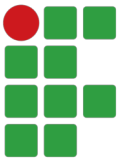

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DA PARAÍBA - CAMPUS CAMPINA GRANDE
CURSO SUPERIOR DE TECNOLOGIA EM TELEMÁTICA

ELAYNE REGINA LIMA SILVA

CLASSIFICAÇÃO AUTOMÁTICA DE QUESTÕES DE CONCURSOS: UM ESTUDO COMPARATIVO UTILIZANDO ALGORITMOS DE MINERAÇÃO DE TEXTOS

Campina Grande

2022

Elayne Regina Lima Silva

CLASSIFICAÇÃO AUTOMÁTICA DE QUESTÕES DE CONCURSOS: UM ESTUDO COMPARATIVO UTILIZANDO ALGORITMOS DE MINERAÇÃO DE TEXTOS

Monografia apresentada à Coordenação do Curso Superior de Telemática do IFPB - Campus Campina Grande, como requisito parcial para conclusão do curso Superior de Tecnologia em Telemática.

Orientador: Prof. Dr. Elmano Ramalho Cavalcanti

Campina Grande

2022

S586c Silva, Elayne Regina Lima.
Classificação automática de questões de concursos: um estudo comparativo utilizando algoritmos de mineração de textos / Elayne Regina Lima Silva. - Campina Grande, 2022.
28 f. : il.

Trabalho de Conclusão de Curso (Curso de Tecnologia em Telemática) - Instituto Federal da Paraíba, 2022.
Orientador: Prof. Dr.Elmano Ramalho Cavalcanti.

1. Mineração de texto 2. Classificação supervisionada 3. Acurácia I. Título.

CDU 004

Elayne Regina Lima Silva

CLASSIFICAÇÃO AUTOMÁTICA DE QUESTÕES DE CONCURSOS: UM ESTUDO COMPARATIVO UTILIZANDO ALGORITMOS DE MINERAÇÃO DE TEXTOS

Monografia apresentada à Coordenação do Curso Superior de Telemática do IFPB - Campus Campina Grande, como requisito parcial para conclusão do curso Superior de Tecnologia em Telemática.

Prof. Dr. Elmano Ramalho Cavalcanti
Orientador

Prof.ª M.ª Iana Daya Cavalcante Facundo Passos
Membro da Banca

Prof. Dr. Marcelo José Siqueira Coutinho De Almeida
Membro da Banca

Campina Grande
2022

Dedicatória

Dedico este trabalho a Deus e a todos que dedicam parte do seu tempo à construção de novos conhecimentos e que buscam a transformação do mundo através da ciência e educação.

Agradecimentos

Agradeço a Deus, a quem sou inteiramente grata por me conduzir até aqui neste novo caminho que escolhi.

A Maria Santíssima, minha fiel intercessora em todos os momentos da minha vida.

Aos meus pais Assis Epifânio da Silva e Tânia Maria Lima Silva por todo amor e apoio incondicionais e a formação do ser humano e profissional que sou hoje. A eles todo o meu amor e consideração.

Aos meus irmãos Aline, Thiago e Matheus, meus parceiros de vida e aos meus cunhados Diego e Kessya que torcem sempre por mim.

A todos os meus familiares e amigos por todo apoio na jornada da vida.

Ao prezado orientador, Professor Elmano Ramalho Cavalcanti que tornou a etapa final mais leve. Agradeço imensamente pelo compromisso, seriedade e parceria ao transmitir seus conhecimentos comigo e por toda empatia e paciência durante execução deste trabalho.

Agradeço à banca examinadora pela contribuição: Professora Ianna Daya, que é meu grande exemplo e minha grande incentivadora desde o início do curso, sempre me dando oportunidade e voz, mostrando que eu poderia ir mais longe do que imaginei e Professor Marcelo Siqueira que é grande referência no Campus e que tenho a honra de compartilhar conhecimento neste trabalho.

Ao IFPB e aos demais docentes, por todo conhecimento compartilhado em todos esses anos de curso.

Aos amigos e colegas que fiz nessa trajetória e a todos os colaboradores e servidores da instituição.

Resumo

No Brasil, o ingresso no funcionalismo público tem como requisito obrigatório a realização de processos seletivos e avaliativos chamados de Concursos Públicos. Esses processos seletivos têm como uma das principais características a extensa quantidade de conteúdo programático nos seus editais, que exige do candidato preparo prévio e dedicação de muitas horas de seu dia aos estudos. Nessa etapa de preparação, uma das alternativas consideradas mais eficazes é estudar resolvendo as questões cobradas pelas bancas elaboradoras em concursos anteriores. Este estudo tem como objetivo identificar qual método de mineração de texto é mais indicado para classificar automaticamente questões de concursos de forma mais eficiente, tendo como justificativa reduzir os altos custos necessários para classificação ao lidar com bancos de questões robustos. A automatização desta tarefa pode ser muito útil para a empresa que fornece o serviço e para os usuários das plataformas de estudos e aplicar algoritmos de classificação em questões de um banco já classificado possibilita o aprendizado dos métodos utilizados para que sejam usados em bancos de questões onde não haja classificação. Para esse estudo, foram utilizadas 20 mil questões relacionadas à área de Informática extraídas da *EdTech* QConcursos, escritas em linguagem natural e que portanto necessitaram passar pelas etapas de mineração de texto. Dentre os modelos, destacaram-se Naïve Bayes, Generalized Linear, Árvore de Decisão, Floresta Aleatória, Gradient Boosted, que apresentaram valores de acurácia superiores a 90% quando aplicados a disciplinas. Quando aplicados a assuntos de questões, os maiores valores foram iguais ou inferiores a 24%.

Palavras-chaves: Mineração de texto, concursos públicos, acurácia.

Abstract

In Brazil, entry into the civil service has as a mandatory requirement the performance of selective and evaluative processes. One of the main characteristics of these selection processes is the extensive amount of program content in their notices, which requires the candidate to prepare in advance and dedicate many hours of his day to studies.At this stage of preparation, one of the alternatives considered to be the most effective is to study by solving the questions charged by the drafting committees in previous contests. This study aims to identify which text mining method is best suited to automatically classify contest questions more efficiently, with the justification of reducing the high costs required for classification when dealing with robust question databases. The automation of this task can be very useful for the company that provides the service and for the users of the study platforms, and applying classification algorithms to questions from an already classified bank makes it possible to learn the methods used to be used in question banks where there is no classification. For this study, 20 thousand questions related to the area of Informatics extracted from EdTech QConcursos were used, written in natural language and that therefore needed to go through the text mining steps. Among the models, Naïve Bayes, Generalized Linear, Decision Tree, Random Forest, Gradient Boosted stood out, which presented accuracy values greater than 90% when applied to disciplines. When applied to question subjects, the highest values were equal to or less than 24%.

