UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
DEPARTAMENTO DE ESTATÍSTICA E INFORMÁTICA
PROGRAMA DE PÓS–GRADUAÇÃO EM INFORMÁTICA APLICADA

BRUNO CLAUDINO PEREIRA DE BRITO

EVASÃO UNIVERSITÁRIA: APLICAÇÃO DE MINERAÇÃO DE DADOS EDUCACIONAIS PARA IDENTIFICAÇÃO DE ATRIBUTOS RELEVANTES

RECIFE – PE

2022

BRUNO CLAUDINO PEREIRA DE BRITO

EVASÃO UNIVERSITÁRIA: APLICAÇÃO DE MINERAÇÃO DE DADOS EDUCACIONAIS PARA IDENTIFICAÇÃO DE ATRIBUTOS RELEVANTES

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática - DEINFO - Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre.

ORIENTADOR: Rafael Ferreira Leite de Mello

COORIENTADOR: Gabriel Alves de Albuquerque Junior

RECIFE – PE

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

B862e Brito, Bruno Claudino Pereira de Brito
EVASÃO UNIVERSITÁRIA: APLICAÇÃO DE MINERAÇÃO DE DADOS EDUCACIONAIS PARA IDENTIFICAÇÃO DE ATRIBUTOS RELEVANTES / Bruno Claudino Pereira de Brito Brito. - 2022.
65 f. : il.

Orientador: Rafael Ferreira de Mello.
Coorientador: Gabriel Alves de Albuquerque Junior.
Inclui referências.

Dissertação (Mestrado) - Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, 2022.

1. Mineração de Dados Educacionais. 2. Random Forest. 3. Árvore de Decisão,. 4. Evasão. I. Mello, Rafael Ferreira de, orient. II. Junior, Gabriel Alves de Albuquerque, coorient. III. Título

CDD 004

BRUNO CLAUDINO PEREIRA DE BRITO

EVASÃO UNIVERSITÁRIA: APLICAÇÃO DE MINERAÇÃO DE DADOS EDUCACIONAIS PARA IDENTIFICAÇÃO DE ATRIBUTOS RELEVANTES

Dissertação submetida à Coordenação do Programa de Pós-Graduação em Informática Aplicada do Departamento de Estatística e Informática - DEINFO - Universidade Federal Rural de Pernambuco, como parte dos requisitos necessários para obtenção do grau de Mestre.

Aprovada em: 15 de Fevereiro de 2022.

BANCA EXAMINADORA

Rafael Ferreira Leite de Mello (Orientador)
Universidade Federal Rural de Pernambuco
Departamento de Computação

Roberta Macedo Marques Gouveia
Universidade Federal Rural de Pernambuco
Departamento de Computação

Cristian Cechinel
Universidade Federal de Santa Catarina
Departamento de Computação

Aos meus pais, meu irmão, meus amigos, meu orientador, e a todos que mesmo de forma indireta me apoiaram e acreditaram em mim, renovando minhas forças para superar as dificuldades.

Agradecimentos

Meu agradecimento ao meu orientador Prof. Dr. Rafael Ferreira, que me acompanhou durante todo o período prestando dicas, informações e feedbacks importantes, se mostrando prestativo e paciente. Sua atuação para mim foi de grande valia.

Aos meus pais Diene Claudino e Armando Brito pela criação e dedicação que me proporcionaram solidez durante os anos iniciais de formação, de forma a criar bases para que pudesse chegar a esse momento. Ao meu irmão que sempre incentivou a leitura e também nunca parou de estudar.

Aos amigos que torceram por mim, em especial os mais próximos que sempre estiveram disponíveis para ajudar e aconselhar: Johny Nunes, Gisele Teresa, Camila Pessoa.

Por último, agradeço aos que, de alguma forma, mesmo indireta, contribuíram para a construção desse trabalho.

O homem não teria alcançado o possível se, repetidas vezes, não tivesse tentado o impossível.

(Max Weber)

Resumo

A evasão está entre um dos maiores problemas pelas quais as universidades enfrentam. Esse problema afeta diversos aspectos do funcionamento universitário assim como possui diversas causas e consequências. Aliado a isso, as universidades armazenam e produzem cotidianamente uma grande de dados de seus alunos. À vista disso, esse trabalho tem como objetivo ajudar os gestores universitários a identificar as principais causas encontradas nos dados de forma que ajude a elaborar estratégias que evitem a evasão. Especificamente, nesse trabalho foram utilizadas técnicas de mineração de dados educacionais para a construção de um modelo que classificasse o aluno com sua possibilidade de evasão. Em seguida, esse modelo foi avaliado pela sua eficácia para numa segunda etapa identificar as principais características encontradas que determinem a evasão. Esse trabalho foi realizado basicamente em três conjuntos de dados, sendo o primeiro o conjunto de dados da UFRPE o qual foi associado ao segundo conjunto, o censo demográfico de 2010, formando o primeiro estudo. Os dados da UFRPE também foram associados ao terceiro conjunto de dados, informações de notas do SISU, caracterizando assim, o segundo estudo. Ambos foram submetidos ao algoritmo classificador baseado em árvore de decisão, Random Forest, e em seguida coletados os valores e discutidos seus resultados. Os resultados mostram a viabilidade do algoritmo classificador para a pesquisa proposta. Além disso, foram disponibilizados os principais fatores nos dados que determinam a evasão segundo o classificador, mostrando que certas características devem possuir maior atenção no momento de elaborar estratégias que mitiguem a evasão.

Palavras-chave: Mineração de Dados Educacionais, Random Forest, Árvore de Decisão, Evasão.

Abstract

Dropping out is among the biggest problems universities face. This problem affects several aspects of university functioning as well as it has various causes and consequences. Allied to this, universities store and produce daily a multitude of data from its students. Therefore, this work has as a goal to help university managers to identify the main causes found in the data in a way that helps devise strategies to avoid evasion. Specifically, in this work educational data mining techniques were used for the construction of a model to classify the student with his/her dropout possibility. In then, this model was evaluated for its effectiveness to identify in a second step the main characteristics found that determine the evasion. This job was carried out basically on two sets of data, one from the demographic census of 2010 and another set of data using SISU information, thus, divided in two studies. Both were submitted to the tree-based classifier algorithm decision, Random Forest, and then collected the values and discussed their results. The results show the viability of the classifier algorithm for the proposed research. In addition, the main factors in the data that determine evasion were made available according to the classifier, showing that certain characteristics should have greater attention at the time of devising strategies to mitigate evasion.

Keywords: Data Mining, Random Forest, Decision Tree, Student Dropout.

Lista de Figuras

Lista de tabelas

Lista de Siglas

SVM	*Support Vector Machine*
MDA	*Mean Decrease in Accuracy*
MDG	*Mean Decrease in Gini*
IDH	Índice de Desenvolvimento Humano
ENEM	Exame Nacional do Ensino Médio
SISU	Sistema de Seleção Unificada
UFRPE	Universidade Federal Rural de Pernambuco
IDHM	Índice de Desenvolvimento Humano Municipal
FECTOT	Taxa de Fecundidade Total
ESPVIDA	Esperança de Vida ao Nascer
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico

Sumário

1 Introdução

Entre alguns dos problemas enfrentados pela educação está o problema da evasão. A evasão ocorre em todos os níveis de ensino, desde a educação infantil até o ensino superior (LOBO, 2012) e em cada nível de ensino traz problemas graves (FILHO et al., 2013). Para o ensino superior, nível de ensino em que ocorre uma educação técnica e voltada ao mercado de trabalho, esse problema traz graves entraves ao desenvolvimento local e social, os quais já foram debatidos por diversos autores (LUCAS, 1988; BARRO, 1991; MANKIW et al., 1992). Todavia, antes de discorrer sobre evasão, é necessário uma visão geral do que significa esse termo.

A palavra evasão pode ser entendida formalmente por qualquer dicionário da língua portuguesa como um processo ou um ato de fugir.[1] Contudo, pode ter conotações diferentes a depender do contexto em que é usado. Quando se refere ao ambiente universitário, possui um significado mais específico. Para alguns autores, a evasão é definida como a perda do vínculo, a saída ou o abandono do aluno no curso, seja intencional ou não (GARCIA; SANTIAGO, 2015; DURSO; CUNHA, 2018). Essa definição segue alinhada ao uma definição também usada por órgãos do governo na definição de evasão universitária. Para o MEC, tecnicamente, a evasão compreende o desligamento do aluno do sistema de ensino ao qual se encontrava (MEC, 1998).

Esse desligamento provoca algumas consequências, que para as universidades públicas envolve, entre outras consequências, a falta de retorno do investimento uma vez que a universidade não consegue cumprir seu objetivo de formar alunos (FILHO et al., 2013). Em relação às universidades privadas, envolve a falta de capital uma vez que tais alunos são responsáveis diretamente pelo orçamento dessa universidade (FILHO et al., 2013). Consequentemente, a evasão tem relação direta com a gestão universitária, uma vez que ela tem como missão garantir ou assegurar a formação adequada do aluno.

Alguns fatores podem estar associados a esse processo de evasão. Pode-se citar como exemplo de fatores associados ao aluno a idade, reprovações, renda ou até mesmo podem ser considerados fatores as características da instituição como capacitação docente ou quantidade de laboratórios. Diante disso, alguns autores organizam esses fatores em grupos. Para alguns autores, esses fatores se resumem em apenas dois conjuntos, fatores

[1] https://michaelis.uol.com.br/, https://www.dicio.com.br/evasao/

internos e externos (DIAS et al., 2010), já para outros pesquisadores, os fatores se agrupam em 3 conjuntos: externos à instituição, internos à instituição e um terceiro que são fatores individuais ao estudante (CHAYM, 2019). Independente dos fatores associados, eles são bastante explorados em busca de soluções que possam reverter a evasão, assim como elaborar estratégias de manutenção do aluno. Entre algumas técnicas computacionais está a Mineração de Dados Educacionais.

As universidades possuem uma coleção enorme de dados institucionais dos alunos, desde notas, endereço, desempenhos, processo seletivo, assim como informações dos alunos com maior vulnerabilidade social. Esses dados vem sendo ultimamente explorados a fim de obter conhecimento ainda não visualizado por esses gestores (BACH; ALESSA, 2014; SCHUHA et al., 2019). Diante desse contexto, a disciplina de mineração de dados educacionais possui técnicas capazes de extrair conhecimento dessas bases de dados e aplicá-las aos problemas relacionado à evasão (BACH; ALESSA, 2014; SCHUHA et al., 2019). Assim, essa pesquisa se insere nesse campo de atuação, em que se busca conhecimento dentro dessas bases de dados em busca de estratégias que evitem o abandono do aluno. Para isso, é preciso conhecer as associações e padrões entre os dados da universidade e como elas se relacionam com outras bases governamentais.

Uma vez delimitada a temática da pesquisa, foi então selecionada os dados para serem analisados, sendo relatada a importância desses dados, e então descrito o algoritmo utilizado e sua eficácia para os resultados e ao fim, obtidos e discutidos os resultados por uma metodologia.

1.1 Problema de Pesquisa

Na mineração de dados educacionais, os dados estudados assumem relevância uma vez que a técnica utilizada pode não trazer grandes informações quando não há dados variados a serem analisados. Nessa pesquisa, tentou-se analisar os dados disponibilizados pela universidade através do seu sistema acadêmico de controle e registro dos alunos e matrículas, junto com demais dados de outras bases públicas em busca de fatores que determinam a evasão. Esses dados incluem características associadas ao aluno como cor da pele e data de nascimento bem como fatores externos ao aluno ou à universidade como índices do IDH de sua cidade natal. Esses fatores, uma vez identificados, podem auxiliar

os gestores universitários a elaborar estratégias que visam a mitigação da evasão em busca de melhores índices na gestão acadêmica.

Para a pesquisa apresentada, foram selecionados duas bases públicas: a do último censo brasileiro produzido em 2010 e a base de notas dos alunos no Sistemas de Seleção Unificada (SISU). Para ambas as bases, a pesquisa teve objetivos gerais e específicos em comuns e metodologia semelhante. Foi avaliado, então, quais fatores de evasão podem estar relacionados quando essas informações dos alunos estão associadas com essas duas bases públicas. Dessa forma, a primeira pergunta de pesquisa foi:

PERGUNTA DE PESQUISA 1:

Qual a eficácia do classificador construído na identificação de alunos com risco de evasão utilizando bases de dados públicas e internas da instituição?

A pergunta acima tem o objetivo determinar a eficácia da predição do classificador ao avaliar um aluno e sua possibilidade de se evadir baseando-se nos dados utilizados. Dessa forma, foi medido por meio da acurácia do classificador uma mensuração numérica da predição da evasão baseado nos dados pesquisados em ambas as bases. Uma vez atingida a eficácia da classificação, foi observado se esses resultados diferem entre as diferentes áreas do conhecimento. Dessa forma, surgiu a segunda pergunta:

PERGUNTA DE PESQUISA 2:

Existe diferença de resultados quando considerados dados de diferentes áreas do conhecimentos

Essa segunda pergunta avaliou a eficiência do classificador e as características mais importantes de acordo com áreas de conhecimento. Os valores apresentados são importantes para dar uma noção de discrepância entre o mais importante e o menos importante para cada área do conhecimento e situá-los dentro do contexto da universidade pesquisada. Por fim, a última pergunta de pesquisa é:

PERGUNTA DE PESQUISA 3:

Quais são as características nos dados que influenciam a evasão em cada base de dados disponibilizada?

A pergunta acima tem como objetivo determinar, à luz de uma técnica formal de mineração de dados, quais características influenciam a evasão dentro daquele conjunto de dados. Essa pergunta foi usada para ambas as bases estudadas. Com isso, foi possível verificar quais fatores são mais influentes na determinação da evasão.

1.2 Objetivos

A seguir são apresentados o objetivo geral e os objetivos específicos que nortearam a condução dessa pesquisa. O objetivo geral define o propósito do estudo e os específicos caracterizam as etapas do projeto.

1.2.1 Objetivo Geral

A presente pesquisa teve como objetivo geral determinar os fatores que levam à evasão, dentro do conjunto de dados estabelecido, assim como ajudar os gestores universitários a identificar quais fatores são mais importantes entre eles.

1.2.2 Objetivos Específicos

Para atingir o objetivo geral foram definidos os seguintes objetivos específicos:

- Coletar a base acadêmica e as bases públicas a fim de associá-las.
- Aplicar o algoritmo Floresta Aleatória com vistas a extrair a informação desejada em ambas as bases.
- Extrair as características de maior e menor importância desses conjuntos de dados segundo o algoritmo.
- Coletar os resultados e disponibiliza-los em gráficos.
- Discutir os principais pontos de destaque encontrados nas características mais importantes.

1.3 Organização do Trabalho

Esse trabalho está organizado em 6 capítulos, de forma que o próximo capítulo, capítulo 2, apresenta fundamentação teórica sobre os tópicos abordados, o algoritmo utilizado, os termos utilizados durante a pesquisa e o contexto computacional em que é usado. Explicações sobre Classificação, Previsão e Random Forest são encontrados nesse capítulo. O capítulo 3 relata os trabalhos relacionados ao tema da pesquisa, descrevendo os trabalhos e suas limitações. O capítulo 4 relata a metodologia utilizada, a coleta dos dados e como eles foram tratados e utilizados para o algoritmo. O capítulo 5 relata os resultados, as discussões e conclusões da pesquisa, e por fim, o capitulo 6 retoma a visão geral da pesquisa abordando também suas limitações e novas perspectivas de estudo.

2 Fundamentação Teórica

Esse capítulo discorre sobre a teoria necessária para a compreensão dos termos e metodologias utilizadas na pesquisa. Está organizado numa sequência da teoria geral para a específica, sendo organizado em 4 seções.

