UNIVERSIDADE DE PASSO FUNDO

Programa de Pós-Graduação em Computação Aplicada

Dissertação de Mestrado

SYSDAE - USANDO MACHINE LEARNING NUMA ANÁLISE DO PERFIL COMPORTAMENTAL DOS ALUNOS DO ENSINO MÉDIO DO IFRS - CÂMPUS SERTÃO

CEDEMIR PEREIRA

UPF

UNIVERSIDADE DE PASSO FUNDO

INSTITUTO DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

SYSDAE - USANDO MACHINE LEARNING NUMA ANÁLISE DO PERFIL COMPORTAMENTAL DOS ALUNOS DO ENSINO MÉDIO DO IFRS - CÂMPUS SERTÃO

Cedemir Pereira

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Computação Aplicada na Universidade de Passo Fundo.

Orientador: Roberto dos Santos Rabello

Passo Fundo

2024

CIP – Catalogação na Publicação

P436s Pereira, Cedemir
SYSDAE [recurso eletrônico] : usando Machine Learning numa analise do perfil comportamental dos alunos do ensino médio do IFRS - Campus Sertão / Cedemir Pereira. – 2024.
2.3 MB ; PDF.

Orientador: Prof. Dr. Roberto dos Santos Rabello.
Dissertação (Mestrado em Computação Aplicada) – Universidade de Passo Fundo, 2024.

1. Inteligência artificial. 2. Machine learning. 3. Estudantes (Ensino médio) - Instituto Federal do Rio Grande do Sul (Campus Sertão) - Comportamento. 4. Algorítmos. I. Rabello, Roberto dos Santos, orientador. II. Título.

CDU: 004.8

Catalogação: Bibliotecária Juliana Langaro Silveira – CRB 10/2427

ATA DE DEFESA DO
TRABALHO DE CONCLUSÃO DE CURSO DO ACADÊMICO

CEDEMIR PEREIRA

Aos trinta dias do mês de agosto do ano de dois mil e vinte e quatro, às dezessete horas, realizou-se, de forma presencial na sala 110 do ITec (Prédio B5). A sessão pública de defesa do Trabalho de Conclusão de Curso "SYSDAE - USANDO MACHINE LEARNING NUMA ANÁLISE DO PERFIL COMPORTAMENTAL DOS ALUNOS DO ENSINO MÉDIO DO IFRS - CÂMPUS SERTÃO", de autoria de Cedemir Pereira, acadêmico do Curso de Mestrado em Computação Aplicada do Programa de Pós-Graduação em Computação Aplicada – PPGCA. Segundo as informações prestadas pelo Conselho de Pós-Graduação e constantes nos arquivos da Secretaria do PPGCA, o aluno preencheu os requisitos necessários para submeter seu trabalho à avaliação. A banca examinadora foi composta pelo professor doutor Carlos Amaral Hölbig e a professora doutora Anubis Graciela de Moraes Rossetto. Concluídos os trabalhos de apresentação e arguição, a banca examinadora considerou o candidato **APROVADO**. Foi concedido o prazo de até quarenta e cinco (45) dias, conforme Regimento do PPGCA, para o acadêmico apresentar ao Conselho de Pós-Graduação o trabalho em sua redação definitiva, a fim de que sejam feitos os encaminhamentos necessários à emissão do Diploma de Mestre em Computação Aplicada. Para constar, foi lavrada a presente ata, que vai assinada pelos membros da banca examinadora e pela Coordenação do PPGCA.

Documento assinado digitalmente
gov.br ROBERTO DOS SANTOS RABELLO
Data: 30/08/2024 21:10:36-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Roberto Rabello - UPF
Presidente da Banca Examinadora
(Orientador)

Documento assinado digitalmente
gov.br CARLOS AMARAL HOLBIG
Data: 02/09/2024 11:28:03-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Carlos Amaral Hölbig - UPF
(Coordenador do PPGCA)

Documento assinado digitalmente
gov.br CARLOS AMARAL HOLBIG
Data: 02/09/2024 11:26:55-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Carlos amaral Hölbig - UPF
(Avaliador Interno)

Documento assinado digitalmente
gov.br ANUBIS GRACIELA DE MORAES ROSSETTO
Data: 02/09/2024 10:32:48-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Anubis Graciela de Moraes
Rossetto - IFSUL
(Avaliadora Externa)

SYSDAE - USANDO MACHINE LEARNING NUMA ANÁLISE DO PERFIL COMPORTAMENTAL DOS ALUNOS DO ENSINO MÉDIO DO IFRS - CÂMPUS SERTÃO

RESUMO

Neste trabalho objetivou-se prever o risco de perda da residência estudantil de alunos do ensino médio, empregando inteligência artificial, a qual utilizou na área de Machine Learning, os algoritmos de aprendizado de máquina. A análise baseia-se em diversos dados relacionados á residência estudantil, incluindo informações sobre o aluno, curso, número das horas orientadas, medida, setor, número de dias suspenso, sexo, série e matrícula para gerar um sistema preditivo. O risco de perder a residência estudantil consiste no fato de o aluno não poder permanecer mais como residente junto ao câmpus, ocasionando diversas dificuldades, como residir fora dos limites da instituição, ter que arcar com aluguel, realidade não contemplada para muitos alunos de nossa instituição. O ato de continuar residente é um reflexo do padrão de comportamento do aluno, é tido como o resultado do modelo. Três algoritmos de aprendizagem de máquina diferentes foram considerados para identificar e classificar os parâmetros que afetam a permanência na residência estudantil: Naive-Bayes, KNN, Árvore de Decisão. Para avaliar o desempenho dos algoritmos de aprendizagem de máquina, três métricas foram utilizadas: precisão, recall e F1-score. Os resultados indicam que o KNN superou as demais técnicas, gerando resultados superiores, seguido pelo Naive-Bayes. Se conclui que há como prever o risco de perda da residência estudantil a partir de um padrão de comportamento dos alunos. Uma aplicação web foi desenvolvida para a apresentação de resultados. Segundo o experimento realizado, a abordagem mostrou-se adequada e pode servir como tomada de decisão em ações que visem o melhor relacionamento entre a Residência Estudantil e os alunos, além de melhorar as habilidades sociais dos estudantes, necessárias para o convívio entre pares.

Palavras-chave: inteligência artificial, machine learning, residência estudantil, sistemas preditivos.