Key-words: *Text mining, Civil Service Exam, accuracy.*

Sumário

1 DEFININDO O PROBLEMA

1.1 INTRODUÇÃO

No Brasil, o ingresso no funcionalismo público tem como requisito obrigatório a realização de processos seletivos e avaliativos chamados de Concurso Público para medir as competências dos candidatos entre si a um cargo efetivo em um determinado órgão público.

Esses processos seletivos têm como uma das principais características a extensa quantidade de conteúdo programático nos seus editais, que exige do candidato preparo prévio e dedicação de muitas horas de seu dia aos estudos. Nessa etapa de preparação, uma das alternativas consideradas é a estratégia de estudar resolvendo as questões cobradas pelas bancas elaboradoras em concursos anteriores.

Assim, os candidatos podem conhecer a forma com que a banca examinadora à frente do concurso elabora a prova e, consequentemente, verificar como os temas podem ser abordados ao longo da avaliação. Estudando por questões de concurso é possível potencializar de uma forma menos cansativa o aprendizado do conteúdo estudado anteriormente e até mesmo estimular a aprender mais sobre as disciplinas predominantes nas provas. Além disso, auxilia no teste da eficácia das rotinas de estudos do candidato, sendo possível monitorar e autoavaliar o seu desempenho.

Nesse cenário, existem atualmente diversas plataformas *Edtech* de Ensino à Distância (EAD) em versões gratuitas ou por assinatura, que proporcionam ambientes completos de estudos para os candidatos terem acesso aos conteúdos presentes nos bancos de questões, assim como videoaulas, guias de estudo, dentre outros recursos. Porém, como existe um grande número dessas questões, é importante saber quais delas são as mais relevantes.

Um modelo adequado para classificar automaticamente as questões desejadas pode tornar possível a criação de um sistema de recomendação de questões de concurso público mais eficaz, considerando o maior número possível de classes que auxilie na redução do custo computacional e melhore o resultado na obtenção de informações relevantes, tornando mais eficiente e ágil o serviço prestado aos usuários, além de impulsionar o uso dessas plataformas.

1.2 OBJETIVOS

1.2.1 **Objetivo Geral**

Obter o(s) modelo(s) de classificação de texto mais indicado(s) para classificar automaticamente questões de concursos de forma mais eficiente.

1.2.2 **Objetivo Específicos**

- Obter modelos de classificação de assuntos de questões da área de informática;
- Obter os valores de acurácia do(s) melhor(es) modelo(s) classificador(es);
- Comparar pelo menos seis algoritmos de classificação supervisionada;
- Identificar o algoritmo de estado da arte de aprendizado que maximize a precisão e acurácia de classificação dos textos;

1.3 JUSTIFICATIVA

A principal justificativa para essa pesquisa se dá pelos custos necessários para classificação ao lidar com bancos de questões robustos. A automatização desta tarefa pode ser muito útil para a empresa que fornece o serviço e para os usuários das plataformas de estudos. Além disso, aplicando algoritmos de classificação em questões de um banco já classificado possibilita o aprendizado dos métodos utilizados para que sejam aplicados a bancos de questões onde não haja classificação.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho está dividido em 5 capítulos. O Capítulo 2 aborda a Fundamentação Teórica e Estado da Arte, o Capítulo 3 Descreve aspectos sobre a solução proposta, o Capítulo 4 apresenta os resultados obtidos e por fim, o Capítulo 5 aponta as conclusões e propostas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta o levantamento do estado da arte, abrangendo o conteúdo de artigos correlatos e alguns conceitos essenciais para compreensão da Mineração de textos, englobando também o processamento, preparo dos textos, os algoritmos de classificação utilizados e os métodos de avaliação dos modelos preditivos gerados.

2.1 MINERAÇÃO DE TEXTO

A mineração de dados é a exploração e análise, automatizada ou semi-automatizada, de grandes quantidades de dados, que busca descobrir padrões e regras significativas (SANTOS, 2013).

Semelhante à mineração de dados, a mineração de texto é uma área da mineração de dados que trabalha com dados não-estruturados. A classificação supervisionada e não-supervisionada de documentos está entre as áreas de pesquisa da mineração de texto (CAVALCANTI et al., 2011).

Envolve várias áreas da informática, sendo um modelo de programação criado para resolver problemas de imprecisão e incertezas nos documentos de texto (PEZZINI, 2017). Também pode ser definido como uma forma de obter conhecimento, usando técnicas de extração e análise de dados a partir de palavras, textos ou frases. Por meio dele é possível extrair padrões importantes, não triviais ou conhecimento a partir de documentos em textos não estruturados. No processo de classificação de textos tem-se como entrada um documento de um conjunto fixo de classes $C = c_1, c_2, ..., c_j$. A saída será determinar a classe sobre à qual o documento d está semanticamente relacionado, ou seja, dado um documento, será assinalada a classe à que o mesmo pertence e (ANDRADE, 2015).

Os modelos de classificação podem ser divididos em single-label, onde cada documento pode pertencer a apenas uma classe ou multi-label, quando um documento pode ser associado a uma ou mais categorias. Usando essa classificação, é possível obter uma solução para classificação de textos em larga escala (SILVA; MEDEIROS, 2020).

2.1.1 Etapas da Mineração de Texto

As etapas gerais que compõem a classificação de textos são: coleta, pré-processamento, indexação e análise da informação(PEZZINI, 2017).

Em seu estudo, (VIANA, 2021), explica cada etapa da seguinte maneira:

- Coleta de dados: tem como função, coletar os dados que formarão a base de dados de textos.
- Pré-processamento de textos: Visa realizar transformações sobre os textos com o objetivo de estruturar tais textos. As transformações nos textos consistem em identificar, compactar e tratar dados que possam estar corrompidos, atributos irrelevantes no processo e valores desconhecidos.
- Indexação: armazena as palavras dos textos em uma estrutura de índices que permitem uma busca rápida por uma palavra-chave em grandes volumes de textos.
- A análise da informação é a fase de interpretação dos resultados, onde estes são analisados, verificando-se depois se a mineração do texto atingiu um bom resultado.