2.1 Mineração de Dados

Mineração de Dados compreende modelos computacionais e estatísticos em grandes massas de dados a fim de obter conhecimento (CORADINE et al., 2011; COELHO, 2006). A mineração de dados nessas grandes bases de dados pode ser conhecida como um processo de procura de associação entre variáveis ou um processo de procura de um subconjunto de dados com determinado padrão. No ramo da administração, é conhecido como uma ciência que busca encontrar padrões ou relacionamentos nos dados em busca de melhores decisões estratégicas ou vantagens competitivas.

Mineração de Dados é uma área interdisciplinar. Entre as áreas compreendidas, a de **Estatística** é uma das mais utilizadas pela mineração. A criação de modelos estatísticos usando conceitos de dispersão, distribuição e probabilidade, além da variância são um dos principais componentes utilizados nas análises dos dados. Uma outra área extensivamente usada é a área de **Aprendizagem de Máquina**. Essa área estuda como um computador pode melhorar seu desempenho baseado nos dados que recebe (MITCHELL, 1997). Basicamente procura realizar decisões inteligentes reconhecendo padrões nos dados. As técnicas de aprendizagem de máquina possuem as seguintes abordagens (HAN; KAMBER, 2006):

1. **Aprendizado supervisionado** - O algoritmo aprende a solucionar o problema baseado nos dados já previamente rotulados do conjunto de treinamento, ou seja, os dados possuem um atributo "classe". O aprendizado supervisionado está em construir um modelo que rotule novos dados que não possuem o atributo classe associado (HAN; KAMBER, 2006; BRAMER, 2013). Algoritmos dessa abordagem necessitam sempre de um conjunto de dados de treinamento e um conjunto de dados para teste (GOLDSCHMIDT, 2006).
2. **Aprendizado não supervisionado** - O algoritmo não possui saída desejada, dessa

forma ele aprende a rotular os dados procurando semelhanças entre eles (HAN; KAMBER, 2006; BRAMER, 2013). Como resultado, formam-se agrupamentos que representarão as classes ou rótulos. As regras de associação também estão inseridas dentro do aprendizado não supervisionado.

3. **Aprendizado semi-supervisionado** - Nessa abordagem o algoritmo utilizado um conjunto de dados rotulados e não rotulados para construir o modelo. Nessa abordagem, alguns dados rotulados são usados para construir o modelo e os dados não rotulados são usados para definir e refinar as fronteiras entre os rótulos (SANCHES, 2003; HAN; KAMBER, 2006).
4. **Aprendizado ativo** - Nessa abordagem, o algoritmo interage com o usuário no processo de aprendizado. Nesse caso, o algoritmo pode pedir para o usuário classificar um conjunto de dados produzidos pelo algoritmo. A ideia é aumentar a qualidade do modelo adquirindo conhecimento do usuário (SETTLES, 2009).

.

Quando se trata do processo de mineração, várias técnicas e estratégias podem ser aplicadas com vistas a um determinado problema. Essas técnicas podem estar associadas a determinados algoritmos que tenham em comum a mesma forma de atuação. Nisso, existem as funcionalidades as quais os problemas se inserem. De maneira geral, a mineração de dados possui duas funcionalidades: a Descritiva e a Preditiva, e para cada uma das funcionalidades, existem suas principais técnicas e algoritmos como mostra a Figura 1.

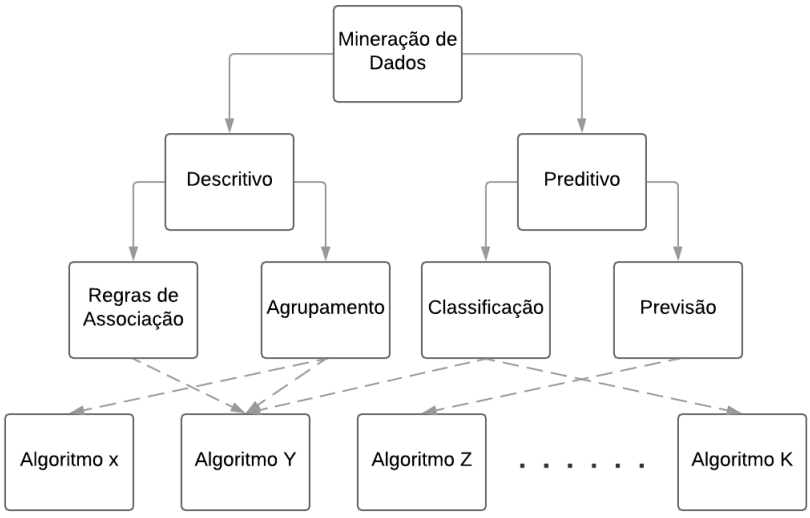


Figura 1 – Divisão das funcionalidades e técnicas da Mineração. Fonte: o autor (2022)

A funcionalidade Descritiva está relacionada à caracterização dos dados que

estão sendo analisados, enquanto na funcionalidade Preditiva está relacionado com a generalização (indução) presentes nos dados para criação de modelos que possam ser usados em dados não observados (COELHO, 2006). A Figura 1 apresenta uma estrutura hierárquica das técnicas, contudo, um mesmo problema pode ser resolvido tanto de maneira preditiva quanto descritiva. Por exemplo, um algoritmo de Redes Neurais pode ser usado tanto na Classificação, cuja função pode ser Preditiva, quanto no agrupamento, cuja função é Descritiva. Assim, as principais técnicas podem ser descritas abaixo.

1. **Regras de Associação** - As regras de associação tem por objetivo encontrar regras de associação entre itens de um determinado subconjunto. De maneira geral, indica que a presença de um item num conjunto pode indicar a presenta de outro item na mesma operação. Essa regra foi utilizada inicialmente na Análise de Cesta de Mercado (*Market Basket Analysis*) (GOLDSCHMIDT, 2006; HAN; KAMBER, 2006; BRAMER, 2013), em que se analisa o comportamento de compra e venda de itens analisando o histórico de vendas. Se baseia na condição Se X Então Y, e estabelece que se um comprador comprar um conjunto de itens X, então deve comprar o conjunto de itens Y. Essa regra possui algumas medidas de qualidade e ficou conhecido como algoritmo Apriori (AGRAWAL; SRIKANT, 1994). Esse algoritmo é utilizado até hoje e foi estendido para outras situações de análise de co-ocorrências, atuando além do seu uso de cestas de mercado.
2. **Agrupamento** - Consiste em agrupar dados minimizando as diferenças dentro do grupo e maximizando as diferenças entre os grupos. O objetivo se dá em agrupar os dados dentro de conjuntos e que esses dados possuam similaridade entre si dentro do conjunto. A determinação de similaridade e diferença são obtidas a partir da acumulação de medidas de todos os elementos do grupo (COELHO, 2006). Geralmente essas medidas são baseadas na distância numérica entre os elementos e podem ser calculadas mediante um tratamento prévio dos dados categóricos. O Agrupamento se dissocia da Classificação na medida em que a Classificação trabalha com rótulos já definidos, ao passo que no Agrupamento, há a necessidade de identificar rótulos, dessa forma ocorre uma indução não supervisionada. Podem ser usadas técnicas hierárquicas, que divide os dados de maneira hierárquica e técnicas aglomerativas como o algoritmo *k-means* (KAUFMAN; ROUSSEEUW, 2009). Nessa seção pode ser citado o estudo dos outliers, isto é, identificação de dados ou elementos

discrepantes dos conjuntos normalmente registrados.

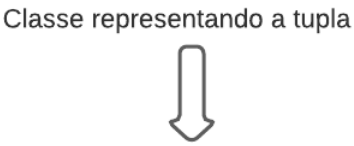
3. **Classificação e Regressão** - Embora tenham resultados distintos, ambos possuem significado parecido, e de certa forma, funções semelhantes. Classificação se entende como a construção de uma função (modelo) que determine um conjunto de registros em um conjunto de rótulos categóricos já pré-definidos. A tarefe então se dá em determinar se determinado item ou dado pertence a uma classe já definida. Na Regressão, a função (modelo) mapeia os registros para determinado valor numérico real. Em resumo, na Classificação, são criados modelos que retornam variáveis categóricas (valores em um conjunto finito, sem ordenação natural, e suficiente pequeno), ao passo que na Regressão são retornados do modelo valores numéricos (conjunto numérico, ordenado e potencialmente infinito) (COELHO, 2006; HAN; KAMBER, 2006). Ambas possuem aplicações também semelhantes. Na Classificação, pode-se, por exemplo, obter se determinado cliente do banco se caracteriza como bom ou mal investidor. Já usando-se a Regressão, pode-se determinar o risco do banco ao emprestar dinheiro a tal cliente. Algoritmos como Redes Neurais, Árvores de Decisão, Redes Baeysianas se enquadram na resolução de problemas de Classificação e Regressão.

Considerando os conceitos apresentados acima, a pesquisa se situa na construção de um modelo Preditivo, utilizando a funcionalidade de Classificação. Por tal motivo, foi dado ênfase nesse tópico.

2.1.1 Etapas da Classificação

A classificação pode ser realizada em duas etapas (HAN; KAMBER, 2006; BRAMER, 2013). A primeira etapa, também chamada de etapa de treinamento, consiste na etapa de aprendizagem em si, em que o algoritmo classificador é construído "aprendendo" num conjunto de tuplas, sendo essas tuplas com n dimensões de atributos, e cada tupla pertencendo a uma determinada classe de atributo ou rótulo de classe, o qual o algoritmo classificador deve aprender. Esse atributo é nominal e serve como a categoria ou a classe. Devido ao número já definido de classes e a predefinição delas, assim como sua ordenação não tem importância, esse passo é uma aprendizagem supervisionada (HAN; KAMBER, 2006). A representação do atributo de classe pode ser visto na Figura 2. Nessa primeira

etapa, chama-se dados de treinamento (*training data*) o conjunto de dados utilizado para a construção do classificador.



Idade	Curso	Área	Estado Civil	Status
24	Agronomia	Agrárias	Solteiro	Evadido
19	Biologia	Biológicas	Solteiro	Evadido
33	Computação	Exatas	Casado	Não Evadido
25	Matemática	Exatas	Divorciado	Não Evadido
24	História	Sociais	Casado	Não Evadido

Figura 2 – Representação de um atributo de classe. Fonte: o autor (2022)

O segundo passo consiste em utilizar o já construído classificador em dados por ele desconhecidos, também chamados de dados de teste. Os resultados devem ser avaliados por medidas de qualidade, e tendo bom desempenho, o classificador pode ser usado em outros dados ou outras aplicações que possuam os mesmos tipos de dados.

A fim de obter uma qualidade no classificador, os dados precisam estar particionados de tal forma que o classificador aprenda o suficiente para classificar novos dados, generalizar a decisão, assim como gerar predições confiáveis. Sendo assim, o particionamento dos dados entre treinamento e teste assume relevância.

O particionamento que utiliza a abordagem **Holdout** tem em vista a construção de um único modelo que possa ser usado em diversos dados. O particionamento Holdout, como relata Goldschmidt (2006), particiona o conjunto de dados da seguinte forma: seja um conjunto de dados, os dados de treinamento possuem uma porcentagem fixa P, sendo $P > 1/2$, restando $(1 - P)$ aos dados de teste. Geralmente é utilizado $P = 2/3$, restando $(1 - P) = 1/3$ aos dados de teste. Não há, ainda, um consenso de qual proporção ideal utilizar para treinamento nessa abordagem, sendo a acurácia um indicador importante da efetividade do particionamento (GOLDSCHMIDT, 2006).

2.1.2 Medidas de Qualidade da Classificação

Um classificador só pode ser considerado adequado se seus resultados possuem um nível de acerto próximo ao real, ou seja, que ocorra a correta classificação nos dados que são ainda desconhecidos pelo classificador. Existem diversas formas de avaliar a qualidade do resultado de um classificador. Existem métricas que envolvem desde a complexidade da solução apresentada pelo classificador, passando pelo tempo de resposta para a classificação ou a escalabilidade do classificador para ser usado em grandes dados. Os métodos mais utilizados correspondem aos que verificam a correta classificação do algoritmo quando aplicado aos dados de teste. Entre os principais métodos avaliativos, podemos destacar a Acurácia, A Precisão, o Revocação, Medida-F, Coeficiente Kappa, entre outros (BRAMER, 2013; GOLDSCHMIDT, 2006).

2.1.2.1 Acurácia

De maneira geral, os resultados da classificação podem ser agrupados numa matriz de confusão (HAN; KAMBER, 2006) conforme a Tabela 1 a seguir.

Tabela 1 – Matriz de confusão da acurácia. Fonte: o autor (2022)

		Predição da Classe	
		Positivo	Negativo
Classe Real	Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Para melhor entendimento da tabela acima, considere um exemplo de uma listagem de alunos com duas classes, a positiva indicando a não evasão, e a negativa indicando a evasão. A Tabela 2 a seguir representaria a matriz de confusão dessas classes do exemplo descrito.

Tabela 2 – Exemplo de matriz de confusão da acurácia. Fonte: Kamber (2006)

		Predição da Classe		
		Positivo	Negativo	Total
Classe Real	Positivo	6954	46	7000
	Negativo	7366	2634	3000
	Total	7366	2634	

A acurácia se refere à quantidade da correta classificação dos registros (VP + VN) sob o total de registros preditos pelo classificador (VP + FN + FP + VN). Em síntese, é possível observar que quanto maior os valores de VP e VN, mais a acurácia tende a ser maior e o classificador tende a classificar corretamente. A acurácia é geralmente utilizada como medida quando os dados estão relativamente balanceados (GOLDSCHMIDT, 2006; HAN; KAMBER, 2006).

A fórmula da acurácia é definida da seguinte maneira.

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

2.1.2.2 Coeficiente Kappa

O coeficiente Kappa foi introduzido como uma técnica de confiabilidade por Jacob Cohen em 1960 (COHEN, 1960). De maneira resumida, representa a taxa de concordância entre dois ou mais avaliadores, assim como permite avaliar se a concordância está além do simples acaso (SILVA; PAES, 2012). No ambiente de mineração de dados, tem a tarefa de analisar se a classificação correta das classes supera a classificação aleatória. Possui a seguinte fórmula (COHEN, 1960; SILVA; PAES, 2012):

$$k = \frac{P(a) - P(e)}{1 - P(e)} \tag{2.2}$$

Em que P(a) representa a proporção de concordância observada na classificação, e P(e) representa a probabilidade hipotética de chances de concordância. Pela fórmula do coeficiente Kappa, obtêm-se um resultado entre -1 e 1. Segundo Cohen (COHEN, 1960; MCHUGH, 2012), valores menores que 0 indicam nenhuma concordância, valores até 0,2 indicam concordância pobre, valores entre 0,21 e 0,4 indicam leve concordância,

valores entre 0,41 e 0,6 indicam concordância moderada, 0,61 até 0,8 indicam concordância substancial, e 0,81 em diante indicam concordância quase perfeita.

Deve-se levar em consideração que o coeficiente Kappa é uma medida conservadora e que podem ser usados pesos em suas avaliações de concordância ou discordância. Por ser uma medida conservadora, é bastante utilizada na área de saúde e diagnósticos devido aos sérios problemas causados por uma classificação incorreta (MCHUGH, 2012).

2.2 Algoritmos de Classificação

Como citado nas seções acima, as técnicas de classificação envolvem rotular um dado em determinada classe já conhecida. Consequentemente, existem alguns algoritmos que realizam essa tarefa de classificação. Todavia, determinados algoritmos utilizam abordagens diferentes para classificar seus dados. Podemos citar alguns desses algoritmos como as Redes Bayesianas, Máquinas de Vetores de Suporte (SVM), Árvores de Decisão, K-vizinhos mais próximos, entre outros. Cada algoritmo citado possui sua especifidade na técnica de classificação. Por exemplo, algoritmos como redes bayesianas são usados quando todos os atributos são categóricos, já para k-vizinhos mais próximos são usados quando todos os atributos são contínuos (GOLDSCHMIDT, 2006).