SYSDAE - USING MACHINE LEARNING IN AN ANALYSIS OF THE BEHAVIORAL PROFILE OF IFRS HIGH SCHOOL STUDENTS - CÂMPUS SERTÃO

ABSTRACT

This study aimed to predict the risk of high school students losing their student residence using artificial intelligence, which used machine learning algorithms in the area of Machine Learning. The analysis is based on several data related to the student residence, including information about the student, course, number of hours taught, grade, sector, number of days suspended, gender, grade and enrollment to generate a predictive system. The risk of losing the student residence consists of the fact that the student can no longer remain as a resident on campus, causing several difficulties, such as living outside the institution's boundaries and having to pay rent, a reality not contemplated by many students at our institution. The act of remaining a resident is a reflection of the student's behavior pattern and is considered the result of the model. Three different machine learning algorithms were considered to identify and classify the parameters that affect the permanence in the student residence: Naive-Bayes, KNN, Decision Tree. To evaluate the performance of the machine learning algorithms, three metrics were used: precision, recall and F1-score. The results indicate that KNN outperformed the other techniques, generating superior results, followed by Naive-Bayes. It is concluded that it is possible to predict the risk of losing the student residence based on a pattern of student behavior. A web application was developed to present the results. According to the experiment carried out, the approach proved to be adequate and can be used to make decisions in actions aimed at improving the relationship between the Student Residence and the students, in addition to improving the social skills of students, which are necessary for coexistence among peers.

Keywords: artificial intelligence, machine learning, student residence, predictive systems.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

DAE	DEPARTAMENTO DE ASSISTÊNCIA ESTUDANTIL
HCA	HIERARCHICAL CLUSTER ANALYSIS
IA	INTELIGÊNCIA ARTIFICIAL
IFRS	INSTITUTO FEDERAL DO RIO GRANDE DO SUL
KNN	K-NEAREST NEIGHBORS
LLE	LOCALLY LINEAR EMBEDDING
LLM	LARGE LANGUAGE MODE
LSTM	LONG SHORT-TERM MEMORY
MEC	MINISTÉRIO DA EDUCAÇÃO
MLA	MACHINE LEARNING ALGORITHM
PA	PREDICTIVE ANALYTICS
PCA	PRINCIPAL COMPONENT ANALYSIS
PNAES	PROGRAMA NACIONAL DE ASSISTÊNCIA ESTUDANTIL
PROEJA	PROGRAMA NACIONAL DE INTEGRAÇÃO DA EDUCAÇÃO PROFISSIONAL COM A EDUCAÇÃO BÁSICA NA MODALIDADE DE EDUCAÇÃO DE JOVENS E ADULTOS
RE	RESIDÊNCIA ESTUDANTIL
TDAH	TRANSTORNO DE DÉFICIT DE ATENÇÃO E HIPERATIVIDADE
TI	TECNOLOGIA DE INFORMAÇÃO
t-SNE	T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING

SUMÁRIO

1. INTRODUÇÃO

O Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) é uma instituição federal de ensino público e gratuito, que se propõe a fornecer ensino humanizado, crítico e cidadão. Oferece cursos gratuitos em 16 municípios gaúchos, sendo estes, cursos de nível médio (técnicos que podem ser cursados de forma integrada, concomitante e subsequente ao Ensino Médio), superiores (de graduação e pós-graduação) e de extensão. No total, são aproximadamente 22.200 alunos e 217 opções de cursos. Tem aproximadamente 1.192 professores e 918 técnicos administrativos.

Dentre os diversos setores presentes no IFRS - Câmpus Sertão, destaca-se o Departamento de Assistência Estudantil (DAE); o mesmo que atende desde alunos com dificuldades pedagógicas, passando também por estudantes com dificuldades psicológicas específicas – como transtornos de ansiedade, TDAH (Transtorno de Déficit de Atenção e Hiperatividade), transtorno depressivo, entre outros. Trabalha também com o desenvolvimento de habilidades necessárias para convívio coletivo, especialmente na Residência Estudantil (RE), conforme o Regulamento de Conduta para estudantes residentes e semiresidentes.

O Instituto Federal do Rio Grande do Sul (IFRS) - Câmpus Sertão oferta residência aos acadêmicos desde que estes habitem em cidades distantes. A residência é o setor responsável pelo gerenciamento desses educandos, que optam ou necessitam se instalar na área interna do Câmpus, promovendo a segurança e o bem-estar, além de tranquilidade aos pais ou responsáveis. Na Coordenadoria de Residência Estudantil são armazenadas e gerenciadas as informações dos alunos. Atualmente este controle é feito manualmente, e todos os formulários de registros são criados em programas de edição de texto, e a cada novo registro é refeito o trabalho de digitar os documentos, com novas informações.

Quando um interno descumpre qualquer regra estabelecida pelo Regulamento de Conduta, o coordenador pode adverti-lo verbalmente ou, se julgar necessário, registrar a ocorrência formalmente por meio de um documento chamado “Atas de Ocorrências da Residência Estudantil”. Nesse documento têm-se os dados escolares do interno, o número do seu quarto, a descrição da infração cometida com maiores detalhes possíveis, a referência no Regulamento que dará respaldo a

advertência e o nome do servidor que registrou o Ato de Indisciplina, e ainda, campos para assinatura do servidor e dos estudantes responsáveis pelo episódio. No Conselho disciplinar são julgadas todas as ocorrências dos alunos e se, o mesmo for reincidente, este poderá perder o direito ao uso da residência estudantil.

O sistema acadêmico utilizado no câmpus não contempla um módulo para a Residência Estudantil, ficando restrito ao cadastro de alunos, suas avaliações e processos estudantis realizados na Coordenadoria de Registros Escolares.

Destaca-se também que tarefas preditivas são ações voltadas para a previsão, que é encontrar uma função, modelo ou hipótese que pode ser utilizada para previsão de uma ocorrência. Por exemplo, digamos que possa-se querer prever um valor de um imóvel ou o possível estado de saúde de um paciente após 5 meses da aplicação de uma cirurgia, podendo este estar doente ou saudável, junto a previsão tem-se uma entrada (geralmente representada por X) e uma saída (geralmente representada por Y).

Este estudo propõe o desenvolvimento de um sistema informatizado utilizando *Machine Learning* para prever o risco de perda da residência estudantil com base em padrões comportamentais dos discentes. A relevância deste trabalho reside na melhoria da gestão da residência estudantil e na garantia da permanência dos estudantes no câmpus. Dessa forma, o problema de pesquisa busca responder a seguinte indagação: “É possível, através do desenvolvimento de um sistema que utilize *Machine Learning*, prever, a partir do padrão de comportamento de alunos, o risco de perder a residência estudantil?”. Frente a isto, no decorrer do presente estudo será desenvolvido de forma mais ampla os aspectos que embasaram a pesquisa, perpassando as instâncias da fundamentação teórica, da coleta de informações e o desenvolvimento do projeto propriamente dito.