Os autores Aranha e Vellasco (2007), descrevem o processo de mineração de textos no seguinte conjunto de etapas:

Figura 1 – Etapas do Processo de Mineração de Textos

Fonte: Adaptado de (MORAIS; AMBRÓSIO, 2007)

2.2 APRENDIZADO DE MÁQUINA SUPERVISIONADO

O aprendizado de máquina significa dar a uma máquina a habilidade de aprender, mesmo que nem tudo esteja explicitamente programado. Podemos dividir o aprendizado de máquina em supervisionado, que busca responder um *target*, ou seja, há uma variável explícita a ser respondida e a não supervisionada, em que busca-se identificar grupos ou padrões a partir dos dados, sem um objetivo específico a ser alcançado. No aprendizado de máquina não supervisionado que não possui rótulos em seus dados e o algoritmo é forçado a compreendê-los sozinho.

De acordo com (SILVA; MEDEIROS, 2020), o aprendizado de máquina supervisionado é um dos tipos mais comuns e é utilizado sempre que se precisa prever um determinado resultado de uma determinada entrada, usando pares de entrada e saída. Nos

modelos de aprendizagem de máquina supervisionada, conseguimos dar pesos ou calibrar o nível de assertividade e de precisão de um modelo.

2.3 MODELOS DE CLASSIFICAÇÃO SUPERVISIONADA

2.3.1 *Máquinas de Vetores de Suporte* - MVS

Máquinas de Vetores de Suporte (MSV), criado por Vapnik é um método matemático capaz de classificar documentos de forma supervisionada. O método consiste em utilizar hiperplanos para separar os dados, deixando a maior margem possível ao separar as classes. Ao encontrar um hiperplano de margem máxima que separa as classes, a classificação de um novo ponto será trivial(ANDRADE, 2015).

MSV abordam os conceitos de uma aprendizagem supervisionada através de uma teoria matemática muito bem fundamentada. O processo de aprendizagem destes classificadores fundamenta-se na busca de minimizar tanto o risco empírico quanto o risco estrutural (MAIA, 2019).

Em seu estudo, (JOACHIMS, 1998) propôs a aplicação do método MSV à classificação de texto. O MSV é considerado um dos classificadores mais populares do tipo linear e tem como objetivo criar um limite de decisão entre duas classes que permite a previsão de rótulos de um ou mais recursos vetores (HUANG et al., 2018).

De acordo com a proposta de aplicação supracitada, as MSV conseguem um ótimo desempenho em tarefas executadas para categorizar texto, superando consideravelmente os métodos existentes por sua capacidade de generalizar bem em espaços de recursos de alta dimensão, eliminando a necessidade de seleção de recursos, facilitando a aplicação da categorização de texto (SILVA; MEDEIROS, 2020).

2.3.2 Regressão Logística

Segundo (SANTOS, 2013) o modelo de Regressão Logística é um algoritmo de classificação frequentemente usado por muitos anos na área de estatística, mas que começou a ser utilizado na área de aprendizado de máquina recentemente, devido à proximidade com o MSV. Com esse modelo, é possível analisar a relação entre uma variável dependente categórica e diversas variáveis independentes, estimando a probabilidade de ocorrência de um evento ajustando os dados a uma curva logística, usada para prever um resultado binário.

O Modelo de Regressão Logística Binário é um caso particular dos modelos lineares generalizados, em que o componente aleatório tem distribuição de Bernoulli e a função de ligação é o *logit* (KANASHIRO, 2022).

2.3.3 *Naïve Bayes*

O modelo de *Naïve Bayes* tem como base o teorema de Bayes, criado pelo matemático Thomas Bayes e é muito utilizado para grande quantidade de dados e por sua simplicidade de aplicação em sistemas complexos obtendo resultados relativamente(ROZA; PEGORARO, 2020).

É um algoritmo probabilístico de classificação capaz de estimar a probabilidade para as classes. Seu objetivo é calcular a probabilidade de um dado documento para escolher a melhor classe em que ele se encaixa. Pode ser chamado de classificador ingênuo, por assumir independência entre atributos, dado um valor da classe.

Dois modelos de classificadores *Naïve Bayes* são conhecidos: o modelo binário e o modelo multinomial. O modelo binário, também chamado de básico, especifica um documento representado por um vetor de atributos binários, indicando quais palavras ocorrem ou não no documento, descrevendo, dessa forma, uma distribuição baseada em um modelo de evento Bernoulli multivariável. Já o segundo modelo especifica que um documento é representado pelo conjunto de ocorrências de palavras do documento, denominado modelo de evento multinomial (SILVA; MEDEIROS, 2020; ANDRADE, 2015).

2.3.4 *Gradient Boosting*

Para entender o modelo de aprendizado de máquina *Gradient Boosting* faz-se necessário compreender que este método de treinamento de *ensemble* que são métodos que usam múltiplos algoritmos de aprendizado para obter um desempenho preditivo considerado ótimo. Ele pode ser usado com qualquer tipo de classificador, para juntar aqueles considerados mais fracos, tornando-se em um modelo mais robusto (BARBOSA et al., 2021).

Para isso, ele utiliza uma técnica chamada de *boosting* que distribui diferentes pesos para cada modelo aumentando a influência do modelo com melhor desempenho no resultado final. O ganho de desempenho obtido ocorre por que os modelos anteriores influenciam no peso dos posteriores, treinados para correção de erros dos antecessores. Quando utilizado para classificação, o *Gradient Boosting* soma a predição de diversos modelos sobre as probabilidades de um dado ser categorizado em uma determinada classe, ajustadas pelas taxas de aprendizagem.Sendo assim, caso um dado tenha mais de 50% de chance de pertencer a uma determinada classe, ela é classificada como tal (BARBOSA et al., 2021; LOPES, 2022).

2.3.5 Árvores de Decisão ou *Decision Tree*

Árvores de decisão (AD ou *Decision Tree*) podem ser definidas como as formas de representar o conhecimento, construindo classificadores para prever em qual a classe

os dados desconhecidos pertencem, sendo essas informações baseadas nos valores de um conjunto de dados, ou seja, métodos de classificação de dados (OKADA; NEVES; SHITSUKA, 2019). É uma estrutura hierárquica na qual cada nó interno representa uma decisão em um dos atributos do conjunto de dados, e cada ramo representa uma tomada de decisão (GUSMÃO; FIGUEIREDO; BRITO, 2021).

De acordo com (NETO et al., 2020), a classificação por AD funciona como um fluxograma com formato de árvore, onde cada nó indica um teste realizado sobre um determinado valor e as ligações entre cada nó representam os valores possíveis do teste do nó superior. As folhas indicam à classe a qual o registro pertence. Sendo assim, após o modelo da árvore ser construído pode-se classificar um novo registro seguindo o fluxo da árvore do nó raiz até a folha. A seguir é possível visualizar um exemplo da arquitetura de uma árvore de decisão.