Em classificações bayesianas, cada valor de um atributo dentro de uma classe é independente do valor de outros atributos (HAN; KAMBER, 2006). Isso que dizer que um determinado valor de uma classe não está relacionado a qualquer outro valor, ou seja, ele é independente. Esse método não usa regras de decisão e sim em probabilidade, tendo como base o teorema de probabilidade de Bayes (BRAMER, 2013).

Para o algoritmo K-vizinhos mais próximos, ele assume a classificação para uma instância desconhecida como a classificação de uma instância ou instâncias mais próximas a ela (BRAMER, 2013). Em resumo, cada tupla representa um ponto num espaço de n dimensões, sendo n a quantidade de atributos. Dessa forma, as tuplas de treinamento são armazenadas num espaço n dimensional padrão. Assim, para uma determinada tupla desconhecida, o algoritmo procura no espaço n dimensional padrão uma tupla próxima a ela (HAN; KAMBER, 2006).

Máquinas de Vetores de Suporte é um método de regressão que também pode ser usado em classificação. Se baseia na construção de um hiperplano para separar as

classes desejadas (HAN; KAMBER, 2006). Por exemplo, tendo 2 classes, o espaço será bidimensional, e tendo a possibilidade das classes serem totalmente separáveis no espaço bidimensional, o hiperplano será uma reta. Contudo, em aplicações complexas com mais classes, o hiperplano pode ter outras formas. Possui a desvantagem de ser lento em grandes quantidades de dados (HAN; KAMBER, 2006; BRAMER, 2013).

Por último, podemos citar as redes neurais nos problemas de classificação. Em suma, é um conjunto de unidades de entrada e saída que estão conectadas entre si e geralmente em camadas, em que cada conexão possui um peso associado (HAN; KAMBER, 2006). Durante a fase de aprendizado, a rede ajusta seus pesos para que ocorra a correta classificação de determinada tupla. Redes neurais podem ser usadas em diversas aplicações e possuem diversas particularidades, entretanto, possuem desvantagens no uso na mineração de dados devido à dificuldade de interpretação dos significados dos seus pesos durante a fase de aprendizado (HAN; KAMBER, 2006).

Tendo alguns dos algoritmos de classificação resumidos acima, e tendo a presente pesquisa utilizado o algoritmo de classificação Random Forest, foi dado ênfase, então, ao algoritmo de classificação baseado em árvore de decisão explicitado melhor a seguir.

2.3 Árvores de Decisão

Árvores de decisão são técnicas comumente usadas para a classificação. Uma árvore de decisão é um grafo onde cada nó representa uma condição sobre determinado atributo, e que o melhor atributo é escolhido para ser o nó divisor, sendo esse atributo escolhido através de algum mecanismo de seleção de atributos (HAN; KAMBER, 2006). Idealmente cada nó representa uma decisão a ser tomada num conjunto de valores de determinado atributo, e assim recursivamente até o nó terminal, gerando uma árvore. Do caminho do nó raiz até o nó folha, formam-se regras sobre os nós do tipo SE $<$*condição*$>$ ENTÂO $<$*caminho*$>$, sendo o conjunto dessas regras chamado de regras de decisão. Ao final, tem-se a classificação de determinado registro baseado nessas regras (GOLDSCHMIDT, 2006). Considere o conjunto de dados da Tabela 3, de uma lista alunos com seus atributos e sua classe.

Tabela 3 – Registros de alunos e sua classe. Fonte: o autor (2022)

Idade	Área	Estado Civil	Classe
24	Exatas	Divorciado	Não Evadido
21	Humanas	Casado	Evadido
23	Exatas	Solteiro	Não Evadido
25	Humanas	Solteiro	Evadido
28	Exatas	Casado	Evadido
31	Exatas	Divorciado	Não Evadido

A partir desses dados, uma simples árvore de decisão pode ser formada, sendo possível verificar visualmente uma simplificação dessa árvore pela Figura 3.

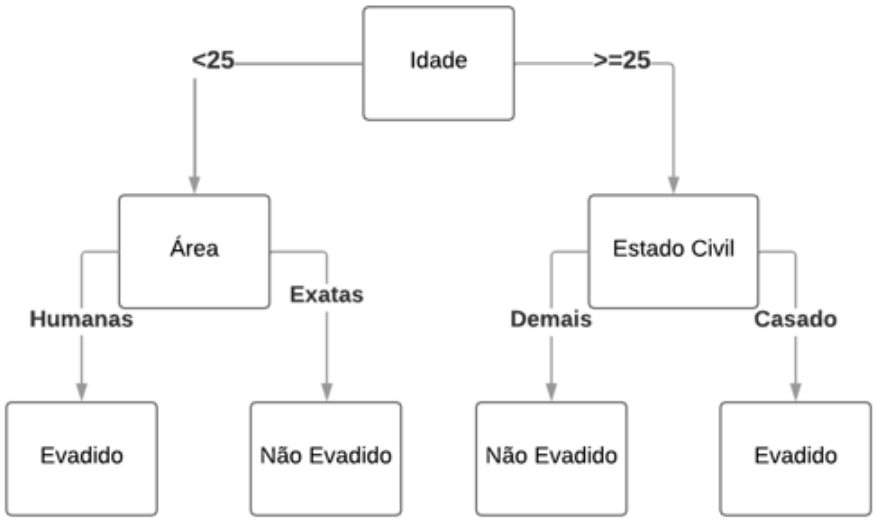


Figura 3 – Exemplo de árvore de decisão para a Tabela 3. Fonte: o autor (2022)

Entre alguns dos algoritmos baseados em árvores de decisão estão os algoritmos ID3, C4.5 e CART, sendo este último um algoritmo baseado em árvore binária, porém todos usam tuplas de treinamento numa abordagem top-down, particionando os dados a cada recursão em pedaços menores, até a construção total da árvore. Como mencionado, o particionamento se dá na melhor escolha do atributo da tupla em busca de partições puras, isto é, que cada partição crie tuplas com a mesma classe (HAN; KAMBER, 2006). Entre os mecanismos de seleção de atributos que melhor representam a tupla, podemos citar o Information Gain e o Gini Index, sendo o primeiro particionando os dados em múltiplas divisões gerando vários caminhos na árvore, e o segundo em apenas em dois, forçando a árvore formada a ser binária (HAN; KAMBER, 2006).

Tendo todos esses conceitos em mente, a seção seguinte discorre sobre o classificador utilizado na pesquisa e que se utiliza desses termos para classificar dados, sendo apresentado a seguir.

2.4 Floresta Aleatória

Um dos algoritmos mais utilizados em pesquisas de classificação supervisionada está o Random Forest, em tradução livre, Floresta Aleatória. Diversas pesquisas atuaram verificando sua eficiência na classificação de dados, entre elas um estudo analisando 179 classificadores, incluindo de diversas famílias e categorias, bem como que sua análise fosse independente da coleção dos dados utilizada, e chegou-se à conclusão que os algoritmos que possuem o maior desempenho de classificação estão as versões que utilizam o Random Forest (DELGADO, 2014), chegando a um máximo de 94% de acurácia dos dados utilizados quando comparado aos outros classificadores.

Uma outra vantagem do uso desse algoritmo está no seu fácil uso em diversas implementações (R Statistics, Python), além de facilmente computar dados, visto a grande abrangência de dados disponíveis atualmente (DEGENHARDT et al., 2017). A implementação usada nessa pesquisa foram as contidas no pacote scikit-learn.org [1].

Random Forest utiliza o conceito de *ensemble classification*, ou seja, utiliza a ideia de classificar através de um conjunto de classificadores, chamado *ensemble classifiers*. De maneira geral, um ensemble é formado por várias instâncias de classificadores base, não vistos um pelo outro, e através de um mecanismo de voto, é retornado a classificação (BRAMER, 2013; DEGENHARDT et al., 2017; HAN; KAMBER, 2006). Os classificadores base podem ser homogêneos (formados pelo mesmo tipo no ensemble) ou heterogêneos (diversos tipos no ensemble), contudo, pesquisas indicam que classificadores ensemble parecem dar respostas mais positivas na classificação e na acurácia. (HO, 1995; DELGADO, 2014; BRAMER, 2013; DEGENHARDT et al., 2017). Em relação ao Random Forest, ele se encontra como ensemble homogêneo de árvores de decisão. Entretanto, na implementação usada nessa pesquisa usando o pacote sklearn, ao contrário do padrão original proposto por Leo Breiman (BREIMAN, 2001) em que ocorre a votação, ocorre aqui a média da probabilidade de classificação de cada classificador base para a classificação final[2].

[1] https://scikit-learn.org/stable/
[2] https://scikit-learn.org/stable/modules/ensemble.htmlb2001

Adicionalmente, há diversas formas pela qual um ensemble pode ser formado, desde o número de árvores até os valores para treinamento. Todavia, foi focado na formação e nos parâmetros utilizados pelo Random Forest para essa pesquisa.

Supondo que o mesmo conjunto de dados de treino para várias árvores podem gerar resultados super otimistas, são utilizados algumas técnicas de seleção de dados de treinos para classificações ensemble. As mais utilizadas nas técnicas de aprendizagem de máquina são bagging, boosting, and stacking (TORABI et al., 2021; BRAMER, 2013). Cada técnica possui sua especificidade, sendo a utilizada pelo Random Forest a técnica de **Bagging**.

Bagging consiste em selecionar aleatoriamente tuplas formando subconjuntos de dados do conjunto de treinamento, mantendo-se os registros originais no conjunto de origem, e então, cada subconjunto ser utilizado para a geração de uma árvore do ensemble. O detalhe aqui é que algumas tuplas podem ser selecionadas mais de uma vez, podendo, inclusive, ter registros apenas da mesma classe no subconjunto. Contudo, estatisticamente, 63.2% dos registos de treinamento são usados para a geração das árvores (GOLDSCHMIDT, 2006), sendo o conjunto total de treinamento usado para validação delas.

Por fim, seu nome Floresta Aleatória decorre de ser constituído de várias árvores (Floresta), em que cada árvore trabalhará um conjunto aleatório de atributos (Random Features). Consequentemente, cada árvore poderá particionar em cada nó um atributo diferente de outra árvore, gerando árvores com baixa correlação entre si, por isso o termo Aleatória.

2.4.1 Importância dos Atributos

Tendo colocado conceitos básicos do Random Forest e como ele funciona, a pesquisa se baseou em verificar a importância que esse classificador deu para cada atributo na classificação final, também chamada Feature Importance. A mensuração dessa importância pode ser dada basicamente de duas formas: Mean Decrease in Accuracy (MDA), e Mean Decrease in Gini (MDG) (BREIMAN, 2001; BELGIU et al., 2014; LIAW; WIENER, 2002). O MDA mede a importância do atributo verificando mudanças na acurácia quando valores de um atributo são aleatoriamente modificados. Para o MDG, calcula a soma de todas as reduções de impureza do Gini, normalizado pelo total de árvores (BREIMAN, 2001;

BELGIU et al., 2014; LIAW; WIENER, 2002). O que deve ser levado em consideração é que a depender do modelo de seleção de atributos, essas importâncias podem ter valores diferentes para o classificador gerado. Como exemplo, uma vez que é usado pelo Random Forest o Gini para partição dos atributos nos nós, pode ser usado, então, o MDG para verificação da importância, refletindo que quanto maior seu valor, mais importante o atributo.

A importância Gini de uma característica é um ranking entre um conjunto de características criado dentro do treinamento da floresta aleatória. A cada nó dentro de uma árvore binária de uma floresta aleatória, uma divisão ideal é realizada usando um cálculo de impureza Gini, medindo o quão boa é a potencial divisão do nó separando as duas classes daquele nó em partições puras (MENZE et al., 2009).

Considerando $P_k = n_k/n$ a proporção de amostras de uma classe binária $k = \{0,1\}$ dentro do nó τ, a impureza de Gini é calculada da seguinte forma (MENZE et al., 2009):

$$i(\tau) = 1 - P_1^2 - P_0^2 \tag{2.3}$$

Sua redução Δi como resultado da separação e do envio das amostras aos dois nós filhos, esquerdo τ_l e direito τ_r, com suas respectivas impurezas $i(\tau)_l$ e $i(\tau)_r$ cria a fórmula da redução a seguir (MENZE et al., 2009):

$$\Delta i(\tau) = i(\tau) - P_l i(\tau_l) - P_r i(\tau_r) \tag{2.4}$$

Assim, após o treinamento da floresta, uma busca na árvore sobre todas as divisões das variáveis nos nós e todos os possíveis valores de redução de impureza Gini alcançado nessas divisões em cada uma delas leva a um menor valor de impureza de Gini. Essa busca pelo menor valor de impureza Gini é acumulado para todos os nós em todas as árvores e individualmente para cada característica gerando o valor de importância Gini para cada característica utilizada (MENZE et al., 2009).

Dessa forma, a Importância Gini, de maneira simplista, indica a frequência que tal característica foi utilizada para uma divisão no nó e o quão importante é o seu valor dentro daquele conjunto de características (MENZE et al., 2009).

3 Trabalhos Relacionados

Nesse capítulo será relatado os trabalhos que se relacionam ao presente trabalho, sejam pelo algoritmo utilizado, sejam pelas técnicas utilizadas ou pela temática envolvida. O objetivo desse capítulo é mostrar que pesquisa segue em consonância com as demais pesquisas realizadas no momento, contudo, se distanciando nos dados utilizados a fim de trazer novas observações sobre a evasão no ambiente universitário.

O primeiro trabalho que se deve mencionar é o de de Bonifro et al. (2020). Esse trabalho é recente e se relaciona ao do presente trabalho uma vez que ele utiliza uma abordagem de segmentar os dados, utilizando os dois anos iniciais para análise. Estatisticamente, na pesquisa relacionada, a evasão acontece em proporções maiores nos primeiros anos de curso. Nesse trabalho, os dados são também parecidos com a presente pesquisa, formados por características pessoais como idade, representado por um range entre 1 e 3, gênero do aluno e características acadêmicas como nota do ensino médio equivalente, créditos cursados no ensino superior e requisitos de estudos adicionais para entrada no curso, totalizando ao fim 7 características analisadas. Essas características e seus intervalos de valores podem ser vistos na Tabela 4.

Tabela 4 – Características utilizadas por Bonifro et al. Fonte: o autor (2022)

Característica	Intervalo de valores
Student gender	1, 2
Student age range	1 a 3
High school id	1 a 10
High school final mark	60 a 100
Additional Learning Requirements	1, 2, 3
Academic school Id	1 a 11
Dropout	0 a 60

Esses dados foram submetidos a 3 algoritmos, Random Forest, Linear Discriminant Analysis e Support Vector Machine, considerados pelo autor como os melhores para o problema da predição da evasão. E entre os objetivos está determinar qual característica se projeta como importante para os algoritmos. Ao fim da pesquisa, entre outros

resultados, chegou-se as duas mais importantes características que são *Additional Learning Requirements* e *High school final mark*. Características que são originadas do aluno anteriormente a sua entrada no curso desejado. Ou seja, características já previamente existentes num aluno ingressante. Essa pesquisa, contudo, teve limitações quanto à quantidade de dados analisados, limitando-se apenas a 15 mil estudantes de apenas 2 anos (2016/2017).