O objetivo geral deste trabalho consiste em desenvolver um sistema preditivo do risco de perder a residência estudantil para os alunos do IFRS - Câmpus Sertão, utilizando algotimos de Machine Learning, visando apoiar a tomada de decisões no gerencimento da residência estudantil. Além disso, tem-se como objetivos específicos os seguintes:

I. Implementar um sistema de gestão para controlar os dados das Atas de Ocorrências.

II. Utilizar algoritmos de *Machine Learning* para identificar padrões de comportamento que possam indicar risco de perda da residência estudantil.
III. Avaliar a eficácia dos MLAs(algoritmos de aprendizado de máquina) na previsão do risco de perda da residência.
IV. Propor recomendações para a melhoria da gestão da residência estudantil com base nos resultados obtidos.

A estrutura do trabalho está organizado com os seguintes capítulos:

- Capítulo 2: Descreve a Fundamentação teórica com conceitos sobre *Machine Learning* e seus usos. *Algoritmos de Machine Learning* utilizados e tipos de aprendizado de máquina. Conceitos de Análise Preditiva e Metodologias de Validação mais comuns. No final do capítulo apresenta os trabalhos relacionados.

- Capítulo 3: Apresenta os Materiais e métodos. A utilização das técnicas de inteligência artificial aplicados sobre *Machine Learning* na previsão da perda da residência estudantil resultam em 25 trabalhos selecionados na revisão sistemática que são brevemente apresentados. Mostra as métricas utilizadas.

- Capítulo 4: Descreve o uso do SYSDAE. O desenvolvimento da aplicação, suas telas, o cálculo de risco do KNN, variáveis de entrada e saída, principais rotinas.

- Capítulo 5: Detalha e discute os resultados encontrados a partir dos modelos de *Machine Learning.* Nesse capítulo é descrito a análise e descrição dos resultados obtidos.

2. FUNDAMENTAÇÃO TEÓRICA

Apresenta-se a seguir um texto introdutório do Câmpus Sertão, seguido dos tópicos Departamento de Assistência Estudantil(DAE), Machine Learning, Análise Preditiva, por fim, Trabalhos Relacionados.

2.1. O IFRS – CÂMPUS SERTÃO

O Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) foi criado em 29 de dezembro de 2008, pela Lei 11.892, que instituiu, no total, 38 Institutos Federais de Educação, Ciência e Tecnologia no país. Por força de lei, o IFRS é uma autarquia federal vinculada ao Ministério da Educação (MEC). Goza de prerrogativas com autonomia administrativa, patrimonial, financeira, didático-científica e disciplinar. Pertence à Rede Federal de Educação Profissional e Tecnológica.

O Câmpus Sertão do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul está situado no Distrito Engenheiro Luiz Englert, município de Sertão, a 25 quilômetros de Passo Fundo, região Norte do Estado do Rio Grande do Sul. A instituição funciona em período integral, com aulas teóricas e práticas, nos turnos da manhã, tarde e noite, incluindo ainda, outras atividades para atendimento da comunidade externa, como cursos de curta duração, que visam à atualização, capacitação e treinamento em áreas diversas e formações de qualificação.

O IFRS – Câmpus Sertão oferece os seguintes cursos técnicos integrados a nível médio: Manutenção e Suporte em Informática, Agropecuária e o Técnico em Comércio do PROEJA. Havendo também o curso técnico pós-médio em agropecuária (subsequente). Disponibilizando a nível superior os cursos de Bacharelado em Agronomia, Bacharelado em Zootecnia, Licenciatura em Biologia, Tecnologia em Análise e Desenvolvimento de Sistemas, Tecnologia em Agronegócio, Tecnologia em Gestão Ambiental e Formação Pedagógica para Graduados não Licenciados. E em nível de Pós-Graduação oferta os cursos em modalidade *Lato sensu* em Teorias e Metodologias da Educação e, Desenvolvimento e Inovação.

2.2. O DEPARTAMENTO DE ASSISTÊNCIA ESTUDANTIL

O Departamento de Assistência Estudantil (DAE) é responsável por planejar e desenvolver a política de assistência estudantil do Câmpus Sertão, de acordo com as diretrizes do Programa Nacional de Assistência Estudantil - PNAES. Na prática, beneficiar os alunos envolve ações que promovam moradia, saúde, alimentação, acompanhamento, acolhimento, transporte e tantos outros itens importantes na vida dos estudantes. A assistência estudantil está dividida basicamente em ações universais, que apoiam todos os estudantes matriculados, sem distinção, e o Programa de Benefícios. Iniciativas voltadas à equidade de oportunidades e melhoria de condições socioeconômicas, tendo como público específico acadêmicos que preencham critérios de vulnerabilidade social, envolvendo o auxílio permanência e auxílio-moradia, por exemplo.

O DAE possui atualmente uma coordenadoria e está subdividido em atendimento geral, restaurante, atendimento psicossocial e pedagógico, ambulatório, residência estudantil e lavanderia. A equipe de atendimento geral de apoio envolve uma telefonista e três auxiliares administrativos.

Na residência estudantil, voltada aos alunos do ensino técnico integrado ao ensino médio, há uma coordenadora, dois assistentes de alunos, um vigilante e um zelador. A equipe atende e dá suporte aos estudantes residentes e semirresidentes.

Abaixo destaca-se as definições de residentes semirresidentes:

I. Estudante Residente: estudante regularmente matriculado nos cursos técnicos integrados ao ensino médio, com frequência às aulas e que esteja morando na área interna da instituição a esse fim destinada, com direito a pernoite e que concorrem à vaga na residência estudantil através de edital específico.

II. Estudante Semirresidente: estudante regularmente matriculado nos cursos técnicos integrados ao ensino médio, com frequência às aulas e que utilize o espaço a este fim destinado, sem direito a pernoite.

De acordo com os dados da Coordenação de Residência Estudantil, ocorre na instituição um total de 145 residentes do sexo masculino e 82 residentes do sexo feminino, havendo 167 alunos semirresidentes.