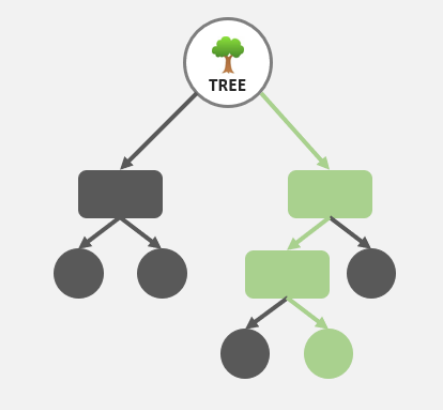

Figura 2 – Esquema gráfico de uma Árvore de Decisão

Fonte: (PESSANHA, 2019)

2.3.6 Floresta Aleatória

No estudo realizado por (MÜLLER; GUIDO, 2016), os autores definem Floresta Aleatória (RF ou *Random Forest*) como uma coleção projetada especificamente para árvores de decisão. Nela, cada árvore apresenta-se um pouco diferente da outra. A ideia desse modelo é combinar as previsões feitas por várias árvores de decisão que foram geradas com base nos valores de um conjunto independente de vetores aleatórios. Floresta aleatória é um método de aprendizagem de máquina amplamente utilizado na atualidade, direcionado a problemas de regressão e classificação. A diversidade representada por várias árvores montadas após selecionadas de forma aleatória de atributos e de exemplos (para evitar correlação entre as árvores), reduz o sobreajuste dos dados de treinamento (*overfitting*).

Geralmente, cada árvore que forma uma floresta tem igual importância na predição final. Na classificação e predição final, a decisão se dá por voto majoritário, diferentemente da regressão, onde a decisão final é obtida através da média entre as decisões tomadas

individualmente (CARVALHO et al., 2021). A seguir, na figura 4, é possível visualizar um exemplo de Floresta Aleatória.

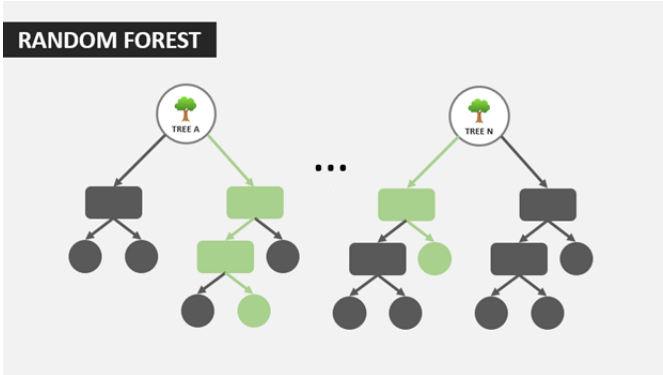

Figura 3 – Esquema gráfico de uma Floresta Aleatória

Fonte: (PESSANHA, 2019)

2.4 AVALIAÇÃO DA PERFORMANCE DE CLASSIFICADORES

Neste estudo será necessária a aplicação de métricas de avaliação para validar o resultado gerado pelos classificadores. Para (SEBASTIANI, 2002), experimentos para avaliar um classificador geralmente medem sua efetividade, ou seja, mensuram a sua capacidade de tomar as decisões mais assertivas de classificação. Na mineração de textos e aprendizagem de máquinas, a avaliação dos resultados é feita utilizando métricas de desempenho, como Acurácia, *Recall*, *F-measure*, *Overfitting*, *Underfitting*, dentre outras. Sendo assim, deve-se considerar as noções de relevância na qualidade de recuperação e acerto dos classificadores.

2.4.1 **Acurácia**

Acurácia (ACC) é a métrica que indica quanto o modelo teve índice de acerto dentro das previsões possíveis, ou seja, a quantidade de acertos dentro do total de predições realizadas no teste, frente aos valores verdadeiros positivos e verdadeiros negativos (SIQUEIRA, 2022; VIANA, 2021).

É a razão entre as classificações feitas corretamente pelo modelo dividido por todas as classificações positivas e negativas que foram feitas, seja classificação de verdadeiros positivos (VP) e verdadeiros negativos (VN). Já as classificações incorretas, são aquelas classificações falso positivo (FP) e falso negativo (FN), e ocorrem quando o modelo classifica incorretamente alguma frase de forma positiva ou negativa, respectivamente (CAMPOS, 2022) . A seguir, pode-se observar a equação que apresenta a fórmula para o cálculo da Acurácia.

$$Acuracia = \frac{VP + VN}{VP + FN + VN + FP} \tag{2.1}$$

2.4.2 *Recall*

Recall ou sensibilidade, é a porcentagem de itens relevantes que foi recuperada. Mede a porcentagem das observações positivas que foram classificadas de forma correta pelo modelo (VIANA, 2021). Isso é possível, calculando a taxa que o método classifica a execução como pertencente a um determinado grupo, podendo ser baixa, média ou alta, em relação a todos que de fato fazem parte dele (OLIVEIRA; BOERES; OLIVEIRA, 2021). Pode-se representar o *recall* com a equação abaixo.

$$Recall = \frac{VP}{VP + FN} \tag{2.2}$$

2.4.3 *F-measure* ou *F-score*

O *F-measure* ou *F-score* , também é conhecida como a pontuação F, é uma média harmônica ponderada entre o *recall* e a precisão (ANDRADE, 2015). Sendo assim, esta métrica determina o quanto o modelo de previsão consegue prever corretamente os valores positivos (OSHITA, 2021).

$$F\text{-}score = 2 * \frac{(RecallX Precisao)}{(Recall + Precisao)} \tag{2.3}$$

2.4.4 *Overfitting* e *Underfitting*

O *Overfitting* ou super-ajuste é um fenômeno que ocorre quando o modelo criado está tão bem ajustado aos dados e não consegue generalizar bem para novas amostras. Atua fazendo com que o modelo se ajuste para tratar casos discrepantes em relação ao restante dos dados, pois o modelo ideal deve apresentar um balanceamento entre o bom ajuste a base de teste, assim como boa generalização para novos dados (SANTOS, 2013).

O *Underfitting* ou sub-ajuste é um modelo que não é capaz de se ajustar e explicar o problema proposto por BARBOSA et al. (2021). Nesse modelo os dados de treinamento tem um desempenho insatisfatório, fato que ocorre porque ele não consegue capturar a relação entre os exemplos de entrada (X) e os valores de destino (Y).

3 SOLUÇÃO PROPOSTA

Este capítulo visa descrever a metodologia utilizada no presente estudo, assim como propor solução para a prova de conceito de classificação de questões utilizando a mineração de texto.

3.1 METODOLOGIA

Pesquisa inicial com abordagem qualitativa de cunho bibliográfico e documental com atenção nas análises sobre os modelos de classificação executados em diversos estudos, para embasar os testes e o processamento e resultados.