Um outro aspecto analisado são quais as técnicas mais usadas por pesquisadores atualmente em buscas de soluções que evitem a evasão, sejam por fatores associados ao aluno ou pelas técnicas empregadas. Nesse contexto, o trabalho de Albán et al. (2019). faz uma revisão sistemática dos trabalhos que contém o tema da evasão e as principais técnicas empregadas nos estudos. Esse trabalho recupera artigos e pesquisas de grandes bases de periódicos a fim de responder algumas perguntas, entre elas, quais são os fatores que mais afetam a evasão e quais são as técnicas empregadas na predição da evasão e seu nível de confiabilidade. Para a pesquisa de Albán et al. (2019), foram consultados trabalhos contendo as palavras "evasão estudantil" e "mineração de dados" no título, abstract ou palavras-chave. Interessante observar que a pesquisa é também recente, publicada em 2019 e compreende os anos de 2006 até 2017. Contudo, dos trabalhos iniciais retornados na consulta, ele considerou apenas trabalhos que envolvessem evasão universitária, predição realizada por mineração de dados e que contivessem métricas avaliativas dos modelos preditivos criados, tendo ao final, após os critérios de seleção, 67 trabalhos analisados.

Entre outros resultados obtidos, a pesquisa mostrou que fatores econômicos são pouco estudados nas pesquisas da evasão. Albán et al. (2019) conseguiu identificar 17 fatores encontrados na sua revisão como relacionados à economia, tais como ajuda financeira, renda familiar, emprego dos pais, etc. Adicionalmente, 21 fatores foram encontradas nos diversos trabalhos considerados como fatores sociais, tais como uso de drogas ou problemas familiares. Todavia, os fatores mais estudados se concentravam nos fatores acadêmicos e pessoais, 64% dos fatores encontrados, tais como notas e data de nascimento. Esses resultados obtidos por Albán et al. (2019) mostram que poucas pesquisas têm se concentrado em outros aspectos dos estudantes além do aspecto pessoal uma vez que foi constatado baixo número de fatores sociais e econômicos estudados.

Uma outra constatação respondida por Albán et al. (2019) foi quais são as técnicas mais empregadas nesses estudos. Em sua pesquisa, foi constatado que 79% dos trabalhos

relacionados envolvendo mineração de dados e evasão usavam algoritmos baseados em árvore de decisão devido a sua flexibilidade de processar dados numéricos e categóricos.

Quanto ao contexto brasileiro, um trabalho que se aproxima bastante da presente pesquisa é o de Rego (2016). Esse trabalho segue a mesma linha de pesquisa tentando prever o aluno com grandes chances de evasão, bem como identificar as principais características que estão associados ao fenômeno da evasão de tais alunos. Para isso, foi coletado uma base dados de uma universidade pública contendo dois grupos de dados, um chamado de Acadêmico e outro Socioeconômico. Após a coleta e tratamentos desses dados, restaram 22 atributos especificados nas Tabelas 5 e 6.

Tabela 5 – Características Acadêmicas utilizadas por Rego (2016). Fonte: o autor (2022)

Característica	Descrição	Tipo
nota_calculo_1	Nota obtida na disciplina	Numérico (0 a 10)
nota_calculo_vet	Nota obtida na disciplina	Numérico (0 a 10)
nota_fisic_1	Nota obtida na disciplina	Numérico (0 a 10)
nota_intro_prog	Nota obtida na disciplina	Numérico (0 a 10)
nota_intro_comp	Nota obtida na disciplina	Numérico (0 a 10)
escola_ens_fund	Categoria do ensino fundamental	Nominal
escola_ens_medio	Categoria do ensino médio	Nominal
turno_ens_medio	Turno no ensino médio	Nominal
mediageral	Média de admissão para a Universidade	Numerico (0 a 1000)

Uma vez os dados prontos para uso, foi então submetido ao algoritmo de classificação *Naive Bayes*, obtendo uma acurácia máxima de 85,48% para a classificação da evasão, obtendo, também, uma ordem de importância de tais atributos. É importante destacar que o trabalho se restringiu apenas a um curso de graduação com apenas 241 tuplas. Um número bem baixo de instâncias se comparado ao da presente pesquisa. Os resultados de Rego (2016) projetam os dados acadêmicos como mais importantes sobre os dados socieconômicos, sendo as duas características socioeconômicas mais importantes a escolaridade da mãe e sua situação empregatícia. Entretanto, devido a sua limitação na variedade dos dados, sendo apenas um curso com 241 registros, não é possível afirmar que tais conclusões possam ser aplicadas a outros cursos de outras áreas do conhecimento, assim como a sua pequena quantidade de dados deixam margem para uma projeção mais segura dessas

Tabela 6 – Características Socioeconômicas utilizadas por Rego (2016). Fonte: o autor (2022)

Característica	Descrição	Tipo
sexo	Sexo do estudante	Nominal
cor_pele	Cor da pele do estudante	Nominal
estado_civil	Estado civil do estudante	Nominal
renda_familiar	Renda familiar do estudante	Nominal
trabalha	Situação de trabalho	Nominal
possu_pc	Se o estudante possui computador próprio	Nominal
acessa_internet	Se o estudante acessa internet	Nominal
sit_pai	Situação empregatícia do pai do estudante	Nominal
sit_mae	Situação empregatícia da mãe do estudante	Nominal
prof_pai	Tipo de profissão do pai	Nominal
prof_mae	Tipo de profissão da mãe	Nominal
instr_pai	Grau de instrução educacional do pai	Nominal)
instr_mae	Grau de instrução educacional da mãe	Nominal

características.

Outro trabalho que pode ser referenciado é o trabalho de Gonçalves (2019), uma vez que utiliza a mesma temática da pesquisa. Esse trabalho coleta dados do curso de Bacharelado em Administração com vistas a identificar alunos com possibilidade de evasão, contudo, não é mencionado no trabalho a quantidade de dados trabalhados ou a partir de quando os dados se iniciam ou terminam, citando apenas que o curso iniciou em 2014.

Esse trabalho recupera 10 características desses dados e os submetem a diversos algoritmos, 9 ao total, entre eles o Random Forest. O autor considera a característica de frequência como a única de rendimento escolar, e as demais como socioeconômicas, e dessa forma, realizou dois experimentos em separado, um experimento incluindo a característica frequência e outro removendo-a. Esse trabalho possui várias lacunas de informação, pois não informou quais os valores que as características podem possuir ou como se deu a construção da característica frequência.

Como resultado, obteve uma acurácia grande quando incluiu a característica frequência nos seus dados, atingindo 81,5% para o Random Forest. Para o experimento sem a frequência, no algoritmo Random Forest, obteve-se 59,7%, tendo também, estabelecido

uma ordem de importância dessas características, colocando a frequência no topo, seguido de características da cidade e escola de origem. Um detalhe importante é que o autor afirma em sua pesquisa que alguns dados possuem lacunas e dados faltantes para certos atributos, o que pode provocar a falta de validação do modelo para esses atributos. De maneira geral, mostrou-se uma pesquisa incipiente, mas que tratou da temática de maneira parecida.

Diversos outros autores podem ser referenciados nesse tema de mineração de dados e evasão, desde trabalhos com características de ambientes virtuais de aprendizagem para o ensino a distância como o trabalho de Sepúlvida (2016), em que analisa as características nos ambientes virtuais que levam à evasão ou trabalhando com o tema da evasão no ensino básico como o trabalho de Ataíde (2016) em que observa características do ensino básico das escolas públicas e privadas de Pernambuco. Todavia, será dado destaque àqueles citados que possuem informações mais completas sobre seus dados e que mais se aproximam da presente pesquisa. Sendo assim, uma tabela contendo os dados trabalhados e suas principais informações podem ser vistas na Figura 4.

Citação	Algoritmos utilizados	Tipos de características	Resultado da classificação	Fez análise das melhores características?	Fez análise por diferentes áreas?
(BONIFRO et al., 2020)	Random Forest, Linear Discriminant Analysis e Support Vector Machine	Apenas dados institucionais.	87%	Apenas disponibilizou o resultado sem contextualizar.	Não.
(REGO, 2016)	Naive Bayes	Apenas dados institucionais.	85,48%	Apenas disponibilizou o resultado sem contextualizar.	Não.
(GONÇALVES; BELTRAME, 2019)	9 ao total, entre eles o Random Forest.	Apenas dados institucionais.	Máximo de 81,5% para o Random Forest	Apenas disponibilizou o resultado sem contextualizar.	Não

Figura 4 – Principais trabalhos relacionados e suas principais informações. Fonte: o autor (2022)

Muito além da pesquisa de Bonifro et al. (2020), a presente pesquisa analisa duas bases de dados, sendo a segunda base composta por mais de 41 mil alunos e diversos cursos compreendendo os anos de 2012 a 2019, assim como considerando 27 características, um número bem maior que o utilizado por Rego (2016). Outro ponto adicional referente aos dados é que eles possuem uma variedade maior que os trabalhos relacionados, sendo

utilizado além da base interna institucional, outras bases de dados a fim de analisar essas características associadas a outras bases. Adicionalmente, o presente trabalho difere dos demais trabalhos na medida em que analisa as características por área de conhecimento, o que até então nenhum trabalho referenciado analisou. Sendo assim, é possível constatar pelos trabalhos relacionados que, embora utilizem técnicas parecidas, não abordaram os mesmos dados pesquisados. Outra justificativa que deve ser mencionada é que os dados abordados possuem informações relativas aos programas de reserva de vagas, assim como índices de vulnerabilidade social dos alunos e da região. Esses indicadores são importantes para avaliar as políticas de cota da universidade estudada, limitando-se, todavia, aos dados produzidos pelo seu sistema acadêmico.

4 Metodologia

Esse capítulo descreve a metodologia utilizada para a coleta e avaliações dos dados selecionados para a pesquisa. Como citado na introdução, foram dois estudos que possuem em comum as mesmas perguntas de pesquisa. Adicionalmente, cada uma das perguntas foi destinada a bases de dados diferentes, especificadas mais a frente. Entretanto, para ambas as perguntas e bases utilizadas, foram realizadas as mesmas etapas metodológicas descritas na seção seguinte. Sendo assim, para o primeiro estudo, **Estudo 1**, os dados se referem ao censo demográfico, e para o segundo estudo, **Estudo 2**, os dados se referem ao SISU.

4.1 Etapas da Pesquisa

Ao todo, as tarefas podem ser resumidas em 4 etapas descritas na Figura 5. Importante destacar que alguns resultados de certas etapas agiram de maneira a retroalimentar a etapa anterior, de forma que as tarefas se deram de maneira iterativa até atingir resultados satisfatórios e consistentes.

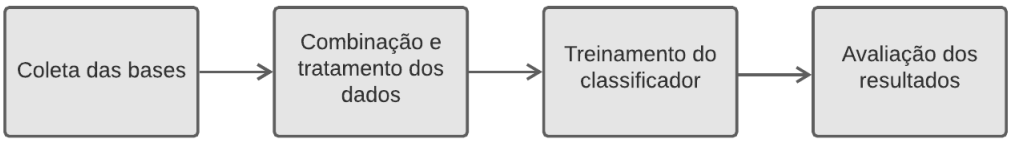


Figura 5 – Fluxo das etapas realizadas. Fonte: o autor (2022)

A descrição de cada etapa se encontra a seguir:

1. **Coleta dos dados** - Essa etapa compreendeu a coleta dos dados para a pesquisa. Os dados institucionais foram fornecidos pela UFRPE. As demais bases são de portais e domínios públicos, os quais ficam disponíveis aos demais pesquisadores. Uma vez que os dados foram coletados, seguiu-se à etapa seguinte.
2. **Combinação e tratamento dos dados** – Essa etapa teve como objetivo juntar os dados institucionais dos alunos com os demais dados de forma a criar um único conjunto de dados. Nesse ponto, ao final, formou-se um único conjunto de dados para análise. Nessa etapa também houve o tratamento dos valores, convertendo-os para valores numéricos quando necessário, assim como removendo valores nulos e em

branco, a fim de manter a consistência dos dados. Nessa etapa, foram descartadas tuplas e colunas que pudessem estar repetidas e mantidas apenas os dados e colunas que representassem o objeto da pesquisa, o que tornou o conjunto de dados mais enxuto. Adicionalmente, nessa etapa, os dados foram agrupados em áreas do conhecimento do curso, e submetidos ao classificador por esse agrupamento. Esse agrupamento facilitou a análise e a comparação dos resultados.

3. **Treinamento do classificador** – Essa etapa compreendeu a criação do código em python utilizando o pacote sklearn para a geração de um modelo, ou seja, de um classificador baseado na ideia do Random Forest. Essa etapa, uma vez concluída, foi usada para todos os conjuntos de dados formados na etapa anterior.
4. **Avaliação dos resultados** – Nessa etapa, os dados foram submetidos ao classificador a fim de obter uma ordem de importância para cada característica. Os resultados da submissão foram coletados e disponibilizados em gráficos para uma melhor visualização dos resultados.

4.2 Ferramentas Utilizadas

As ferramentas incluem todos os tipos de programas e suportes utilizados para a pesquisa. De maneira geral, foi focado nas configurações para a construção do algoritmo.

Para a construção do algoritmo, foi utilizado a linguagem python[1], escolhida por apresentar bibliotecas que possibilitam utilizar abordagens necessárias para essa pesquisa, assim como também utilizadas em pesquisas semelhantes (BARBOSA et al., 2021; CAVALCANTI et al., 2020; FERREIRA et al., 2018). Entre essas bibliotecas, destaca-se a biblioteca de mineração de dados e aprendizagem de máquina chamada sklearn[2]. Conforme consta em sua documentação, possui características de ser open source, possuir ferramentas simples e eficiente de predição de dados, reusável em vários contextos e acessível a todos. Em relação aos dados, foram coletados em formato Microsoft Excel .xls e convertidos para formato csv para utilização do algoritmo.

[1] https://www.python.org/
[2] https://scikit-learn.org/stable/

4.3 Metodologia da Pesquisa do Estudo 1

Nas próximas subseções estão descritas a coleta, tratamento e processamento dos dados do Estudo 1. Esse estudo compreende a análise da eficácia do classificador e seus fatores que podem determinar a evasão compreendendo os dados da universidade em associação aos dados do Censo demográfico.

4.3.1 Coleta dos Dados

Como citado no capítulo introdutório, o primeiro estudo se deu em duas bases de dados: uma acadêmica onde contém informações de formação dos alunos e seus dados pessoais, e outra base demográfica, onde contém informações demográficas e socioeconômicas sobre as cidades do Brasil[3]. Essas duas bases foram associadas para, então, determinar quais as características se tornam importantes segundo o classificador, para em seguida, analisá-las. Importante mencionar que a base demográfica contém dados do último censo brasileiro realizado em 2010. Muito embora o censo seja de 2010, o marco temporal dos dados acadêmicos são de 2010 ate 2019, ou seja, durante esse período, o censo de 2010 é o único realizado e disponível para esse período. Os dados acadêmicos foram retirados do sistema acadêmico da Universidade Federal Rural de Pernambuco, a qual possui diversos cursos de graduação, dentre cursos presenciais e à distância. Foram selecionados os alunos cuja última situação acadêmica tenha sido Formado a partir de 2010. Ou seja, alguns alunos podem ter se matriculado antes de 2010, entretanto, a consulta se deu em alunos cuja conclusão no curso tenha sido dada a partir de 2010. Assim como, visando atingir o anonimato dos dados acadêmicos, de forma que não se possa identificar os alunos, foram omitidos na consulta seus dados pessoais, tais como nome, endereço, CPF, deixando apenas sua situação acadêmica, cidade natal, sexo, estado civil e o ano do seu ingresso.