O DAE também auxilia com o desenvolvimento das habilidades necessárias para convívio coletivo, especialmente na Residência Estudantil. Com relação aos cuidados com a saúde da comunidade acadêmica, o departamento oferece um ambulatório para atendimentos básicos em saúde e conta, em sua equipe, com uma técnica de enfermagem, uma enfermeira, uma médica e uma dentista. São realizados atendimentos de baixa complexidade e encaminhamentos de casos de maior complexidade para órgãos de saúde externos. O DAE desempenha um importante papel institucional, sendo uma rede de apoio multiprofissional no suporte aos acadêmicos em seu processo de ingresso, permanência e êxito estudantil.

Além da rede de atendimento, também há o Programa de Benefícios, que colabora imensamente com a permanência de estudantes em situação de vulnerabilidade social através de auxílios estudantis, hoje divididos em duas categorias: o auxílio permanência e o auxílio-moradia.

Segundo os dados da Assistente Social, no Auxílio Moradia do Técnico em Agropecuária Integrado têm-se 11 alunos beneficiados, e no PROEJA, 2 estudantes. Já o Auxílio Permanência beneficia 174 alunos, onde destes, 18 são do PROEJA.

De acordo com o Regulamento de Conduta para Estudantes Residentes e Semirresidentes[18], em seu capítulo VII, é proibido aos estudantes:

> I – Usar, portar ou depositar dentro das dependências da Residência Estudantil quaisquer substâncias psicoativas consideradas lícitas, como bebidas ou cigarros, ou consideradas ilícitas pela legislação penal;
>
> II - Guardar ou utilizar qualquer espécie de arma, inclusive réplicas de brinquedo;
>
> III - Utilizar indevidamente substâncias inflamáveis, explosivas de qualquer natureza que represente perigo para si e para a comunidade escolar;
>
> IV - Namorar no interior dos apartamentos;
>
> V – Os estudantes do sexo masculino entrar ou permanecer nos apartamentos das estudantes do sexo feminino e vice-versa;
>
> VI - Levar para as dependências da Residência Estudantil pessoas estranhas ou não autorizadas pelo DAE. As pessoas autorizadas

pelo DAE estarão sob a responsabilidade dos estudantes do apartamento visitado, não podendo ultrapassar às 22h;

VII - Guardar ou trafegar com veículos bicicletas, motos, skates, patins, ou outros similares nas dependências da Residência Estudantil, sem a devida autorização;

VIII - Riscar, pintar e/ou colar quaisquer materiais, assim como colocar pregos, parafusos ou similares, nas portas, paredes e camas, interna ou externamente;

IX - Levar para a Residência Estudantil qualquer espécie de animal ou vegetais ornamentais, sem a devida autorização;

X - Mudar de cama, armário ou quarto sem a devida autorização;

XI - Permanecer na Residência Estudantil nos horários de aula, exceto em caso de doença diagnosticada pelo serviço de saúde ou com autorização do DAE;

XII - Perturbar o repouso noturno das 22h às 7h;

XIII - Promover reuniões, festas ou encontros com barulho excessivo, em qualquer horário, sem a devida autorização. Não será permitido volume alto, em nenhum horário do dia, caso isso aconteça, os/as estudantes do apartamento responsável ficarão proibidos de utilizar aparelho de som e este aparelho será recolhido no DAE e no final de semana o proprietário deverá levá-lo para casa.

XIV - Utilizar os equipamentos de combate a incêndio para outros fins que não sejam os de segurança;

XV - Instalar fogões, fogareiros, fornos a gás ou elétricos (micro-ondas), máquinas de lavar e secadoras de roupas, torneiras elétricas, impressoras, freezer e panelas, cafeteiras, torradeiras, fritadeiras elétricas, estufas ou aquecedores, roteadores;

XVI - Depositar lixo fora dos locais apropriados;

XVII - Promover jogos que envolvam dinheiro;

XVIII - Comercializar qualquer tipo de produto;

XIX - Os/as estudantes residentes e semirresidentes só poderão frequentar os apartamentos de outros estudantes com a autorização dos moradores do apartamento;

> XX - Apresentar conduta desrespeitosa e/ou violenta com colegas e servidores, incluindo a prática de bullying.
> Paragrafo único: É permitido o uso dos seguintes eletroeletrônicos: secadores, chapinhas de cabelo, barbeadores, carregadores de celular e notebooks, desde que utilizados nos ambientes autorizados e de forma correta.

Quanto as tarefas da Comissão Disciplinar, ainda segundo o Regulamento de Conduta para Estudantes Residentes e Semirresidentes[18], assim foi definido:

> Art. 27. A Comissão Disciplinar da Residência Estudantil é o colegiado responsável pela avaliação disciplinar e da aplicação das ações pedagógicas aos estudantes em regime de residência ou semirresidência, em consonância com o Regulamento de Direitos e Deveres dos Estudantes, previsto na Organização Didático-Pedagógica, sendo composta por:
> I – Coordenador(a)-Geral do DAE, que a presidirá;
> II – Coordenador(a) da Residência Estudantil, que exercerá a vice-presidência;
> III – 01 (um) membro da APS - Associação de Pais e Servidores;
> IV – 01 (um) Técnico em Assuntos Educacionais;
> V – 01 (um) representante da Direção de Ensino;
> VI - coordenações dos cursos técnicos;
> VII – 01 (um) assistente de alunos;
>
> Art. 28. A Comissão Disciplinar da Residência Estudantil obedecerá aos seguintes fluxos:
> I - Análise dos registros de ocorrência;
> II - Análise dos encaminhamentos pedagógicos efetuados pelos setores responsáveis;
> III - Convocação de pessoas para esclarecimento dos registros, quando necessário;

IV – Registro escrito da versão alegada pelos envolvidos, como forma de apresentação de defesa;
V – Comunicação escrita com convocação para comparecimento dos pais e/ou responsáveis para ciência da situação;
VI - Convocação de profissional de atendimento especializado (pedagogo/a, psicólogo/a, assistente social, enfermeiro/a, técnico/a em assuntos educacionais, nutricionista ou outro servidor que tenha prestado atendimento especializado ao (s) estudante(s) envolvido(s) cujo parecer se avalie pertinente para o caso;
VII - Proposição de encaminhamentos de ações pedagógicas, processos administrativos e/ou civis, conforme exigência do caso;
VIII - Retorno aos envolvidos, mediante parecer.
Parágrafo único: Todas as discussões e encaminhamentos efetuados pela Comissão Disciplinar da Residência Estudantil deverão seguir a legislação vigente e observar o zelo para com o tratamento das informações, objetivando preservar a dignidade, evitar a exposição desnecessária dos envolvidos e garantir o direito ao contraditório e à ampla defesa.