As etapas aplicadas neste estudo de caso seguem a metodologia do processo de mineração de dados apresentada a seguir, a qual é composta pelas seguintes etapas de seleção dos dados, pré-processamento e modelagem.

Posteriormente, realizou-se a análise quantitativa dos resultados obtidos no que concerne a performance dos classificadores, a partir dos cenários observados após processamento do banco de questões.

3.1.1 Seleção dos Dados

Para esse estudo, foram utilizados bancos de questões relacionadas à área de Informática da *EdTech* QConcursos. A base de dados selecionada possui em sua totalidade uma amostra mais de 20 mil questões, escritas em linguagem natural e portanto necessitam passar por modificações para que os algoritmos que serão utilizados nas próximas etapas sejam capazes de manipular todo seu conteúdo, conforme será possível observar no decorrer do capítulo.

3.1.2 Pré-Processamento

O banco de dados gerado com as questões de concursos foi extraído com o Sistema Gerenciador de Banco de Dados (SGBD) *MySQL Workbench*, carregados e processados no *software* de mineração de texto *RapidMiner*.A escolha destas ferramentas para mineração se deu por ser de maior praticidade na manipulação e entendimento dos resultados obtidos, assim como a possibilidade de utilização em máquina local.

Sabendo-se que o pré-processamento objetiva identificar, compactar e tratar dados possivelmente corrompidos, atributos considerados irrelevantes no processo e valores

desconhecidos, esta etapa foi executada imediatamente após o agrupamento dos dados. O conteúdo dos bancos encontrava-se separado por 09 disciplinas da área de informática:

1. Algoritmos e Estrutura de Dados;
2. Arquitetura de Computadores;
3. Arquitetura de *Software*;
4. Bancos de Dados;
5. Engenharia de *Software*;
6. Gerência de Projetos;
7. Programação;
8. Segurança da Informação; e
9. Sistemas Operacionais.

Os dados foram agrupados em um banco único convertido de *.sql* para o formato.*csv* e posteriormente preparados para que pudessem ser processados pelos modelos escolhidos. O banco de questões, após agrupamento, totalizou 91.048 questões e visando reduzir a dimensionalidade dos textos, fez-se inicialmente a limpeza e a padronização do conteúdo dos arquivos. Foram removidos acentuação e espaçamentos, questões duplicadas, sendo também tratadas as colunas e linhas que continham valor vazio (*NULL*). Essa ação resultou em uma redução do banco para 81.048 registros.

Para análise, o banco sofreu uma nova redução, contendo o valor total de aproximadamente 20 mil questões, subdivididas em amostras menores de 15mil e 4mil durante o processo. Desse modo, seria possível a melhor observação dos processos e resultados.

3.1.3 Modelagem

A modelagem foi realizada importando o banco de questões no Software RapidMiner após a etapa de processamento, os dados foram inseridos na função Automodel do sistema e a partir disso, selecionados alguns modelos disponíveis na aplicação como *Naïve Bayes*, *Generalized Linear*, Máquina de Suporte de Vetores, Floresta Aleatória, Árvore de Decisão e *Gradient Boosted*, a fim de obter modelos de classificação das questões da área de informática tanto por disciplinas quanto por assuntos.

3.1.4 ***RapidMiner***

O Software RapidMiner é uma ferramenta utilizada para análise de dados de diversas áreas do conhecimento, que faz uso de aprendizagem de máquina que tem como principal objetivo, acelerar o processo de criação de análises preditivas e facilitar sua aplicação em cenários práticos de negócios. Possui uma biblioteca com mais de 1500 máquinas de aprendizagem de algoritmos e funções para construir o modelo preditivo mais forte possível para qualquer caso de uso. É de código aberto e extensível para fácil integração com aplicações existentes, dados e programação de línguas como *Python*e R.

4 RESULTADOS

Neste capítulo, são apresentados os achados mais importantes que foram identificados após a execução dos modelos de classificação. Os resultados alcançados foram validados de acordo com algumas das métricas consideradas pela literatura.

4.1 CLASSIFICAÇÃO DE DISCIPLINAS DE QUESTÕES DA ÁREA DE INFORMÁTICA

A seguir, pode-se observar na Figura 4 os valores de acurácia encontrados após aplicação dos modelos de classificação com base nas 09 disciplinas contidas no banco de questões, correspondente a amostra de aproximadamente 15 mil questões. Os modelos com melhores performances foram Decision Tree e Gradient Boosted Tree ambos com 100% de acurácia, seguido do Generalized Linear com 83%, os demais modelos obtiveram valores inferiores.

Nas figuras 5, 6 e 7 podem ser observadas respectivamente as matrizes de Confusão dos modelos Árvore de Decisão e *Gradiente Boosted Tree* com acerto de 100% Generalized Linear com 83%.