As situações acadêmicas encontradas são citadas a seguir: *matriculado, matrícula vínculo, trancamento, formado, complementação curricular, desligamento, transferência externa, desvinculado, transferência interna, trancamento de semestre anterior, titulado, matriculado sub judice, desistência, reintegração, intercâmbio, excluído, integralizado, mobilidade estudantil*. Contudo, o trabalho tem foco em evasão acadêmica. Daí então,

[3] http://www.atlasbrasil.org.br/consulta

essas situações foram agrupadas em apenas dois grupos: *Evadido* e *Não Evadido*. Foram considerados alunos Evadidos todos aqueles que não concluíram o curso com sucesso. Situação de *desistência*, *desvinculado*, *desligamento* e *excluído* foram agrupados como Evadidos. Para o grupo do Não Evadido, foram considerados os demais alunos das situações restantes, citadas a seguir: matriculado, matrícula vínculo, trancamento, formado, complementação curricular, transferência externa, transferência interna, trancamento de semestre anterior, titulado, matriculado sub judice, reintegração, intercâmbio, integralizado, mobilidade estudantil. Essas situações consideram alunos que ainda possuem vínculo com a universidade estudada ou que possuem o curso concluído com sucesso, por isso, considerados Não Evadidos.

Os dados demográficos se referem ao site do Atlas do Desenvolvimento Humano no Brasil e traz índices como o Índice de Desenvolvimento Humano Municipal (IDHM) e outros mais de 200 indicadores de demografia, educação, renda, trabalho, habitação e vulnerabilidade para as cidades brasileiras. Essa base foi colocada numa tabela cuja coluna principal é o nome da cidade, coluna em que determina semanticamente as demais colunas contendo os seus respectivos índices.

4.3.2 Combinação e Tratamentos dos Dados

Uma vez os dados disponíveis para uso, as bases foram unidas de forma a relacionar cada aluno a sua cidade, com seus respectivos indicadores demográficos e socioeconômicos. Assim, os dados foram colocados no mesmo esquema, e em seguida fez-se uma consulta entre ambos os dados com a junção se dando na cidade natal do aluno e sua unidade federativa com a cidade do atlas brasileiro e sua unidade federativa. Foi escolhido a cidade natal do estudante de forma a representar a cidade de origem, dando maior variabilidade as cidades analisadas. Gerando, consequentemente, um conjunto de dados cuja tupla contém informações do grupo acadêmico, evadido ou não evadido, com seus respectivos índices demográficos. Vale ressaltar, mais uma vez, que os índices são relativos aos dados do último censo brasileiro realizado em 2010. Ao total, obteve-se 37157 registros de alunos para análise, divididos em 13360 alunos considerados evadidos e 23797 alunos considerados não evadidos. Um resumo das características acadêmicas e demográficas podem ser visualizadas pela Tabela 7.

Nesse ponto, os dados foram divididos por área de conhecimento, conforme as áreas definidas pelo CNPq[4]. Ou seja, foram agrupados os cursos pelas suas áreas, quer sejam áreas de ciências humanas, ciências sociais, exatas e da terra, etc. Esse agrupamento se fez necessário para dar maior granularidade à pesquisa, dando maior especificidade aos resultados obtidos. Dessa forma, foram formados 6 grupos de dados representando a área de conhecimento dos cursos afins. Estas áreas estão citadas a seguir: ciências sociais, ciências agrárias, ciências biológicas, engenharias, ciências humanas e exatas e da terra.

Uma vez os dados agrupados e disponíveis para uso, estes foram tratados para a aplicação do Random Forest, que inclui tratamentos de valores de string para números e remoção de valores em branco e colunas duplicadas. E, assim, aplicando o algoritmo para cada grupo de dados especificado acima.

Tabela 7 – Resumo do conjunto de dados. Fonte: o autor (2022)

Grupo de características	Quantidade	Valores
Características do estudante internas da instituição	06	Gênero(categórico), Ano de Admissão(numérico), cidade natal(categórico), Situação Acadêmica (categórico), Estado Civil(categórico) e Cor/Raça(categórico)
Dados demográficos municipais	237	Indicadores de demografia, educação, renda, trabalho, habitação dos municípios brasileiros tais como expectativa de vida, taxa de fecundidade, renda entre outros.(Todos em valores numéricos)

Adicionalmente, nessa etapa, devido ao grande número de características dessa base, para cada grupo aplicado ao algoritmo, foram realizados 3 experimentos. Essa divisão foi feita com o objetivo de analisar as características em diferentes cenários. O primeiro experimento foi realizado contendo todas as características do conjunto de dados. Lembrando que nesse cenário, existem 237 características. Este primeiro experimento possui uma mineração crua dos dados, sem nenhuma remoção ou escolha. O segundo experimento foi realizado removendo-se a característica mais marcante dentre a primeira experimentação. A referida característica foi ANO_ADMISSAO. Essa característica foi removida. Vale mencionar, nesse ponto, que não há como esta característica se repetir para anos futuros. Essa remoção, explicada nas seções posteriores, trouxe resultados mais coerentes. E um terceiro experimento, escolhendo-se certas características também foi

[4] http://lattes.cnpq.br/documents/11871/24930/TabeladeAreasdoConhecimento.pdf/d192ff6b-3e0a-4074-a74d-c280521bd5f7

realizado. Esse terceiro cenário foi realizado contendo um total de 8 características. Um número baixo a fim de fazer um balanço mais equânime entre as características pessoais e demográficas. Um resumo dos experimentos pode ser melhor entendido na Tabela 8.

Tabela 8 – Resumo dos experimentos. Fonte: o autor (2022)

6 grupos de dados por área do conhecimento		
Cada grupo de dados submetidos a 3 experimentos		
1º Experimento	2º Experimento	3º Experimento
Todas as características	Todas as características, excetuando a mais relevante da 1º experimentação	Escolhendo as características

Uma observação a ser mencionada é que no terceiro experimento, ao ser escolhida as características, foram escolhidas todas relativas ao aluno e algumas características relativas a índices socioeconômicos dos municípios. Dos alunos, optou-se por trazer todas características por terem pouca quantidade. Em contrapartida, dos índices demográficos foram escolhidos, no 3º experimento, Taxa de Fecundidade Total representado pela sigla FECTOT, Esperança de Vida ao Nascer representado por ESPVIDA, Expectativa de Anos de Estudo representado por E_ANOSESTUDO e, por último, a porcentagem de pessoas em domicílios com abastecimento de água e esgotamento sanitários inadequados representado por AGUA_ESGOTO. Optou-se por escolher essas características por representar parâmetros mais gerais e amplos da cidade nos temas de níveis sanitários, educacionais e humanos. Estes índices acima citados não estão restringidos apenas a faixas de valores, como os demais índices demográficos estão nestes referidos temas. Ressalte-se que outros poderiam ser escolhidos, todavia, a pesquisa se limitou a estes na terceira medição.

Outro ponto a ser considerado é que as características dos alunos não envolveram características como médias ou coeficientes de rendimento ou número de matrículas. Foram analisadas somente características já existentes num aluno ingressante, uma vez que demais características serão produzidas no decorrer do curso. Dessa forma, para o gestor universitário, a fim de evitar evasão, necessita de antecipação quanto a fatores de vulnerabilidade iminente, tornando o monitoramento do abandono escolar mais eficiente.

4.3.3 Treinamento do Algoritmo

O algoritmo Random Forest contem alguns parâmetros fundamentais para seu funcionamento. Sendo assim, as configurações dos principais parâmetros usados para a construção do classificador implementado em python baseado no pacote sklearn[5] foram *n_estimators* com valor de 100, indicando o número de árvores do ensemble, *criterion* para gini, indicado o método de escolha do atributo divisor no nó e *max_depth* como ilimitado indicando que os nós serão expandidos até as folhas serem puras.

Adicionalmente, como definido no particionamento Holdout, os dados foram divididos, para cada grupo, em dois conjuntos: treinamento e teste, com 70% dos dados reservados para treinamento visando a construção do modelo e os outros 30% para teste e validação desse modelo. E em seguida, finalizando, o algoritmo foi ajustado para receber os tipos de dados mencionados com sua devida quantidade de atributos.

A acurácia e os resultados desses experimentos podem ser encontrados no próximo capítulo, onde também é discutido e interpretados seus valores.

4.4 Metodologia da Pesquisa do Estudo 2

Para esse segundo estudo, Estudo 2, os dados são originários de duas bases distintas, a primeira sendo do sistema acadêmico e a segunda provinda do sistema de seleção universitária (SISU).

4.4.1 Conjunto de Dados

Os dados da primeira base se referem às informações do sistema acadêmico da Universidade Federal Rural de Pernambuco. Os dados dessa base abrangem 31 cursos presenciais compreendidos entre os anos de 2012 até 2019. Informações como etnia, estado civil, cidade onde o aluno mora e seus *status* de formado/matriculado ou evadido foram extraídos. Para esse estudo, foi feito um recorte dos dados dos dois anos iniciais após o ingresso, pois trabalho semelhante recente apontou que esse é o período com maior evasão no ensino superior (BONIFRO et al., 2020). Assim como, para classificação de um aluno evadido utilizou-se a seguinte regra:

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

- Se o aluno estiver com o status Matriculado, classificamos seu status como *Não Evadido*;
- Se seu status, após dois anos do seu ingresso, estiver com outro status (Desistência, Excluído, Desvinculado ou Desligado), sua classificação foi *Evadido*.

O segundo conjunto de dados foi extraído do sistema de seleção unificada para ingresso nas universidades públicas brasileiras (SISU)[6]. Esse sistema é responsável pelo registro dos candidatos às vagas das universidades públicas brasileiras. Entre outras informações coletadas para a análise estão as notas obtidas no exame, tais como nota do ENEM nas diferentes disciplinas (redação, matemática, linguagens, ciências naturais), assim como informações de endereço, na época da realização da prova do ENEM, e das cotas sociais ao qual o candidato participava, como por exemplo, cotas destinadas a vagas de deficiente físico, renda familiar baixa ou se candidato de cor negra. Esses dados são importantes, pois fazem parte da política estatal brasileira de tentar equilibrar as vagas das universidades públicas para candidatos com maior grau de vulnerabilidade social, visando garantir quantitativos mínimos de formação de candidatos desses segmentos.

4.4.2 Combinação e Tratamento dos Dados

As duas bases foram unidas entre si pelo identificador único que cada aluno possuía em ambas as bases. É importante ressaltar que todas as colunas que identificavam a pessoa foram removidas a fim de manter o anonimato dos dados analisados. Uma vez unidas e disponíveis para uso, os dados foram tratados, removendo valores em branco ou lacunas entre eles. Para utilização dessa versão do algoritmo de classificação, os dados categóricos precisam ser transformados em numéricos. Para isso, foram aplicados transformações das características utilizando o *Label Encoder Algorithm*[7]. Esse algoritmo, transforma atributos categóricos em uma lista crescente de inteiros. Por exemplo, a característica Gênero, Masculino ou Feminino, foi transformada para 0 ou 1.

Ao final, o conjunto de dados obteve 27 características dos alunos e 41.460 instâncias divididos em 4.521 alunos evadidos e 36.939 alunos não evadidos. Embora os dados se encontrem desbalanceados, eles representam a realidade. Por isso, o uso do índice Kappa e da matriz de confusão visa dar mais segurança aos resultados apresentados, pois

[6] https://sisu.mec.gov.br
[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

são utilizados em trabalhos semelhantes na areá de educação (BARBOSA et al., 2021; CAVALCANTI et al., 2020; FERREIRA et al., 2018). A Tabela 9 apresenta o conjunto de características extraídas de cada base com média e desvio padrão para os atributos numéricos e número de opções para os atributos categóricos. É importante destacar que para cada nota do ENEM o aluno também tem uma nota "curso escolhido"que representa a nota do ENEM multiplicado por um fator de importância que cada curso da universidade pode atribuir para as competências do ENEM durante a seleção dos alunos.

Adicionalmente, visando uma análise mais detalhada, os dados foram agrupados por área de conhecimento do curso[8], totalizando 6 áreas de conhecimento, citadas a seguir: Ciências Sociais, Ciências Humanas, Ciências Agrárias, Exatas e da Terra, Biológicas e Engenharias. Cada área submetida ao classificador, registrando seus resultados para compará-los entre si.

4.4.3 Treinamento do Classificador

Por fim, da mesma forma que no Estudo 1, o algoritmo implementado em python baseado no pacote sklearn foi treinado tendo os principais parâmetros de funcionamento definidos como *n_ estimators* com valor de 100, indicando o número de árvores do ensemble, *criterion* para gini, indicando o método de escolha do atributo divisor no nó e *max_ depth* como ilimitado indicando que os nós serão expandidos até as folhas serem puras.

Adicionalmente, os dados foram divididos em dois conjuntos: treinamento e teste, com 70% dos dados reservados para treinamento visando a construção do modelo e os outros 30% para teste e validação desse modelo.

Por fim, a acurácia, os resultados e análises desse estudo podem ser visualizados no capítulo seguinte.

[8] http://lattes.cnpq.br/documents/11871/24930/TabeladeAreasdoConhecimento.pdf/d192ff6b-3e0a-4074-a74d-c280521bd5f7

Tabela 9 – Características utilizadas para o estudo 2. Fonte: o autor (2022)

Características do SISU				
Características	Tipo	Significado	Média	σ
NU_NOTA_M	Numérico	Nota de Matemática do ENEM	570.379	101.600
NU_NOTA_CURSO_M	Numérico	Nota de Matemática para o curso escolhido	886.566	670.739
NU_NOTA_L	Numérico	Nota de Linguagens do ENEM	555.146	55.913
NU_NOTA_CURSO_L	Numérico	Nota de Linguagens para o curso escolhido	981.946	698.456
NU_NOTA_CH	Numérico	Nota de Ciências Humanas do ENEM	593.394	60.598
NU_NOTA_CURSO_CH	Numérico	Nota de Ciências Humanas para o curso escolhido	889.135	406.680
NU_NOTA_CN	Numérico	Nota de Ciências Naturais do ENEM	532.554	67.288
NU_NOTA_CURSO_CN	Numérico	Nota de Ciências Naturais para o curso escolhido	1177.059	733.848
NU_NOTA_R	Numérico	Nota de Redação do ENEM	613.416	120.538
NU_NOTA_CURSO_R	Numérico	Nota de Redação para o curso escolhido	1087.566	422.111
NU_NOTA_INSCRITO	Numérico	Nota do inscrito do ENEM	584.413	56.412
ST_LEI_OPTANTE	Categórico	Indica se a oferta modalidade é da lei.	Valores de Sim ou Não	
ST_LEI_ETNIA_P	Categórico	Indica se é optante da etnia Preto ou Pardo.	Valores de Sim ou Não	
ST_LEI_ETNIA_I	Categórico	Indica se é optante da etnia Indígena.	Valores de Sim ou Não	
ST_LEI_RENDA	Categórico	Indica se renda inferior a 1,5 salário mínimo.	Valores de Sim ou Não	
NO_MOD_CONCORREN	Categórico	Nome da Modalidade de Concorrência.	25 modalidades	
Características do Sistema Acadêmico				
Características	Tipo	Significado	Média	σ
ANO_NASCIMENTO	Numérico	Ano de nascimento	1992.697	6.841
IDADE_PROVA	Numérico	Idade do aluno no momento da prova	23.055	6.717
NM_COR_RACA	Categórico	Cor autodeclarada pelo aluno.	6 valores	
NM_SEXO	Categórico	Gênero do aluno	Mas ou Fem	
NM_EST_CIVIL	Categórico	Estado civil	6 estados civis	
NO_BAIRRO	Categórico	Bairro de residência	1661 bairros	
NO_MUNICIPIO	Categórico	Município do aluno	613 municípios	
NO_CURSO	Categórico	Curso do aluno	31 cursos	
DS_TURNO	Categórico	Turno do curso	4 turnos	
NO_CAMPUS	Categórico	Campus do curso	4 campus	
SG_UF_INSCRITO	Categórico	Estado Federativo do aluno	26 estados	

5 Resultados

Esse capítulo relata os resultados obtidos nos experimentos descritos no capítulo anterior. Dessa forma, para as questões de pesquisa citadas na introdução desse trabalho são apresentados seus resultados e em seguida a discussão desses resultados.