Art. 29. Dos Recursos:

I - Os responsáveis legais terão o prazo de 48h após a entrega da medida disciplinar atribuída ao estudante pela Comissão Disciplinar para protocolar formulário de recurso junto ao Gabinete da Direção Geral, caso queiram recorrer da decisão da referida Comissão.
II - O Diretor Geral no uso de suas atribuições legais terá o prazo de 5(cinco) dias úteis para emitir parecer sobre o recurso.

2.3. MACHINE LEARNING

Machine Learning é um ramo da inteligência artificial (IA) e da ciência da computação que se concentra no uso de dados e algoritmos para imitar a maneira como os humanos aprendem, melhorando gradualmente sua precisão [1].

Machine Learning é um componente importante do crescente campo da ciência de dados. Por meio do uso de métodos estatísticos, os algoritmos são treinados para fazer classificações ou previsões, revelando os principais *insights* em projetos de mineração de dados. Esses *insights* subsequentemente conduzem a tomada de decisões em aplicativos e negócios, impactando de forma ideal as principais métricas de crescimento. Conforme o *big data* continua a se expandir e crescer, a demanda do mercado por cientistas de dados aumenta, exigindo que eles auxiliem na identificação das questões de negócios mais relevantes e, posteriormente, os dados para respondê-las.

Segundo Bonnin [1] , *Machine Learning* é definido da seguinte forma: "Diz-se que um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho nas tarefas em T, medido por P, melhora com experiência E.”

Esta definição restabelece os elementos que desempenham um papel em cada projeto de aprendizagem de máquina: a tarefa a realizar, as sucessivas experiências e, uma clara e apropriada medida de desempenho. Em palavras mais simples, temos um programa que melhora como executa uma tarefa baseada sobre experiência e guiada por um certo critério.

De acordo com Netto e Maciel [2], *Machine Learning* (Aprendizado de Máquina) é a disciplina que faz uso de toda uma série de procedimentos e algoritmos para identificar padrões, agrupamentos ou tendências e, então, extrair informações úteis para a análise de dados, de maneira totalmente automatizada. Grosso modo, pode-se dizer que são métodos matemáticos usados para treinar algoritmos que identificam padrões.

Neste estudo, fixou-se em três algoritmos de *Machine Learning*: Naive-Byes, KNN e Árvore de Decisão. Cada um com características distintas que os tornaram adequados para a tarefa de previsão de risco da residência estudantil. A seguir descreve-se os algoritmos:

I. Naive-Bayes: Um classificador probabilístico baseado no teorema de Bayes, que assume independência entre os atributos. É eficaz para tarefas de classificação de texto e problemas com alta dimensionalidade de dados.
II. KNN (K-Nearest Neighbors): Um algoritmo de classificação que atribui a classe de um novo exemplo com base na classe

majoritária de seus vizinhos mais próximos. É simples e intuitivo, adequado para problemas onde a proximidade dos dados é um fator importante.

III. Árvore de Decisão: Um modelo de aprendizado supervisionado que utiliza uma estrutura de árvore para tomar decisões baseadas nos valores dos atributos dos dados. É conhecido por sua interpretabilidade e capacidade de capturar relações não lineares entre as variáveis.

2.3.1. Funcionamento do Machine Learning

Segundo Bonnin [1], pode-se dividir o sistema de aprendizado de um algoritmo de *Machine Learning* em três partes principais:

I. Um processo de decisão: Em geral, algoritmos de *Machine Learning* são usados para fazer uma predição ou classificação. Com base em alguns dados de entrada, que podem ser rotulados ou não rotulados, seu algoritmo produzirá uma estimativa sobre um padrão nos dados.

II. Uma função de erro: Uma função de erro serve para avaliar a predição do modelo. Se houver exemplos conhecidos, uma função de erro poderá fazer uma comparação para avaliar a precisão do modelo.

III. Um processo de otimização de modelo: Se o modelo pode se ajustar melhor aos pontos de dados no conjunto de treinamento, então os pesos são ajustados para reduzir a discrepância entre o exemplo conhecido e a estimativa do modelo. O algoritmo repetirá este processo de avaliação e otimização, atualizando os pesos de maneira autônoma até que um limite de precisão seja atingido.

De acordo com Bonnin [1], o aprendizado de máquina como disciplina não é um campo isolado – isso é emoldurado dentro de um domínio mais amplo, Inteligência Artificial (IA). Mas, como pode-se imaginar, o aprendizado de máquina não apareceu do nada. Como disciplina tem seus predecessores, e vem evoluindo

em estágios de aumentar a complexidade em quatro etapas claramente diferenciadas, descritas a seguir:

I. O primeiro modelo de aprendizado de máquina envolvia decisões baseadas em regras e um simples nível de algoritmos baseados em dados que os inclui em si, e como pré-requisito, todos as possíveis ramificações e regras de decisão, implicando que todas as ramificações possíveis sejam codificadas no modelo de antemão por um especialista na área. Essa estrutura foi implementada na maioria das aplicações desenvolvidas desde as primeiras linguagens de programação que apareceram em 1950. O principal tipo de dado e função tratados por esse tipo de algoritmo é o Booleano, pois ele tratou exclusivamente de decisões sim ou não.

II. Durante o segundo estágio de desenvolvimento do raciocínio estatístico, começou-se a constatar que as características probabilísticas dos dados têm uma palavra a dizer, além das escolhas anteriores previamente definidas. Isso reflete melhor a natureza difusa dos problemas do mundo real, onde pontos fora da curva são comuns e onde é mais importante levar em consideração uma conta às tendências não determinísticas dos dados, do que a abordagem rígida de questões fixas. Essa disciplina adiciona ao mix de elementos de ferramentas matemática da teoria da probabilidade Bayesiana. Os métodos pertencentes a esta categoria incluem ajuste de curva (geralmente linear ou polinomial), os quais têm a propriedade comum de trabalhar com dados numéricos.

III. O estágio de aprendizado de máquina é o domínio no qual foi trabalhado ao longo deste trabalho, e envolve tarefas mais complexas do que os mais simples elementos Bayesianos do estágio anterior.

    a. A característica mais notável dos algoritmos de aprendizado de máquina é a de que eles podem generalizar modelos a partir de dados, mas os modelos são capazes de gerar seus próprios seletores de recursos,

que não são limitados por uma função de destino rígida, pois são gerados e definidos à medida que o processo de treinamento evolui. Outro diferencial de esse tipo de modelo é que eles podem receber uma grande variedade de tipos de dados como entrada, como fala, imagens, vídeo, texto e outros dados suscetíveis de serem representados como vetores.