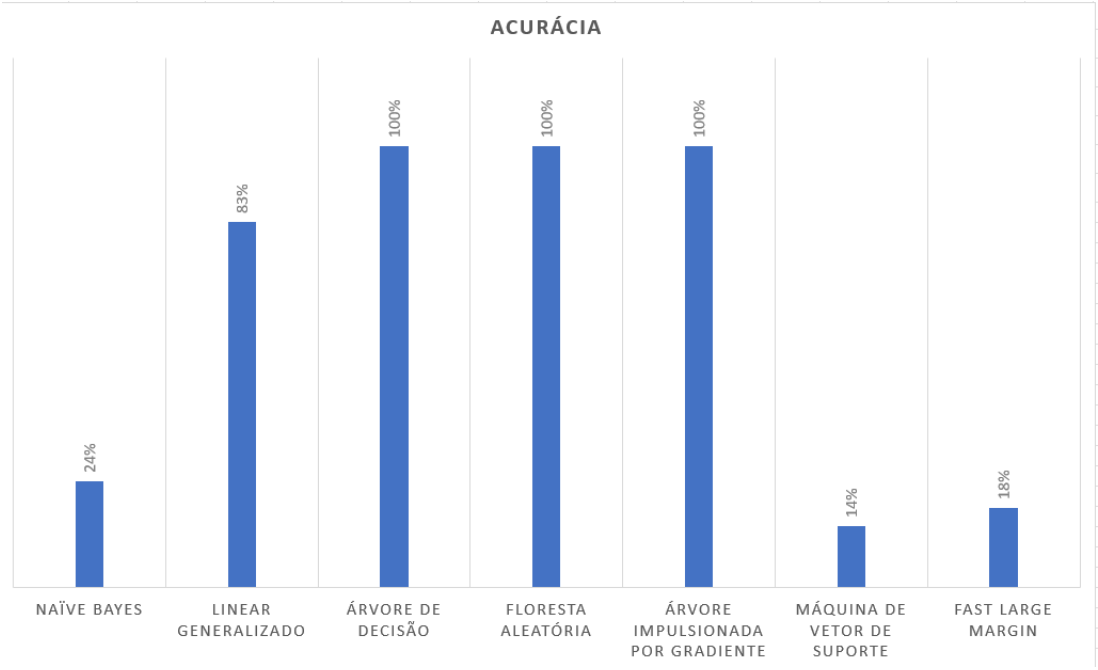

Figura 4 – Acurácia nos Modelos de Classificação de acordo com as disciplinas

Fonte: A autora

Decision Tree - Performance

	true Algoritm...	true Arquitet...	true Arquitet...	true Bancod...	true Engenh...	true Gerenci...	true Progra...	true Seguran...	true Sistema...	class precisi...
pred. Algorit...	482	0	0	0	0	0	0	0	0	100.00%
pred. Arquite...	0	483	0	0	0	0	0	0	0	100.00%
pred. Arquite...	0	0	483	0	0	0	0	0	0	100.00%
pred. Banco...	0	0	0	482	0	0	0	0	0	100.00%
pred. Engen...	0	0	0	0	484	0	0	0	0	100.00%
pred. Gerenc...	0	0	0	0	0	482	0	0	0	100.00%
pred. Progra...	0	0	0	0	0	0	484	0	0	100.00%
pred. Segura...	0	0	0	0	0	0	0	482	0	100.00%
pred. Sistem...	0	0	0	0	0	0	0	0	483	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

Figura 5 – Matriz de Confusão do Modelo *Decision Tree*

Gradient Boosted Trees - Performance

	true Algoritm...	true Arquitetu...	true Arquitetu...	true Bancode...	true Engenh...	true Gerenci...	true Program...	true Seguran...	true Sistema...	class precisi...
pred. Algorit...	482	0	0	0	0	0	0	0	0	100.00%
pred. Arquitet...	0	483	0	0	0	0	0	0	0	100.00%
pred. Arquitet...	0	0	483	0	0	0	0	0	0	100.00%
pred. Bancod...	0	0	0	482	0	0	0	0	0	100.00%
pred. Engen...	0	0	0	0	484	0	0	0	0	100.00%
pred. Gerenc...	0	0	0	0	0	482	0	0	0	100.00%
pred. Progra...	0	0	0	0	0	0	484	0	0	100.00%
pred. Segura...	0	0	0	0	0	0	0	482	0	100.00%
pred. Sistem...	0	0	0	0	0	0	0	0	483	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

Figura 6 – Matriz de Confusão do Modelo *Gradient Boosted Tree*

Generalized Linear Model - Performance

	true Algoritm...	true Arquitet...	true Arquitet...	true Bancod...	true Engenh...	true Gerenci...	true Progra...	true Seguran...	true Sistema...	class precisi...
pred. Algorit...	419	4	2	4	0	1	1	0	2	96.77%
pred. Arquite...	1	111	1	0	1	0	0	0	2	95.69%
pred. Arquite...	61	365	477	136	0	0	2	0	119	41.12%
pred. Banco...	2	0	2	342	0	0	0	0	2	98.28%
pred. Engen...	0	0	1	0	482	1	0	0	0	99.59%
pred. Gerenc...	0	0	0	0	0	481	0	0	0	100.00%
pred. Progra...	0	0	0	0	0	0	480	0	0	100.00%
pred. Segura...	0	0	0	0	0	0	0	483	0	100.00%
pred. Sistem...	0	2	0	1	0	0	0	0	357	99.17%
class recall	86.75%	23.03%	98.76%	70.81%	99.79%	99.59%	99.38%	100.00%	74.07%	