5.1 Resultados do Estudo 1

Como relatado, esse estudo envolveu os dados institucionais dos alunos da universidade pesquisada em associação com os dados do último censo demográfico de 2010. Seus resultados são descritos nas seções abaixo, seguido de sua discussão ao fim.

5.1.1 Questão da Pesquisa 1

Inicialmente foi analisado se o modelo proposto de classificação de um aluno como evadido teria eficácia nos seus resultados. Sendo assim, o resultado alcançado poderia ser medido pela acurácia diante da classificação dos dados. Os resultados da acurácia geral por experimento podem ser observados na Tabela 10. É possível constatar nos resultados que o nível de acurácia gira próximo a 70% em todos os experimentos. Este valor é importante pois revela que é possível analisar e discutir os possíveis casos de evasão usando a abordagem proposta.

Tabela 10 – Resultados da Acurácia e do Kappa para o Estudo 1. Fonte: o autor (2022)

	1 Experimento	2 Experimento	3 Experimento
Acurácia	0.713401	0.686670	0.685414
Kappa	0.357625	0.228148	0.235676

É possível também verificar a matriz de confusão dos resultados dos três experimentos na Tabela 11 logo a seguir.

Tabela 11 – Matriz de confusão para o Estudo 1. Fonte: o autor (2022)

		Predição da Classe					
		1 Experimento		2 Experimento		3 Experimento	
		Positivo	Negativo	Positivo	Negativo	Positivo	Negativo
Classe Real	Positivo	1956	2055	1389	2616	1379	2666
	Negativo	1046	6090	908	6235	975	6128

5.1.2 Questão da Pesquisa 2

Em seguida, respondendo a segunda pergunta, foram verificados a acurácia e o kappa por área de conhecimento. De maneira a analisar e comparar melhor os resultados, eles foram disponibilizados em formato de gráfico. Sendo assim, a Figura 6 mostra o resultado geral da acurácia por área do conhecimento avaliada nesse estudo. O classificador conseguiu alcançar uma taxa de acurácia maior para a área de engenharia enquanto a área de ciências sociais e ciências agrárias foram as que obtiveram os menores resultados.

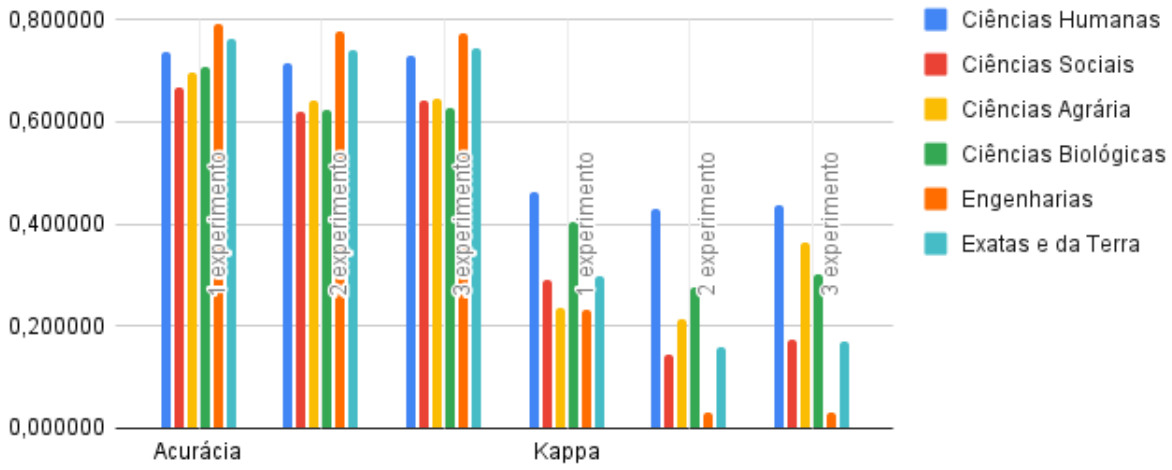


Figura 6 – Resultado da acurácia por área de conhecimento. Fonte: o autor (2022)

5.1.3 Questão de Pesquisa 3

Essa seção mostra os resultados obtidos após a execução dos experimentos a fim de responder a terceira pergunta. Como já mencionado, os resultados estão divididos em três experimentos diferentes e o principal foco é identificar as características mais relevantes para cada contexto analisado, não apenas o resultado final do classificador. Devido a

restrições de espaço, os resultados foram agrupados e disponibilizados em formato de tabelas listando as 8 primeiras características. O resultado da aplicação do algoritmo no primeiro experimento pode ser observado na Figura 7. O resultado da aplicação do algoritmo no segundo experimento é observado na Figura 8. Lembrando que neste segundo experimento foi removida a característica ANO_ADMISSAO. Por fim, a Figura 9 apresenta resultado do terceiro experimento. Ressaltando que, neste experimento, foram selecionadas características predeterminadas para análise conforme descrito na seção de metodologia.

Importância das Características											
1º Experimento											
Ciências Sociais		Ciências Agrárias		Ciências Biológicas		Engenharias		Ciências Humanas		Exata e da Terra	
Accuracy: 0.66685		Accuracy: 0.69737		Accuracy: 0.70656		Accuracy: 0.78989		Accuracy: 0.73514		Accuracy: 0.76194	
ANO ADMISSAO	0.5150	ANO ADMISSAO	0.5437	ANO ADMISSAO	0.5414	ANO ADMISSAO	0.4440	ANO ADMISSAO	0.3471	ANO ADMISSAO	0.5255
COR_RACA	0.1421	COR_RACA	0.1228	COR_RACA	0.1273	COR_RACA	0.2035	COR_RACA	0.0836	COR_RACA	0.1096
EST_CIVIL	0.0782	SEXO	0.0650	SEXO	0.0642	EST_CIVIL	0.0805	EST_CIVIL	0.0638	EST_CIVIL	0.0733
SEXO	0.0719	EST_CIVIL	0.0493	EST_CIVIL	0.0504	NM_SEXO	0.0692	NM_SEXO	0.0358	NM_SEXO	0.0574
MULH0A4	0.0041	PREN20RICOS	0.0101	PREN80	0.0105	T_ATIV1824	0.0045	MULH15A19	0.0191	REN3	0.0050
PESO13	0.0040	RDPC10	0.0066	R2040	0.0068	T_FLPRE	0.0041	HOMEM20A24	0.0156	T_ANALF18A24	0.0045
PESO1824	0.0028	PREN60	0.0066	PREN20RICOS	0.0067	T_ATIV18M	0.0038	PESO610	0.0155	T_FREQ25A29	0.0043
PREN80	0.0027	T_FUMED	0.0058	PREN60	0.0058	T_FUND12A14	0.0037	PESOM15M	0.0153	CORTE1	0.0043

Figura 7 – Resultado da aplicação do algoritmo no 1º experimento. Fonte: o autor (2022)

Importância das Características											
2º Experimento											
Ciências Sociais		Ciências Agrárias		Ciências Biológicas		Engenharias		Ciências Humanas		Exata e da Terra	
Accuracy: 0.62106		Accuracy: 0.64296		Accuracy: 0.62162		Accuracy: 0.77777		Accuracy: 0.71447		Accuracy: 0.73975	
COR_RACA	0.3253	COR_RACA	0.3414	COR_RACA	0.3019	COR_RACA	0.3123	COR_RACA	0.1316	COR_RACA	0.2610
EST_CIVIL	0.1410	NM_SEXO	0.1161	NM_SEXO	0.1407	NM_SEXO	0.1507	EST_CIVIL	0.1049	EST_CIVIL	0.1398
NM_SEXO	0.1241	EST_CIVIL	0.1008	EST_CIVIL	0.1107	EST_CIVIL	0.1173	SEXO	0.0565	NM_SEXO	0.1199
PREN20RICOS	0.0107	PREN60	0.0130	TRABPUB	0.0159	T_FORA4A5	0.0074	HOMEM20A24	0.0282	T_RMAXIDOSO	0.0081
R1040	0.0085	PPOB	0.0126	RDPC5	0.0140	PREN40	0.0069	PEA1517	0.0238	PEA1014	0.0076
PREN10RICOS	0.0083	RDPC10	0.0124	CORTE4	0.0133	T_ATIV	0.0067	PEA1014	0.0237	THEIL	0.0074
PREN80	0.0061	PREN20RICOS	0.0114	RDPCT	0.0123	T_ATIV2529	0.0063	MULH30A34	0.0235	RAZDEP	0.0074
HOMEM40A44	0.0053	R1040	0.0113	PREN10RICOS	0.0089	T_LIXO	0.0059	PESOTOT	0.0229	REN5	0.0073

Figura 8 – Resultado da aplicação do algoritmo no 2º experimento. Fonte: o autor (2022)

Importância das Características											
3º Experimento											
Ciências Sociais		Ciências Agrárias		Ciências Biológicas		Engenharias		Ciências Humanas		Exata e da Terra	
Accuracy: 0.64167		Accuracy: 0.64605		Accuracy: 0.62741		Accuracy: 0.77373		Accuracy: 0.73062		Accuracy: 0.74238	
COR_RACA	0.2788	COR_RACA	0.3125	COR_RACA	0.2629	COR_RACA	0.2914	AGUA_ESGOTO	0.2052	COR_RACA	0.1898
EST_CIVIL	0.1291	FECTOT	0.1279	FECTOT	0.1381	SEXO	0.1300	FECTOT	0.1803	FECTOT	0.1533
FECTOT	0.1122	AGUA_ESGOTO	0.1078	AGUA_ESGOTO	0.1230	EST_CIVIL	0.1204	ESPVIDA	0.1792	AGUA_ESGOTO	0.1347
ESPVIDA	0.1121	ESPVIDA	0.1018	EST_CIVIL	0.1058	E_ANOSESTUDO	0.1013	COR_RACA	0.1201	ESPVIDA	0.1280
AGUA_ESGOTO	0.1117	E_ANOSESTUDO	0.0991	NM_SEXO	0.1036	ESPVIDA	0.0957	E_ANOSESTUDO	0.0959	EST_CIVIL	0.1233
NM_SEXO	0.0857	E_ANOSESTUDO	0.0903	ESPVIDA	0.0947	FECTOT	0.0890	EST_CIVIL	0.0933	E_ANOSESTUDO	0.0963
MUNICIPIO	0.0856	MUNICIPIO	0.0853	E_ANOSESTUDO	0.0886	AGUA_ESGOTO	0.0875	MUNICIPIO	0.0813	MUNICIPIO	0.0952
E_ANOSESTUDO	0.0845	NM_SEXO	0.0748	MUNICIPIO	0.0827	MUNICÍPIO	0.0843	NM_SEXO	0.0444	NM_SEXO	0.0790

Figura 9 – Resultado da aplicação do algoritmo no 3º experimento. Fonte: o autor (2022)

Para melhor destacar a importância das características analisadas no terceiro experimento, a Figura 10 apresenta as sete principais características encontradas nesse experimento divididas por área do conhecimento.

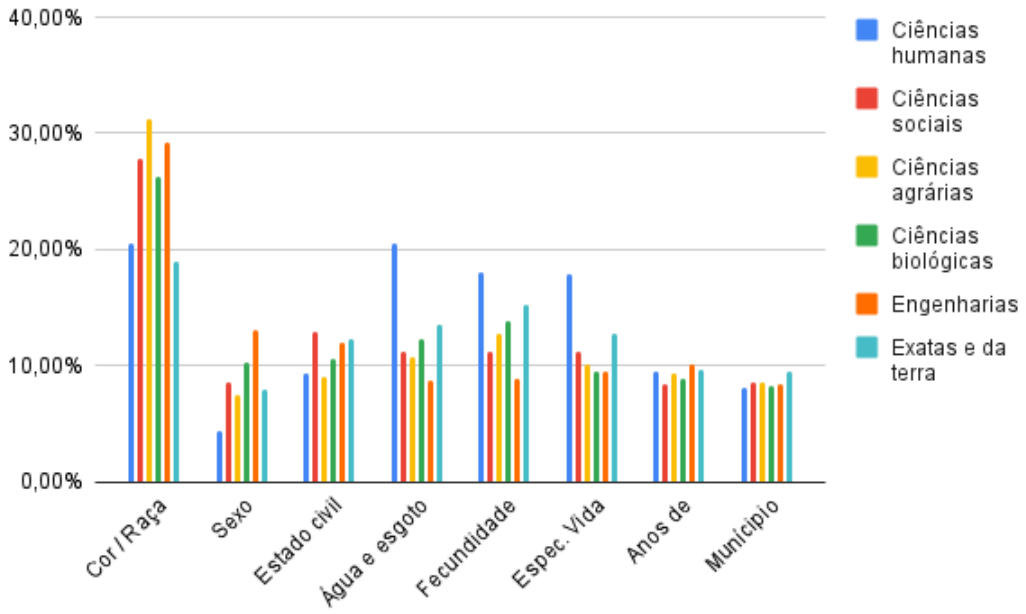


Figura 10 – Ordem de importância das características do 3º experimento. Fonte: o autor (2022)

5.1.4 Discussões da Questão de Pesquisa 3

É possível constatar no primeiro experimento que o ano de admissão do aluno possui grande influência na determinação do grupo acadêmico. Acredita-se que este fato decorra da grande quantidade de um determinado valor para um grupo acadêmico, o que pode ter ocasionado um desbalanceamento no algoritmo durante o experimento. O ano de admissão ficou acima de mais de 50% de influência em muitos grupos, sendo preponderante na determinação no grupo acadêmico. Ressalte-se neste primeiro ensaio que, seguido do ano de admissão, as características próprias dos alunos, tais como cor e estado civil seguem tendo mais influência em relação às características demográficas e, de certa medida, em todos os grupos, com distanciamento e influência maior entre características do aluno sobre características demográficas. Destaca-se aqui que a característica de cor e raça estão presentes em todos os grupos como segundo mais influente.

No segundo experimento, destaca-se a influência forte de cor e raça ao se remover o

ano de admissão. Há, neste ponto, um fator a ser enfatizado. A característica de ano de admissão é uma característica que não há como se repetir para alunos novos ou calouros. É uma característica do momento passado, não sendo válidas para o momento presente, assim como são valores que não são levados em consideração para programas de auxílio estudantil. Dessa forma, removê-la trará mais coerência à análise dos dados. Sendo assim, neste segundo experimento mostra-se uma equidade maior entre as características do aluno, mantendo-se a preponderância sobre as características demográficas. Ainda neste experimento, mantêm-se uma influência maior para a cor e raça contendo valor numérico maior sobre os demais em todos os grupos.

É possível perceber, neste momento, que não há nenhuma característica demográfica que seja relevante ou que se mantenha constante nas primeiras medições. Ao passo que, no terceiro experimento, ao se colocar características demográficas selecionadas, o resultado passa a ter um valor diferenciado, como mostrado na Figura 10.

No terceiro experimento, a característica de cor e raça se mantém constante como a de maior influência no grupo acadêmico , excetuando na área de ciências humanas. Adicionalmente, todas as características ainda conservam uma equidade entre os valores, não possuindo um sobressalente ou que se distancie bastante sobre os demais. Contudo, há aqui um ponto de destaque a ser realizado. Outras características pessoais perdem espaço para características demográficas. Características como Água e Esgoto ou Taxa de Fecundidade Total tornam-se relevantes, acima de características pessoais como gênero ou estado civil.

Com base nos resultados apresentados, as principais conclusões são:

1. Característica de cor e raça é determinante em todos os cenários, exceto em ciências humanas. O que sugere que esses elementos de cor, raça e etnia estão intrinsecamente ligados a fatores de evasão quando associados ao dados do censo demográfico.
2. Características socioeconômicas da cidade natal do estudante influenciam na determinação da evasão, a depender de quais características foram levadas em consideração para análise. As características intencionalmente selecionadas nessa pesquisa se revelaram mais importantes que características como gênero ou estado civil.