IV. A IA é o último passo na escala de capacidades de abstração que, de certa forma, incluem todos os tipos de algoritmos anteriores, mas com uma diferença fundamental: os algoritmos de IA são capazes de aplicar o conhecimento aprendido para resolver tarefas que nunca foram consideradas durante o treinamento. Os tipos de dados com os quais este algoritmo trabalha são ainda mais genéricos do que os tipos de dados suportados pelo aprendizado de máquina, e eles devem ser capazes, por definição, de transferir capacidades de resolução de problemas de um tipo de dados para outro, sem um retreinamento completo do modelo. Desta forma, poder-se-ia desenvolver um algoritmo para detecção de objetos em imagens em preto e branco e o modelo abstrato poderia abstrair o conhecimento para aplicar o modelo para colorir imagens.

A Figura 1 representa esses quatro estágios de desenvolvimento em direção à IA real.

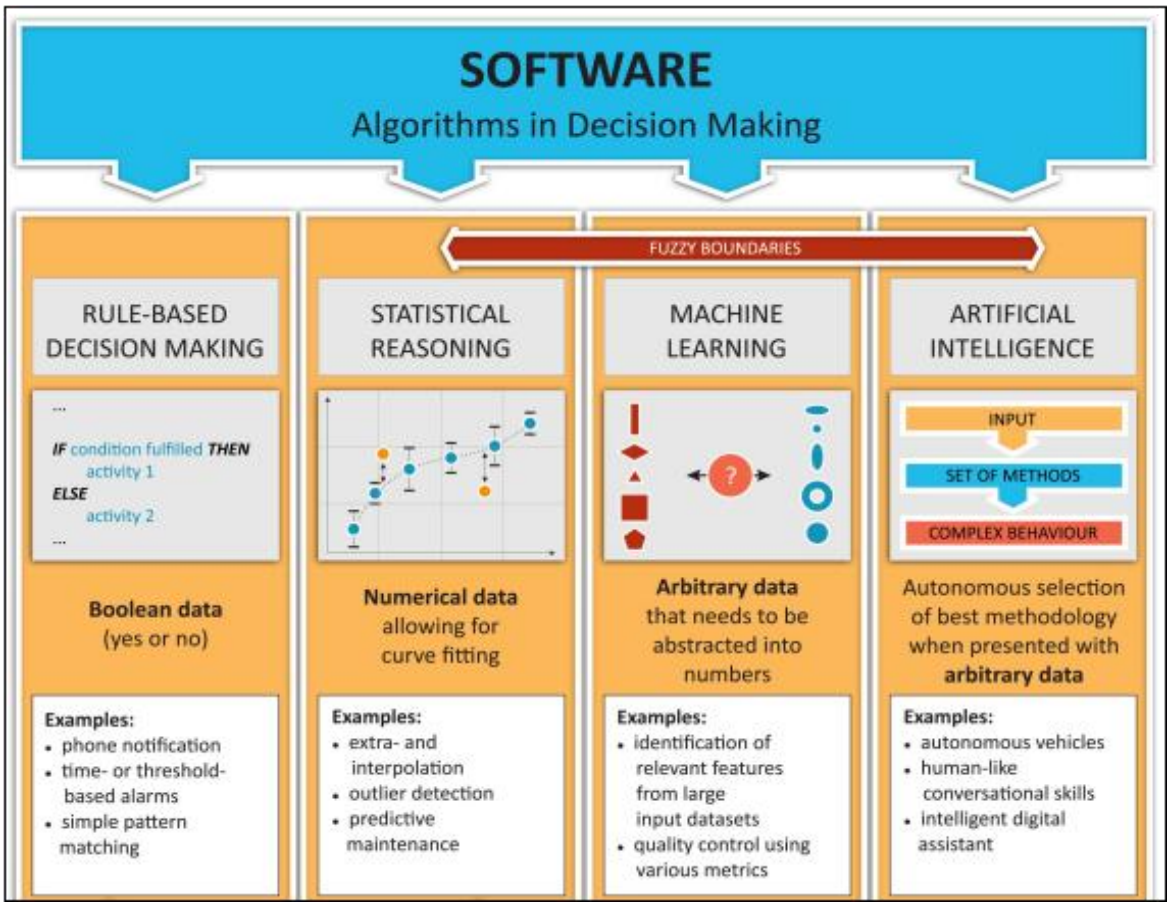


Figura 1. Os algoritmos na tomada de decisão [1].

De acordo com Campesato [3] , em termos de alto nível, o aprendizado de máquina é um subconjunto da IA que pode resolver tarefas inviáveis ou muito complicadas com linguagens de programação mais tradicionais. Um filtro de spam para e-mail é um dos primeiros exemplos de aprendizado de máquina. O aprendizado de máquina geralmente substitui a precisão dos algoritmos mais antigos.

Apesar da variedade de algoritmos de aprendizado de máquina, os dados são indiscutivelmente mais importantes do que o algoritmo selecionado. Muitos problemas podem surgir com os dados, como informações insuficientes, baixa qualidade dos mesmos, dados incorretos, ausentes, irrelevantes, valores de dados duplicados e assim por diante.

2.3.2. Tipos de aprendizado de máquina e algoritmos

Abaixo busca-se elencar os diferentes tipos de projeto de aprendizado de máquina, conforme explica Bonnin [1], começando pelo grau de conhecimento prévio do ponto de vista do implementador. O projeto pode ser dos seguintes tipos:

2.3.2.1. Aprendizagem supervisionada

Este tipo de aprendizagem recebe um conjunto de amostras de dados, acompanhado do resultado que o modelo deve informar depois de aplicá-lo. Em termos estatísticos, tem-se os resultados de todos os conjuntos de experimentos do treinamento.

2.3.2.2. Aprendizagem não supervisionada

Esse tipo de aprendizagem fornece apenas os dados de amostras do domínio do problema, além da tarefa de agrupar dados semelhantes e aplicar uma categoria que não tem informação prévia das quais ela pode ser inferida.

2.3.2.3. Aprendizagem por reforço

Esse tipo de aprendizagem não possui um conjunto de amostras rotuladas e tem um número diferente de elementos participantes, que incluem um agente, um ambiente, um aprendizado e uma política ótima ou um conjunto de etapas, maximizando uma abordagem orientada a objetivos, usando recompensas ou penalidades (o resultado de cada tentativa).

Destaca-se a figura 2, onde vemos as principais áreas do Machine Learning.