Figura 7 – Matriz de Confusão do Modelo *Generalized Linear*

Fonte: A autora

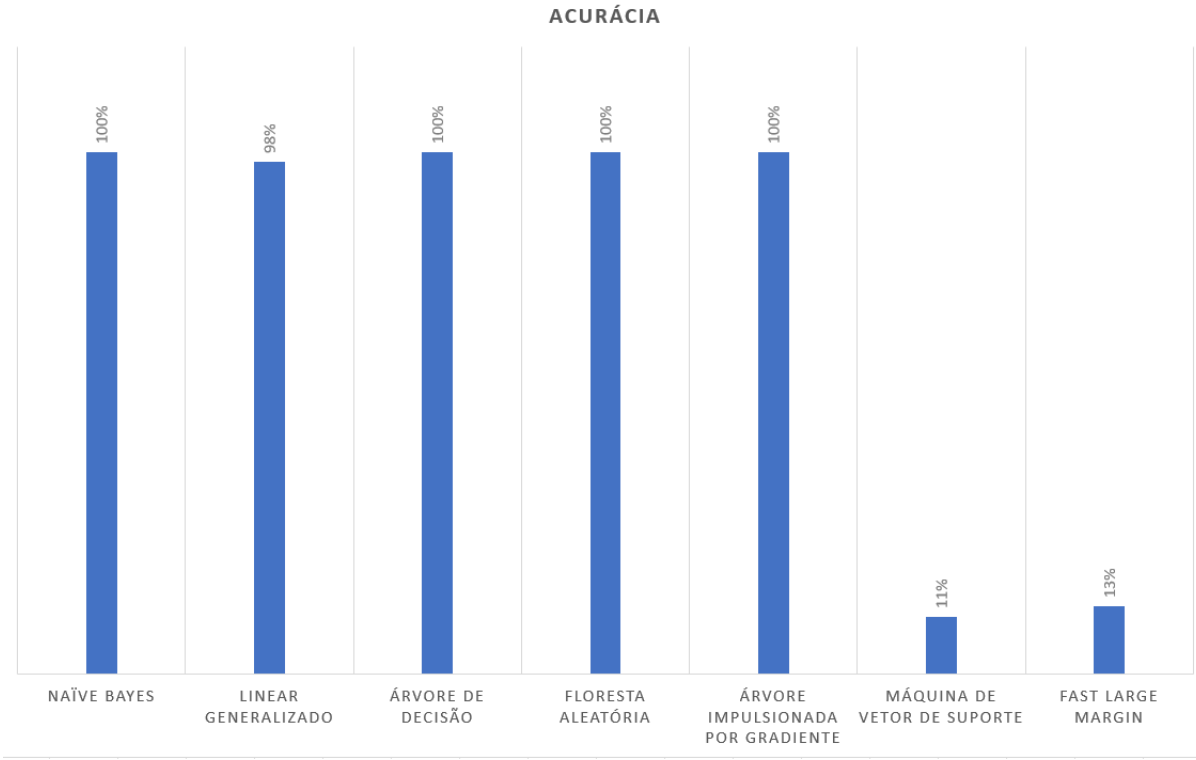

Figura 8 – Acurácia nos Modelos de Classificação de acordo com 20 mil questões

Fonte: A autora

Aplicando a mesma análise a aproximadamente 20 mil questões tendo as disciplinas como parâmetro, foi possível observar que com o aumento da amostra quatro (04) modelos apresentaram altos valores de acurácia.Dentre os modelos com melhor desempenho geral, o Naïve Bayes com 100%, com melhor performance, melhor ganho e melhor tempo total de execução, seguidos dos modelos *Decision Tree*, *Random Forest* e *Gradient Boosted Tree* que também apresentaram 100% de acurácia, porém tiveram valores de ganho e tempos inferiores ao primeiro modelo mencionado. Outro modelo que apresentou bom desempenho foi o *Generalized Linear* com 98% de acurácia. Esse resultado é observado exposto no gráfico da figura 8.

4.2 CLASSIFICAÇÃO DE ASSUNTOS DE QUESTÕES DA ÁREA DE INFORMÁTICA

Aplicando os modelos de classificação tendo como base a busca por assuntos das disciplinas, foi utilizada uma amostra de aproximadamente 4 mil questões da disciplina de Banco de Dados que contém os seguintes assuntos (classes): Diagrama de Entidade e Relacionamento, *SQL Server*, Concorrência em Banco de Dados, Bancos de Dados Paralelos e Distribuídos, Conceitos Básicos em Banco de Dados, Gerência de Transações,

Modelagem de Dados, *Data Warehouse*, Diagramas de Entidade e Relacionamento (DER) e Arquitetura de Banco de Dados, conforme a figura 9.

Assuntos de Banco de Dados	Quantidade de Questões
SQL Server	13.21%
Backup em Banco de Dados	2.48%
DW-Data Warehouse	12.22%
DER- Diagrama de Entidade e Relacionamento	19.72%
Bancos de Dados Paralelos e Distribuídos	3.28%
Arquitetura de Banco de Dados	21,01%
Modelagem de Dados	9,64%
Concorrência em Banco de dados	1,94%
Gatilhos (Triggers)	3,73%
Gerência de Transações	7,25%
Concorrência em Banco de dados	1,94%

Figura 9 – Assuntos da Disciplina de Banco de Dados

Fonte: A autora

Após aplicação dos modelos de classificação ao assunto da disciplina de Banco de Dados, od modelos obtiveram o resultados apresentados na Figura 10.

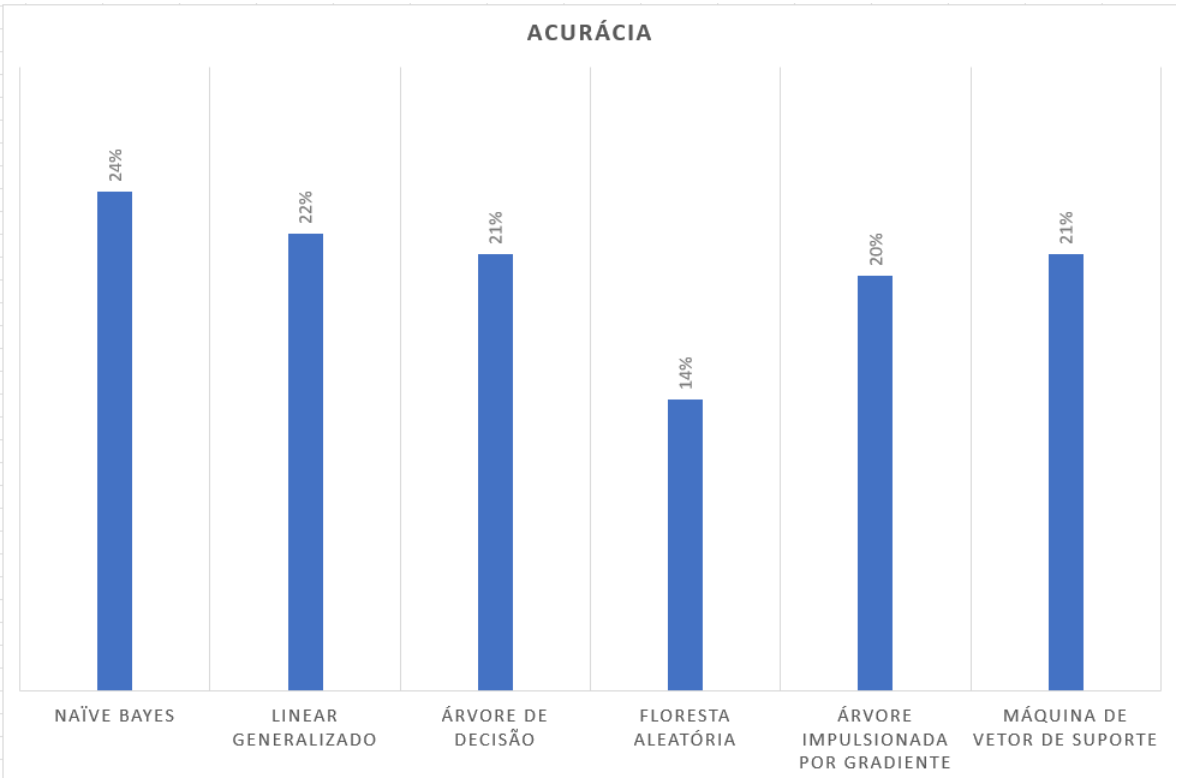

Figura 10 – Modelos de Classificação de acordo com assuntos

Fonte: A autora

Nesse aspecto, dentre os resultados apresentados os modelos Naïve Bayes obteve 24% de acurácia, seguidos de Generalized Linear com 22%, Árvore de Decisão e o MSV com 21%, Gradient Boosted Tree (Árvore Impulsionada por Gradiente) com 20% e por fim

o Floresta Aleatória obteve 14% de acurácia. Esse comportamento pode ter sido decorrente do grande número de classes ou devido à proximidade entre grupos de assuntos (ex: SGBD e Banco de dados Relacionais). É importante ressaltar que, embora uma questão possa ter mais de um assunto, foi considerado apenas o primeiro assunto indicado pelo especialista.

5 CONCLUSÕES E TRABALHOS FUTUROS

São apresentadas neste capítulo as conclusões gerais do trabalho assim como alguns aspectos que podem ser explorados em estudos futuros.