A seguir, tem-se os resultados do Estudo 2.

5.2 Resultados do Estudo 2

Como relatado, esse estudo envolveu os dados institucionais em associação com os dados do SISU. Seus resultados são descritos nas seções abaixo.

5.2.1 Questão de Pesquisa 1

Inicialmente foi analisado se o classificador possui eficácia necessária para utilizar a abordagem de classificação utilizando todos os dados disponíveis do Estudo 2. O resultado final atingiu acurácia e Kappa de 0.93 e 0.59, respectivamente. Os resultados Kappa atingem um nível considerado de concordância moderada (LANDIS; KOCH, 1977), o que atesta a eficácia do modelo proposto para esse estudo.

É possível também observar a matriz de confusão desses resultados logo abaixo na Tabela 12.

Tabela 12 – Matriz de confusão para o Estudo 2. Fonte: o autor (2022)

		Predição da Classe	
		Positivo	Negativo
Classe Real	Positivo	11076	4
	Negativo	644	714

5.2.2 Questão de Pesquisa 2

Após esse resultado inicial, a Tabela 13 apresenta os resultados para acurácia e kappa do classificador dividido por área do conhecimento. É possível perceber que todas as áreas tiveram acurácia acima de 90% e índice Kappa acima de 0.54, tendo a área de Ciências Agrárias atingido 0.69. Mais uma vez os resultados Kappa atingem um nível acima dos resultados aleatórios, considerado concordância moderada (LANDIS; KOCH, 1977). Esses resultados mostram que é possível utilizar o classificador proposto com a essa abordagem dos dados.

Tabela 13 – Acurácia e Kappa do classificador por área do conhecimento. Fonte: o autor (2022)

	Biológicas	Engenharias	Cie. Sociais	Cie. Exatas	Humanas	Agrárias
Acurácia	0,9432860	0,9020501	0,9307465	0,9303148	0,9544217	0,9553571
Kappa	0,5941582	0,5521091	0,5413974	0,6055057	0,5737286	0,6909492

5.2.3 Questão de Pesquisa 3

Nessa seção são disponibilizadas as respostas para a questão de pesquisa 3 relacionadas ao Estudo 2. Devido ao conjunto e extensão dos dados, os resultados são apresentados em formato de gráfico para uma melhor comparação entre suas importâncias a fim de facilitar o entendimento. Dessa forma, a Figura 11 apresenta a importância das características do conjunto de dados no resultado do classificador.

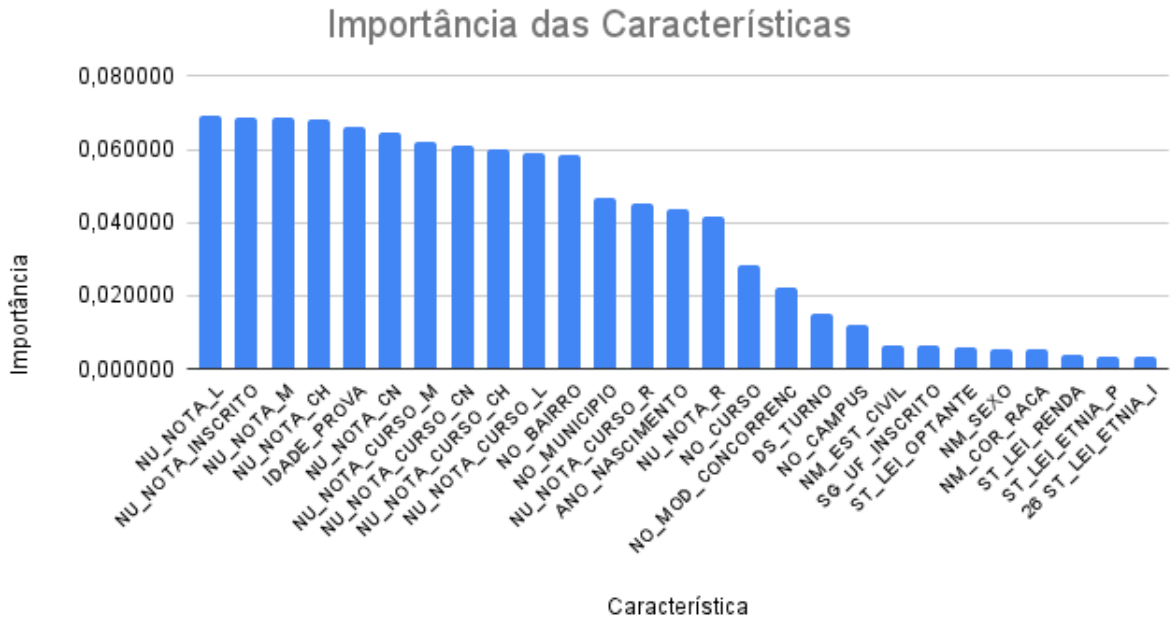


Figura 11 – Importâncias das características para o estudo 2. Fonte: o autor (2022)

5.2.4 Discussões da Questão de Pesquisa 3

A partir da Figura 11, é possível identificar que as 15 primeiras características de maior importância para o classificador são todas relativas às notas do ENEM, inclui-se também características de informação da idade do aluno na prova, cidade, bairro e ano de

nascimento.

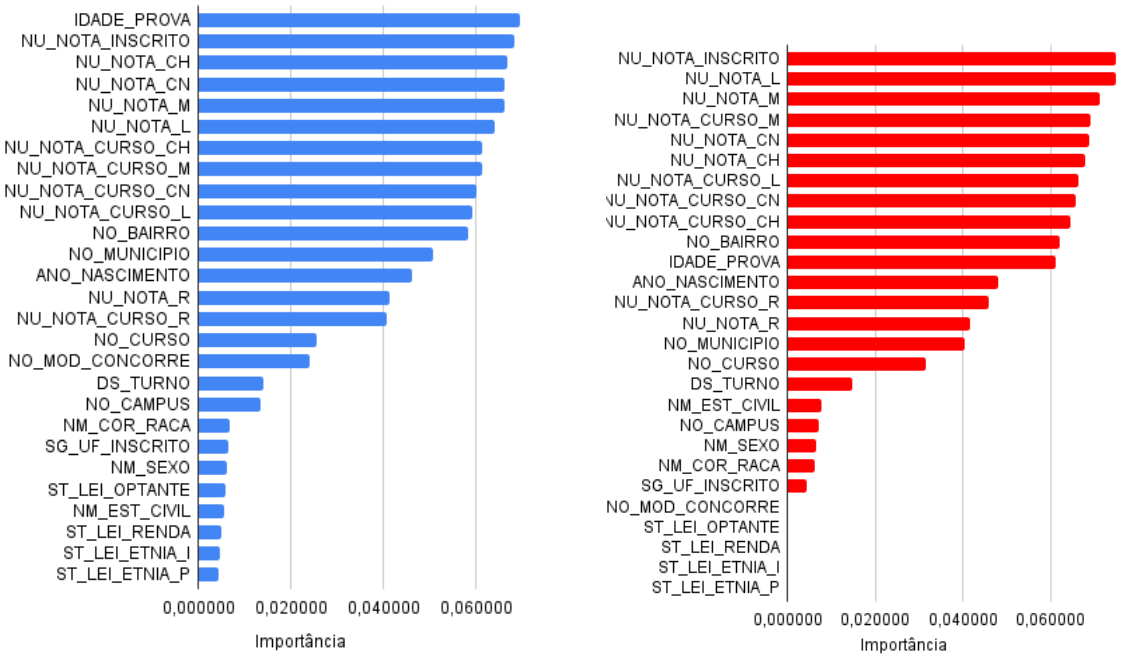
Outra característica importante é idade_prova, pois essa supera todas as outras que não são características de notas, se equiparando a valores parecidos às notas. A partir desse momento, fez-se uma consulta no grupo de evadidos dos valores de Média, Mediana e Moda da idade_prova. O que gerou os seguintes valores para a característica idade_prova: *Média de 23 anos, Mediana de 20 anos e Moda de 18 anos*. Tendo esses valores, destaca-se o valor de moda. Pois sendo a moda o valor que mais se repete no grupo de evasão, o grupo de 18 anos foi considerado como o grupo de maior risco e probabilidade de evadir-se. Vale ressaltar que 18 anos é uma idade em que as pessoas estão começando o contato com o meio acadêmico de universidade. Ou seja, a universidade é um ambiente totalmente novo para esse grupo.

Além disso, faz-se uma inferência muito importante no gráfico. O grupo das características de baixa importância estão as características demográficas, curso escolhido, turno, campus assim como também estão as características do SISU relativos à modalidade de concorrência (tipo de cota). Ou seja, para o algoritmo considerando esse conjunto de dados, a evasão não está relacionada a essas características socialmente colocadas como críticas na determinação da evasão. Tendo a cota indígena, parda ou a cota de renda consideradas como as últimas de importância para classificar um aluno. Para explicar esse resultado é importante destacar que a universidade que está sendo analisada tem uma forte política de bolsas de permanência para alunos cotistas. Isso pode estar influenciando na importância dessas características. Para qualquer inferência sobre esse ponto é importante realizar uma análise mais detalhada sobre o assunto.

Para finalizar a primeira etapa da análise, também foram avaliadas as características mais importantes para os grupos de cotistas e não cotistas. Alunos que concorreram à modalidade Ampla Concorrência foram agrupados como *Não Cotistas*. Os demais que escolheram alguma lei de cota, seja de renda, escola pública ou étnico foram agrupados como *Cotistas*. A Figura 12 apresenta a lista de características ordenadas pela importância.

É possível perceber que em ambos os casos as características mais importantes são similares, com uma diferença para a posição da idade e da cidade, que tem uma importância maior para os cotistas. Também é possível observar que a modalidade de concorrência, característica que determina as cotas, não foi relevante para os não cotistas.

Por fim, foi feita uma análise para avaliar a importância das características das



(a) Cotista (b) Não cotista

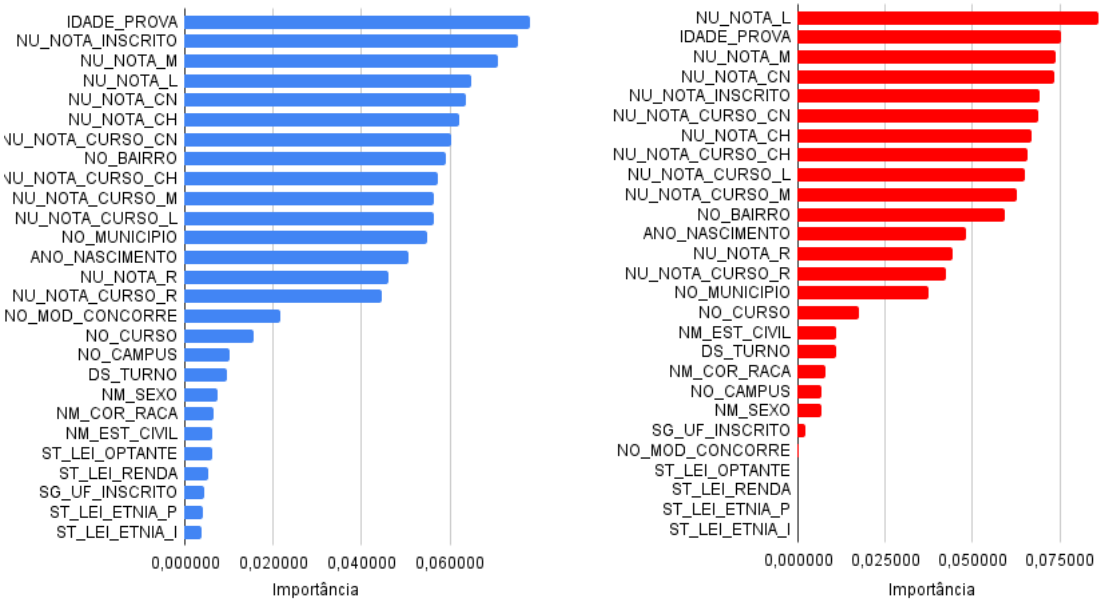
Figura 12 – Análise das características mais importantes para alunos cotistas e não cotistas. Fonte: o autor (2022)

áreas que possuem mais cursos com maior evasão e cursos com menor evasão, o quais foram identificados, dentro da base do sistema acadêmico, como sendo as áreas de exatas e a área de humanas, respectivamente. As Figuras 13 e 14 apresentam uma maior tendência de diferença entre a mesma característica entre o grupo cotista e não cotista.

Na análise da área de Ciências Exatas (Figura 13), segue a mesma tendência do gráfico da Figura 12. Para ambos os grupos, as mesmas características étnicas e de renda tiveram pouca relevância. Mostrando, assim, que as cotas ou a ausência delas, não impactam diretamente na classificação da evasão diante deste aspecto. A idade na prova, as notas do SISU e o bairro seguem como características importantes. Para o gráfico da área de Ciências Humanas (Figura 14), segue também o padrão de importância da figura 12. Nesse, todavia, o município adquire importância maior para os não cotistas, e a idade na prova maior para os cotistas. Outra diferença, já esperada, entre as áreas é que as notas em matemática e linguística/redação são mais importantes para os cursos de exatas e humanas, respectivamente.

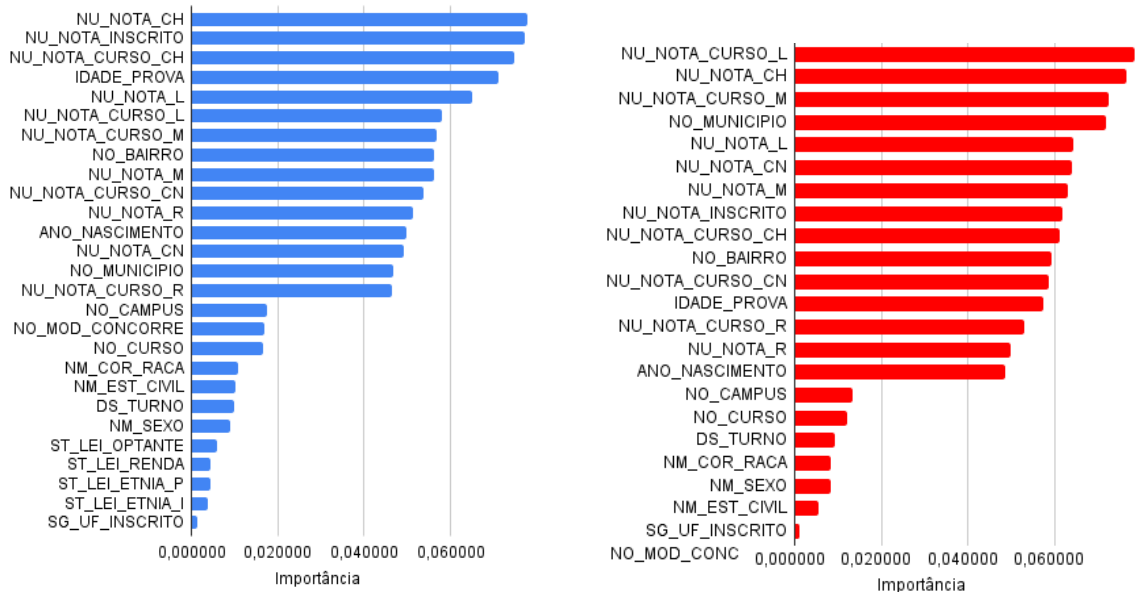
Com isso, para esse classificador e com base nos resultados apresentados, pode-se tirar a conclusão que:

1. A modalidade de concorrência, ou melhor, o tipo de cota escolhida pelo aluno para



(a) Cotista (b) Não cotista

Figura 13 – Análise das características mais importantes para alunos cotistas e não cotistas - área de exatas. Fonte: o autor (2022)



(a) Cotista (b) Não cotista

Figura 14 – Análise das características mais importantes para alunos cotistas e não cotistas - área de humanas. Fonte: o autor (2022)

o ingresso, quer seja renda, deficiente ou outra cota é pouco relevante para prever a evasão, quando se observa os dados dos alunos associados aos dados do SISU.

