição 2**	**Repetição 3**
Classificador 1 - MLP	67,19%	70,89%	68,78%
Classificador 2 - MLP	69,84%	63,41%	63,45%
Classificador 3 - MLP	68,75%	79,16%	79,16%
3º Execução			
Modelos	**Acurácia**		
	Repetição 1	**Repetição 2**	**Repetição 3**
Classificador 1 - MLP	74,07%	75,66%	68,78%
Classificador 2 - MLP	68,84%	69,91%	64,61%
Classificador 3 - MLP	69,79%	67,71%	78,12%
4º Execução			
Modelos	**Acurácia**		
	Repetição 1	**Repetição 2**	**Repetição 3**
Classificador 1 - MLP	68,25%	71,95%	73,54%
Classificador 2 - MLP	69,93%	72,01%	60,17%
Classificador 3 - MLP	70,83%	64,58%	78,12%
5º Execução			
Modelos	**Acurácia**		
	Repetição 1	**Repetição 2**	**Repetição 3**
Classificador 1 - MLP	73,54%	74,60%	75,66%
Classificador 2 - MLP	75,25%	76,36%	63,52%
Classificador 3 - MLP	65,62%	69,79%	76,04%

Fonte: Elaborado pelo autor (2022).

APÊNDICE E - ESTIMATIVAS GERADAS PELOS MODELOS DURANTE TREINO/VALIDAÇÃO NA REGRESSÃO

Tabela E.1: Estimativas médias dos modelos para Classificador 1 e Regressores 2 e 3, utilizando o conjunto com unigramas.

Classificador 1			
Modelos	**Repetição 1**	**Repetição 2**	**Repetição 3**
	Acurácia Média		
MLP	62,52%	65,66%	67,16%
Regressor 2			
	Repetição 1	**Repetição 2**	**Repetição 3**
	Média Métrica MSE		
SVR	0,4984	0,5570	0,6529
K-NNR	0,5049	0,5855	0,7400
MLPR	0,4865	0,6355	0,6400
	Média Métrica MAPE		
SVR	32,65%	35,26%	38,73%
K-NNR	32,98%	36,51%	42,03%
MLPR	32,34%	37,75%	39,22%
Regressor 3			
	Repetição 1	**Repetição 2**	**Repetição 3**
	Média Métrica MSE		
SVR	0,3487	0,3854	0,4399
K-NNR	0,3503	0,3429	0,4895
MLPR	0,3669	0,3429	0,4074
	Média Métrica MAPE		
SVR	13,08%	13,61%	15,06%
K-NNR	12,24%	12,22%	15,27%
MLPR	13,71%	14,06%	14,86%

Fonte: Elaborado pelo autor (2022).

ANEXOS

ANEXO A - NÚMEROS DO CENSUP

Figura A.1: Número de Instituições de Educação Superior, por Organização Acadêmica e Categoria Administrativa - Brasil - 2009-2019.

Ano	Instituições								
	Total	Universidade		Centro Universitário		Faculdade		IF e Cefet	
		Pública	Privada	Pública	Privada	Pública	Privada	Pública	Privada
2009	**2.314**	100	86	7	120	103	1.863	35	n.a.
2010	**2.378**	101	89	7	119	133	1.892	37	n.a.
2011	**2.365**	102	88	7	124	135	1.869	40	n.a.
2012	**2.416**	108	85	10	129	146	1.898	40	n.a.
2013	**2.391**	111	84	10	130	140	1.876	40	n.a.
2014	**2.368**	111	84	11	136	136	1.850	40	n.a.
2015	**2.364**	107	88	9	140	139	1.841	40	n.a.
2016	**2.407**	108	89	10	156	138	1.866	40	n.a.
2017	**2.448**	106	93	8	181	142	1.878	40	n.a.
2018	**2.537**	107	92	13	217	139	1.929	40	n.a.
2019	**2.608**	108	90	11	283	143	1.933	40	n.a.

Fonte: MEC/Inep [36].

Figura A.2: Número de Cursos de Graduação, por Modalidade de Ensino e por Grau Acadêmico - Brasil - 2009-2019.

Ano	Cursos de Graduação										
	Total Geral	Modalidade de Ensino/Grau Acadêmico									
		Presencial					A distância				
		Total	Bacharelado	Licenciatura	Tecnológico	Bacharelado/ Licenciatura	Total	Bacharelado	Licenciatura	Tecnológico	Bacharelado/ Licenciatura
2009	**28.671**	27.827	15.663	6.697	4.491	976	844	157	485	200	2
2010	**29.507**	28.577	16.401	7.401	4.775	n.a.	930	185	521	224	n.a.
2011	**30.420**	29.376	16.832	7.352	5.192	n.a.	1.044	199	559	286	n.a.
2012	**31.866**	30.718	17.486	7.613	5.619	n.a.	1.148	217	581	350	n.a.
2013	**32.049**	30.791	17.665	7.328	5.798	n.a.	1.258	240	592	426	n.a.
2014	**32.878**	31.513	18.319	7.261	5.933	n.a.	1.365	290	595	480	n.a.
2015	**33.501**	32.028	18.938	7.004	6.086	n.a.	1.473	316	625	532	n.a.
2016	**34.366**	32.704	19.795	6.693	6.216	n.a.	1.662	387	663	612	n.a.
2017	**35.380**	33.272	20.578	6.501	6.193	n.a.	2.108	525	771	812	n.a.
2018	**37.962**	34.785	21.882	6.419	6.484	n.a.	3.177	855	996	1.326	n.a.
2019	**40.427**	35.898	23.083	6.391	6.424	n.a.	4.529	1.319	1.234	1.976	n.a.

Fonte: MEC/Inep [36].