O trabalho apresentado realizou estudo e aplicação das técnicas de mineração de textos com o objetivo de identificar um modelo de classificação automática de pelo menos 20 mil questões de concursos da área de informática. Foram para isso comparados 06 modelos de classificação supervisionada, tendo se destacado entre eles os modelos *Naïve Bayes*, *Generalized Linear*, Árvore de Decisão, Floresta Aleatória e Gradient Boosted Tree, por terem apresentado valores de acurácia superiores a 90

Porém, deve-se considerar novos testes em trabalhos futuros, com ferramentas de computação em nuvem como o *Google Collab* e *softwares* de código aberto como o *Jupyter Notebook*, dentre outros disponíveis na atualidade, pois o grande volume de dados pode causar limitações de *hardware* que podem impactar nos resultados consideravelmente, principalmente na redução do tempo gasto na execução dos modelos. Além disso, executando dessa maneira pode-se considerar também a melhor investigação em outras disciplinas de outras áreas com um volume maior de dados.

Referências

ANDRADE, P. H. M. A. d. Aplicação de técnicas de mineração de textos para classificação de documentos: um estudo da automatização da triagem de denúncias na cgu. 2015. Citado 4 vezes nas páginas 12, 14, 15 e 18.

BARBOSA, I. C. et al. Reconhecimento de mensagens com teor transfóbico no twitter. Universidade Federal de Campina Grande, 2021. Citado 2 vezes nas páginas 15 e 18.

CAMPOS, J. M. R. Avaliação de desempenho e de satisfação do usuário do assistente virtual ifes. talk. Serra, 2022. Citado na página 17.

CARVALHO, D. et al. A machine learning approach to interpolating indoors trajectories. In: SBC. **Anais do IX Symposium on Knowledge Discovery, Mining and Learning**. [S.l.], 2021. p. 145–152. Citado na página 17.

CAVALCANTI, E. R. et al. Detecção e avaliação de cola em provas escolares utilizando mineração de texto: um estudo de caso. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 56, 2011. Citado na página 12.

GUSMÃO, C.; FIGUEIREDO, K.; BRITO, W. A. Técnicas de processamento de linguagem natural em denúncias criminais: Automatização e classificação de texto em português coloquial. In: SBC. **Anais do XLVIII Seminário Integrado de Software e Hardware**. [S.l.], 2021. p. 172–182. Citado na página 16.

HUANG, S. et al. Applications of support vector machine (svm) learning in cancer genomics. **Cancer genomics & proteomics**, International Institute of Anticancer Research, v. 15, n. 1, p. 41–51, 2018. Citado na página 14.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In: SPRINGER. **European conference on machine learning**. [S.l.], 1998. p. 137–142. Citado na página 14.

KANASHIRO, L. H. Aplicação do modelo de regressão logística na identificação das causas que levam a acidentes com vítimas fatais nas rodovias br-101 e br-116. Universidade Federal de São Paulo, 2022. Citado na página 14.

LOPES, R. S. Comparação de métodos de aprendizado de máquina para análise de risco de crédito. Serra, 2022. Citado na página 15.

MAIA, P. P. N. Classificação de descontinuidades em juntas soldadas utilizando máquinas de vetores-suporte treinadas a partir de sinais de ultrassom simulados numericamente. 2019. Citado na página 14.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007. Citado na página 13.

MÜLLER, A. C.; GUIDO, S. **Introduction to machine learning with Python: a guide for data scientists**. [S.l.]: "O'Reilly Media, Inc.", 2016. Citado na página 16.

NETO, F. et al. Computação em nuvem e aprendizado de máquina para análise de grandes volumes de dados educacionais. In: SBC. **Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2020. p. 58–69. Citado na página 16.

OKADA, H. K. R.; NEVES, A. R. N. das; SHITSUKA, R. Análise de algoritmos de indução de árvores de decisão. **Research, Society and Development**, v. 8, n. 11, p. e298111473–e298111473, 2019. Citado na página 16.

OLIVEIRA, L. F.; BOERES, C.; OLIVEIRA, D. de. Definição de parâmetros do spark por meio de aprendizado de máquina: um estudo com dataflows de astronomia. In: SBC. **Anais do XV Brazilian e-Science Workshop**. [S.l.], 2021. p. 25–32. Citado na página 18.

OSHITA, I. T. **Classificação de fake news por mineração de texto**. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2021. Citado na página 18.

PESSANHA. **Random Forest: como funciona um dos algoritmos mais populares de ML**. 2019. Disponível em: <https://medium.com/cinthiabpessanha/random-forest-como-funciona-um-dos-algoritmos-mais-populares-de-ml-cc1b8a58b3b4>. Acesso em: 21 de maio de 2022. Citado 2 vezes nas páginas 16 e 17.

PEZZINI, A. Mineração de textos: conceito, processo e aplicações. **Revista Eletrônica do Alto Vale do Itajaí**, v. 5, n. 8, p. 58–61, 2017. Citado na página 12.

ROZA, B. E.; PEGORARO, M. A. G. Classificador de phishing utilizando algoritmo de naive bayes. 004, 2020. Citado na página 15.

SANTOS, F. L. d. Mineração de opinião em textos opinativos utilizando algoritmos de classificação. 2013. Citado 3 vezes nas páginas 12, 14 e 18.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 34, n. 1, p. 1–47, 2002. Citado na página 17.

SILVA, E. C. M. d.; MEDEIROS, B. A. Comparação de métodos de mineração de texto para classificação de documentos jurídicos. **Ciência da Computação-Tubarão**, 2020. Citado 4 vezes nas páginas 12, 13, 14 e 15.

SIQUEIRA, G. N. Iot aplicada ao monitoramento da saúde de pessoas idosas: um sistema para identificação de quedas. Serra, 2022. Citado na página 17.

VIANA, W. M. O. Comparativo de alguns modelos de machine learning utilizando dados de domínio público e a linguagem python. Universidade Estadual Paulista (UNESP), 2021. Citado 3 vezes nas páginas 12, 17 e 18.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Restrito

Entrega de trabalho de conclusão de curso

Assunto: Entrega de trabalho de conclusão de curso
Assinado por: Elayne Regina
Tipo do Documento: Projeto
Situação: Finalizado
Nível de Acesso: Restrito
Hipótese Legal: Informação Pessoal (Art. 31 da Lei no 12.527/2011)
Tipo do Conferência: Cópia Simples

Documento assinado eletronicamente por:

- **Elayne Regina Lima Silva**, **ALUNO (201811210038) DE TECNOLOGIA EM TELEMÁTICA - CAMPINA GRANDE**, em 26/09/2022 12:37:59.

Este documento foi armazenado no SUAP em 26/09/2022. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 634714
Código de Autenticação: b48ec840cd