2. A autodeterminação da etnia, seja pardo ou indígena, ou a cota de renda também tiveram pouca relevância na classificação, quando se observa os dados associados

ao SISU, mantendo-se a idade na prova e as notas do SISU entre as maiores na determinação da evasão.

Os resultados desses dois estudos mostram que as características possuem determinada importância quando associadas a determinados conjuntos de dados, ou seja, podem adquirir ou perder importância quando interagidas com outras características. De maneira geral, para esses dois estudos, percebemos que os elementos de cor, raça e etnia se projetam como importantes na determinação da evasão quando associados aos dados demográficos das cidades brasileiras, ao passo que adquire pouca importância quando associados aos dados de notas no SISU.

Essas conclusões entram em sintonia com as políticas afirmativas de cotas raciais nas universidades públicas. Existem diversos estudos que atuam nesse tópico analisando desde aspectos jurídicos, sociais e históricos (SILVA, 2013). Os resultados desse estudo trazem, através da mineração de dados, outro aspecto que pode apoiar trabalhos nesse tema. Por fim, os resultados mostram aos gestores universitários uma maior necessidade de atenção ao ingressante jovem de idade próxima à maioridade civil, uma vez que possui, segundo o classificador, maior peso para a evasão, independente da cota associada.

6 Considerações Finais

A presente pesquisa apresentou uma análise de dados entre bases de dados em busca de fatores que pudessem predizer a evasão, de forma a ajudar os gestores universitários na busca de melhores resultados nesse tema. Para isso, foram utilizados conceitos de Mineração de Dados Educacionais. Dessa forma, foram feitos dois estudos em bases de dados, sendo elas do censo demográfico de 2010 e do sistema de seleção unificada (SISU), ambas associadas aos dados do sistema acadêmico da Universidade Federal Rural de Pernambuco.

Inicialmente, para o primeiro estudo, foram coletados os dados dessas bases. Uma, acadêmica, da própria universidade e outra, demográfica, de índices socioeconômicos, de renda, educacionais, sanitários e de vulnerabilidade dos municípios brasileiros. Esses dados foram agrupados por áreas de conhecimento dos cursos e submetidos ao algoritmo Random Forest. O objetivo foi identificar fatores que pudessem predizer a evasão. Os dados foram disponibilizados e analisados seus principais pontos de destaque, o que sugere que a característica de cor e raça está relacionada diretamente à evasão escolar, características demográficas possuem relevância moderada, não sendo fundamentais na evasão final, assim como demais características pessoais da pesquisa em questão não demostrou fundamental atuação na evasão escolar.

Para o segundo estudo, foram extraídos dados de duas bases de dados, uma da própria universidade estudada, em que se obteve dados pessoais dos estudantes e uma outra do sistema de seleção universitária governamental, de onde foram extraídos dados de seleção como notas e cotas sociais desses estudantes. Esses dados foram tratados e submetidos ao Random Forest, algoritmo classificador baseado em árvores de decisão, para que pudessem ser verificados padrões nos dados e analisá-los segundo o contexto social e econômico do local. Como resultado, obteve-se uma ordem de importância das características desses dados em que se observa diversos pontos. Entre alguns pontos que se pode destacar a importância do bairro para a classificação da evasão, a cota social tendo baixa relevância e a idade no momento da prova se projetando como importante.

Esses resultados mostram que o modelo de classificação possui grande eficácia na determinação da evasão e que pode ser usado pelos gestores para se analisar dados a fim de obter melhores decisões de gestão assim como elaborar melhores estratégias que a

diminuam a evasão.

A fim de divulgar o tema e a pesquisa, foi então submetido e aprovado até o momento o seguinte artigo em conferência e periódico da área de Ciência da Computação:

- BRITO, Bruno Claudino Pereira de ; MELLO, Rafael Ferreira Leite de; ALVES, Gabriel. Identificação de Atributos Relevantes na Evasão no Ensino Superior Público Brasileiro. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31. , 2020, Online. **Anais**[...]. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 1032-1041. DOI: https://doi.org/10.5753/cbie.sbie.2020.1032.

6.1 Limitações e Trabalhos Futuros

O presente trabalho teve algumas limitações. Entre algumas limitações, pode-se destacar a classificação ser realizada apenas por um classificador. Decerto, o classificador escolhido possui grande eficácia cientificamente comprovada na classificação, contudo, utilizando-se outros classificadores poderia ser feito uma comparação das eficácias.

Outra limitação se constitui na quantidade de características trabalhadas. Diferente de outras pesquisas, nessa foram coletadas mais características que nos trabalhos relacionados, contudo, não houve nenhuma característica relacionada ao auxílio ou apoio financeiro recebido pelo aluno ou características relacionadas ao desempenho acadêmico do aluno no decorrer do curso, constituindo, assim, uma perspectiva de trabalho futuro.

Entre algumas das perspectivas de continuidade desse estudo está a de analisar outro conjunto de características, entre elas características mais acadêmicas, como notas, faltas, número de matrículas no semestre, ou até mesmo número de reprovações a fim de determinar a importância que cada atributo possui no processo de evasão. Outra perspectiva se constitui, também, em analisar outro conjunto de dados sociais como dados de auxílio financeiro recebidos pelos alunos, problemas pessoais como desemprego ou renda familiar.

Adicionalmente, outra perspectiva de trabalho está em colocar outras medidas avaliativas do modelo, de forma manter consistente os resultados e suas discussões, como o F-score, Precisão e AUC (Area Under the ROC Curve), entre outros, assim como utilizar algoritmos automáticos de seleção de atributos e redução da dimensionalidade dos dados.

Referências

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: **Proceedings of the 20th International Conference on Very Large Data Bases**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (VLDB '94), p. 487–499. ISBN 1-55860-153-8. Disponível em: <http://dl.acm.org/citation.cfm?id=645920.672836>.

BACH, C.; ALESSA, A. Data mining and knowledge management for marketing. **International Journal of Innovation and Scientific Research**, v. 2, p. 321–328, 06 2014.

BARBOSA, A.; FERREIRA, M.; Ferreira Mello, R.; Dueire Lins, R.; GASEVIC, D. The impact of automatic text translation on classification of online discussions for social and cognitive presences. In: SCHEFFEL, M.; DOWELL, N.; JOKSIMOVIC, S.; SIEMENS, G. (Ed.). **LAK21 Conference Proceedings - The Impact we Make: The contributions of learning analytics to learning**. United States of America: Association for Computing Machinery (ACM), 2021. p. 77–87. International Learning Analytics amp; Knowledge Conference 2021, LAK 2021 ; Conference date: 12-04-2021 Through 16-04-2021. Disponível em: <https://www.solaresearch.org/events/lak/lak21/,https://dl.acm.org/doi/proceedings/10.1145/3448139>.

BARRO, R. J. Economic Growth in a Cross Section of Countries*. **The Quarterly Journal of Economics**, v. 106, n. 2, p. 407–443, 05 1991. ISSN 0033-5533. Disponível em: <https://doi.org/10.2307/2937943>.

BELGIU, M.; TOMLJENOVIC, I.; LAMPOLTSHAMMER, T. J.; BLASCHKE, T.; HöFLE, B. Ontology-based classification of building types detected from airborne laser scanning data. **Remote Sensing**, v. 6, n. 2, p. 1347–1366, 2014. ISSN 2072-4292. Disponível em: <https://www.mdpi.com/2072-4292/6/2/1347>.

BONIFRO, F. D.; GABBRIELLI, M.; LISANTI, G.; ZINGARO, S. P. Student dropout prediction. In: BITTENCOURT, I. I.; CUKUROVA, M.; MULDNER, K.; LUCKIN, R.; MILLÁN, E. (Ed.). **Artificial Intelligence in Education**. Cham: Springer International Publishing, 2020. p. 129–140. ISBN 978-3-030-52237-7.

BRAMER, M. **Principles of Data Mining**. 3. ed. London: Springer, 2013. (Undergraduate Topics in Computer Science). ISSN 1863-7310. ISBN 978-1-4471-7306-9.

BREIMAN, L. Random forests. **Machine Learning**, Kluwer Academic Publishers, v. 45, n. 1, p. 5–32, 2001. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A%3A1010933404324>.

CAVALCANTI, A. P.; DIEGO, A.; MELLO, R. F.; MANGAROSKA, K.; NASCIMENTO, A. C. A.; FREITAS, F. L. G. de; GAEVIĆ, D. How good is my feedback?: a content analysis of written feedback. **Proceedings of the Tenth International Conference on Learning Analytics & Knowledge**, 2020.

CHAYM, L. D. e C. Evasão universitária: Um modelo para diagnóstico e gerenciamento de instituições de ensino superior. **Revista de Administração IMED**, v. 9, n. 1, p.

167–186, 2019. ISSN 2237-7956. Disponível em: <https://seer.imed.edu.br/index.php/raimed/article/view/3198>.

COELHO, P. S. de S. Data mining: Algumas questões epistemológicas. Simpósio de Pesquisa Operacional e Logística da Marinha, Rio de Janeiro – RJ, 2006. Disponível em: <https://www.marinha.mil.br/spolm/sites/www.marinha.mil.br.spolm/files/arq0007_0.pdf>.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37, 1960.

CORADINE, L. C.; LOPES, R. V. V.; MACIEL, A. F. Mineração de dados: Uma introdução. **Learning Nonlinear Models**, SBRN, v. 9, n. 3, p. 168–184, 2011.

DEGENHARDT, F.; SEIFERT, S.; SZYMCZAK, S. Evaluation of variable selection methods for random forests and omics data sets. **Briefings in Bioinformatics**, v. 20, n. 2, p. 492–503, 10 2017. ISSN 1477-4054. Disponível em: <https://doi.org/10.1093/bib/bbx124>.

DELGADO, M. F. e. a. "do we need hundreds of classifiers to solve real world classification problems?". **J. Machine Learning Research**, Wiley Online Library, v. 15, n. 1, p. 3133 – 3181, 2014.

DIAS, E. C. M.; THEÓPHILO, C. R.; LOPES, M. A. S. Evasão no ensino superior: Estudo dos fatores causadores da evasão no curso de ciências contábeis da universidade estadual de montes claros – unimontes – mg. **Anais do 7º Congresso USP de Iniciação Científica em Contabilidade**, 2010.

DURSO, S.; CUNHA, J. Determinant factors for undergraduate student's dropout in an accounting studies department of a brazilian public university. **Educação em Revista**, v. 34, 05 2018.

FERREIRA, R.; KOVANOVIĆ, V.; GAŠEVIĆ, D.; ROLIM, V. Towards combined network and text analytics of student discourse in online discussions. In: ROSÉ, C.; MARTINEZ-MALDONADO, R.; HOPPE, H.; LUCKIN, R.; MAVRIKIS, M.; PORAYSKA-POMSTA, K.; MCLAREN, B.; du Boulay, B. (Ed.). **Artificial Intelligence in Education**. Springer, 2018. (Lecture Notes in Artificial Intelligence), p. 111–126. ISBN 9783319938424. International Conference on Artificial Intelligence in Education 2018, AIED 2018 ; Conference date: 27-06-2018 Through 30-06-2018. Disponível em: <https://link-springer-com.ezproxy.lib.monash.edu.au/book/10.1007/978-3-319-93843-1,https://link.springer.com/book/10.1007/978-3-319-93846-2>.

FILHO, R. L. L. e. S.; MOTEJUNAS, P. R.; HIPóLITO, O.; LOBO, M. B. C. M. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641–659, jun. 2013. Disponível em: <http://publicacoes.fcc.org.br/index.php/cp/article/view/346>.

GARCIA, F. C.; SANTIAGO, E. F. B. Mecanismo de enfretamento à evasão no ensino superior público: Inserção do conteúdo sobre profissões no ensino médio. **Revista Gestão Pública Práticas e Desafios**, v. 6, n. 01, 2015.

GOLDSCHMIDT, E. P. R. Data mining: um guia prático. Revista Eletrônica de Sistemas de Informação, v. 5, n. 1, p. 1–2, 2006.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. [S.l.]: Elsevier, 2006.

HO, T. K. Random decision forests. In: **Proceedings of 3rd International Conference on Document Analysis and Recognition**. [S.l.: s.n.], 1995. v. 1, p. 278–282 vol.1.

KAUFMAN, L.; ROUSSEEUW, P. **Finding Groups in Data: An Introduction to Cluster Analysis**. [S.l.: s.n.], 2009. ISBN 9780470317488.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, JSTOR, p. 159–174, Mar 1977. Doi: <10.2307/2529310>.

LIAW, A.; WIENER, M. Classification and regression by randomforest. **R News: The Newsletter of the R Project**, v. 2, n. 3, p. 18–22, 2002. Disponível em: <http://cran.r-project.org/doc/Rnews/>.

LOBO, M. B. de C. M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. **Cadernos ABMES**, n. 25, 2012. Disponível em: <https://abmes.org.br/arquivos/publicacoes/Cadernos25.pdf>.

LUCAS, R. On the mechanics of economic development. **Journal of Monetary Economics**, v. 22, n. 1, p. 3–42, 1988. Disponível em: <https://EconPapers.repec.org/RePEc:eee:moneco:v:22:y:1988:i:1:p:3-42>.

MANKIW, N. G.; ROMER, D.; WEIL, D. N. A Contribution to the Empirics of Economic Growth. **The Quarterly Journal of Economics**, v. 107, n. 2, p. 407–437, 1992. Disponível em: <https://ideas.repec.org/a/oup/qjecon/v107y1992i2p407-437..html>.

MCHUGH, M. L. Interrater reliability: the kappa statistic. Biochemia medica, v. 22, 2012.

MEC. **Informe estatístico do MEC revela melhoria do rendimento escolar**. 1998. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/informe-estatistico-do-mec-revela-melhoria-do-rendimento-escolar>. Acesso em: 24 agosto 2021.

MENZE, B.; KELM, B.; MASUCH, R.; HIMMELREICH, U.; BACHERT, P.; PETRICH, W.; HAMPRECHT, F. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. **BMC bioinformatics**, v. 10, p. 213, 08 2009.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2.

SANCHES, M. K. Aprendizado de máquina semi-supervisionado: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados. Instituto de Ciências Matemáticas e de Computação, São Paulo – SP, 2003. Disponível em: <https://teses.usp.br/teses/disponiveis/55/55134/tde-12102003-140536/pt-br.php>.

SCHUHA, G.; REINHARTB, G.; PROTEA, J.-P.; SAUERMANNA, F.; HORSTHOFERB, J.; OPPOLZERA, F.; KNOLL, D. Data mining definitions and applications for the management of production complexity. **52nd CIRP Conference on Manufacturing Systems**, v. 81, p. 874–879, 2019.

SETTLES, B. **Active Learning Literature Survey**. [S.l.], 2009. Disponível em: <http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf>.

SILVA, R. de S.; PAES Ângela T. Teste de concordância kappa. Educ Contin Saúde einstein, v. 10, 2012.

SILVA, R. M. F. D. Ações afirmativas e direito fundamental à educação: uma análise à luz das cotas raciais universitárias. Revista Jurídica da Presidência, v. 14, n. 104, 2013.

TORABI, M.; UDZIR, N. I.; ABDULLAH, M. T.; YAAKOB, R. A review on feature selection and ensemble techniques for intrusion detection system. **International Journal of Advanced Computer Science and Applications**, The Science and Information Organization, v. 12, n. 5, 2021. Disponível em: <http://dx.doi.org/10.14569/IJACSA.2021.0120566>.