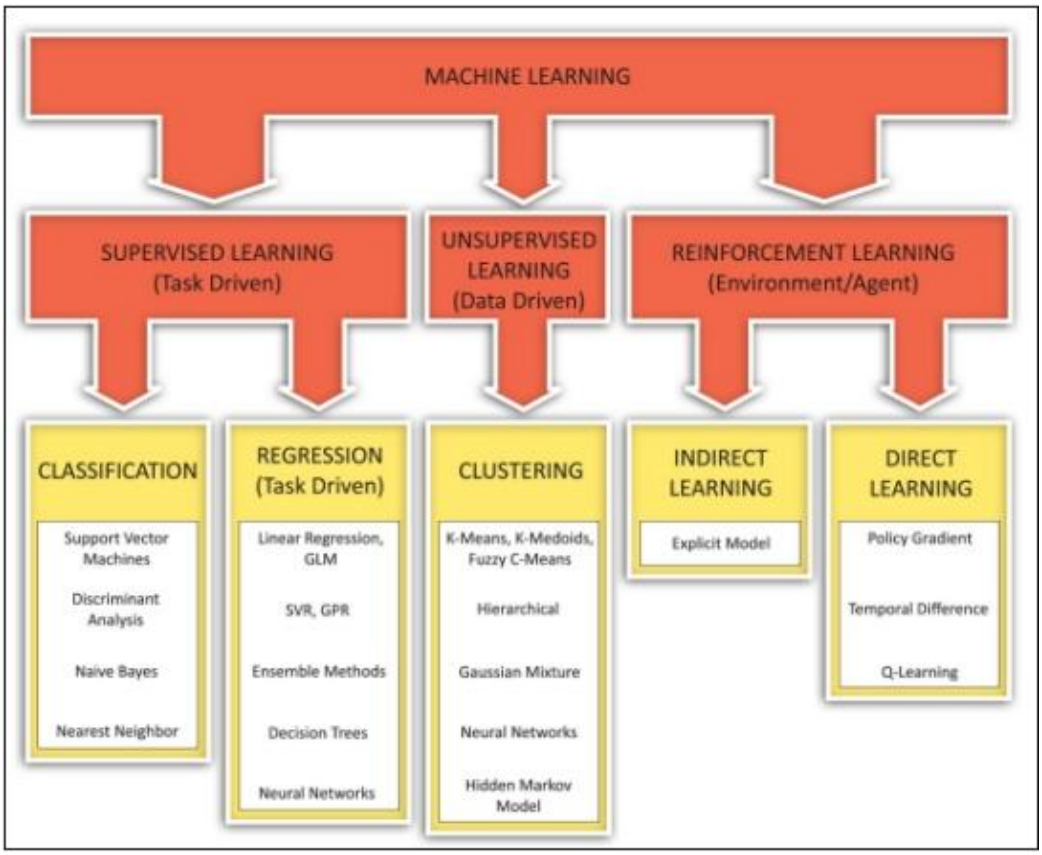


Figura 2. Principais áreas do *Machine Learning* [1]

Campesato [3] também elucida sobre os tipos de aprendizado. De acordo com o autor, aprendizado supervisionado significa que os pontos de dados em um conjunto de dados possuem um rótulo que identifica seu conteúdo. Por exemplo, o conjunto de dados MNIST contém arquivos PNG 28x28, cada um contendo um único dígito desenhado à mão (ou seja, 0 a 9 inclusive). Toda imagem com o dígito 0 tem o rótulo 0; toda imagem com o dígito 1 tem o rótulo 1; todas as outras imagens são rotuladas de acordo com o dígito exibido nessas imagens. Como outro exemplo, as colunas no conjunto de dados do Titanic são características dos passageiros, como sexo, classe da cabine, preço da passagem, se o passageiro sobreviveu ou não e assim por diante. Cada linha contém informações sobre um único passageiro, incluindo o valor 1 se o passageiro sobreviveu. O conjunto de dados MNIST e o conjunto de dados Titanic envolvem tarefas de classificação: o objetivo é treinar um

modelo com base em um conjunto de dados de treinamento e, em seguida, prever a classe de cada linha em um conjunto de dados de teste.

Em geral, os conjuntos de dados para tarefas de classificação têm um pequeno número de valores possíveis: um dos nove dígitos no intervalo de 0 a 9, um dos quatro animais (cachorro, gato, cavalo, girafa), um dos dois valores (sobreviveu versus pereceu, comprado versus não comprado). Como regra geral, se o número de saídas puder ser exibido em um número relativamente pequeno de valores (que é um número subjetivo) em uma lista *drop-down*, provavelmente é uma tarefa de classificação.

No caso de um conjunto de dados que contém dados imobiliários, cada linha contém informações sobre uma casa específica, como número de quartos, metros quadrados da casa, número de banheiros, preço da casa e assim por diante. Neste conjunto de dados, o preço da casa é o rótulo de cada linha. Observa-se que a faixa de preços possíveis é muito grande para encaixar razoavelmente bem em uma lista *drop-down*. Um conjunto de dados imobiliários envolve uma tarefa de regressão: o objetivo é treinar um modelo com base em um conjunto de dados de treinamento e, em seguida, prever o preço de cada casa em um conjunto de dados de teste.

Por sua vez, o aprendizado não supervisionado envolve dados não rotulados, o que normalmente é o caso de algoritmos de agrupamento (discutidos posteriormente). Alguns algoritmos importantes de aprendizado não supervisionado que envolvem agrupamento são os seguintes:

I. k-means;
II. Análise de cluster hierárquico (HCA);
III. Maximização da expectativa.

Alguns algoritmos importantes de aprendizado não supervisionado que envolvem redução de dimensionalidade são os seguintes:

I. Análise de componentes principais (PCA);
II. kernel PCA;
III. Incorporação linear localmente (LLE);
IV. Incorporação estocástica de vizinhança t-distribuída (t-SNE).

Outra tarefa não supervisionada de também importância chama-se detecção de anomalias. Sua relevância está na detecção de fraudes e detecção de *outliers*.

O aprendizado semisupervisionado é uma combinação de aprendizado supervisionado e não supervisionado: alguns pontos de dados são rotulados e outros não rotulados. Uma técnica envolve o uso de dados rotulados para classificar (ou seja, rotular) os dados não rotulados, após o que você pode aplicar um algoritmo de classificação.

Outrossim conforme Netto e Maciel [2], os algoritmos de aprendizagem supervisionada podem ser de Classificação ou de Regressão. Algoritmos de Classificação tratam de problemas em que os dados têm uma classificação prévia e se deseja prever a qual categoria um dado não classificado pertence. Por exemplo, o problema da classificação dos e-mails legítimos e dos spams. Nesse caso, treinamos o algoritmo de classificação para identificar quais são as características de um e-mail legítimo e quais as de um spam. Treinado, o algoritmo pode testar se um e-mail (ou um conjunto de palavras) pertence a uma das duas classes.

Regressão é utilizada se um ou mais atributos são variáveis contínuas (reais) e dependentes. Ou seja, a mudança em uma variável implica a mudança da outra. O exemplo claro aqui é o do investimento em pesquisa, administração e propaganda nas cinquenta startups. O conjunto é formado por variáveis numéricas - reais e contínuas (em contraste as numéricas discretas ou nominais). Pode-se dizer que algoritmos de classificação atribuem rótulos aos dados, também chamados de labels, classificando-os, como o nome indica, em grupos. Diz-se que os algoritmos de classificação produzem como saída um atributo classe ou meta. Algoritmos de regressão, por sua vez, produzem valores, que tentam prever alguma característica numérica dos dados.

2.3.2.4. Algoritmos de regressão

São cinco os tipos de algoritmos de regressão mais comumente estudados em *Machine Learning*: Linear, Polinomial, Árvore de Decisão (Decision Tree) e Floresta Randômica ou Floresta Aleatória (Random Forest).

Utiliza-se como exemplo a seguinte pergunta: “Existe uma relação entre antiguidade na empresa e salário recebido?”. Esse tipo de algoritmo tenta estabelecer uma relação entre duas variáveis (no exemplo, tempo de serviço em anos e o salário recebido). Ou em outras palavras: É possível prever o salário a

partir do tempo de serviço? Ou mesmo se pode inferir quantos anos de serviço um funcionário tem a partir do salário?

Uma maneira visual de entender o conceito por trás da regressão é montar um gráfico com as variáveis em dois eixos e marcar os pontos correspondentes. A regressão linear é a maneira pela qual consegue-se encontrar uma reta (linha) que ligue todos os pontos, ou, mais precisamente, uma reta que tenha a menor distância possível dos pontos.

A Tabela 1 representa parte dos dados utilizados na Regressão..

Tabela 1. Dados utilizados na Regressão.

Anos de experiência	Salário
1,1	39.343,,00
1,3	46.205,00
1,5	37.731,00
2	43.525,00
2,2	39.891,00
2,9	56.642,00
3	60.150,00
3,2	54.445,00
3,2	64.445,00
3,7	57.189,00
3,9	63.218,00
4	55.794,00
4	56.957,00
4,1	57.081,00
4,5	61.111,00
4,9	67.938,00
5,1	66.029,00
5,3	83.088,00
5,9	81.363,00
	Continua

2.3.2.5. Algoritmos de classificação

Como abordado anteriormente, esses algoritmos são utilizados para criar partições de conjuntos, um nome pomposo para subconjuntos próprios. Como visto, são muito úteis em problemas em que uma classe (ou classes) deve ser definida. Pode haver algoritmos de classificação que usem aprendizagem supervisionada ou não supervisionada. São exemplos de algoritmos de classificação: KNN (K-Nearest

Neighbors), SVM (Support Vector Machines), Regressão Logística (que, apesar do nome, é um algoritmo de classificação), entre outros.

Em alguns algoritmos (principalmente nos de aprendizagem supervisionada) é comum a aprendizagem ser dividida em duas etapas: treino e previsão. Usa-se parte dos dados (geralmente entre 70% e 80%) para “treinar” o algoritmo, e a outra parte, para prever os resultados, testar o algoritmo na parte restante e verificar se ele está próximo do resultado correto. Obviamente, quanto maior o tamanho dos dados, melhor será a previsão.

No entanto, o melhor algoritmo utilizado servirá muito bem para o conjunto de testes. No mundo real, onde a quantidade de dados aumenta exponencialmente, é necessário que esse algoritmo sempre seja atualizado e volte a ser testado sobre uma base de dados real. Aplica-se aqui um conceito de *cross-validation*, ou validação cruzada, que é separar as partes de treino e teste em subconjuntos distintos que abranjam todo o conjunto de dados.

2.4. ANÁLISE PREDITIVA

De acordo com Winters [4], a análise preditiva se baseia em um conceito simples: prever a probabilidade de eventos futuros com base em dados históricos. Sua história pode remontar a pelo menos 650 a.C. Alguns dos primeiros exemplos incluem os babilônios, que tentaram prever mudanças climáticas de curto prazo com base no aparecimento de nuvens e halos.

A medicina também tem uma longa história de necessidade de classificar as doenças. O rei babilônico Adad Apla-iddina decretou que os registros médicos fossem coletados para formar um Manual de Diagnóstico. Algumas previsões neste *corpus* listam tratamentos com base no número de dias em que o paciente esteve doente e em sua pulsação. Uma das primeiras instâncias da bioinformática.

Posteriormente, a análise preditiva especializada foi desenvolvida no início das indústrias de subscrição de seguros. Isso foi usado como uma forma de prever o risco associado ao seguro de embarcações marítimas [5]. Mais ou menos na mesma época, as seguradoras de vida começaram a prever a idade que uma pessoa viveria para definir as taxas de prêmio mais adequadas.

Embora a ideia de previsão sempre parecesse enraizada desde cedo na necessidade humana de entender e classificar, não firmou-se até o século 20, e foi com o advento da computação moderna, que ela realmente se consolidou.

Junto com a previsão, veio uma maior compreensão de causa e efeito e como as várias partes do problema estavam inter-relacionadas. A descoberta e o *insight* surgiram por meio da metodologia e da adesão ao método científico. Mais importante, eles surgiram para encontrar soluções para problemas importantes e muitas vezes práticos da época. Foi isso que os tornou únicos.

2.4.1. O que a análise preditiva não faz

A análise preditiva não informa o que acontecerá no futuro, mas sim a criação de modelos preditivos que colocam um valor numérico, ou pontuação, na probabilidade de um determinado evento acontecer no futuro com um nível aceitável de confiabilidade, e inclui cenários hipotéticos e avaliação de riscos.

2.4.2. Por que análise preditiva?

Na área de inteligência de negócios, com a plataforma correta de gerenciamento de operações, os tomadores de decisão são capazes de gerenciar todas as entradas, eventos e dados relacionados aos negócios que fornecem informações em tempo real para o nível corporativo. Posteriormente, modelos preditivos podem ser usados para identificar padrões úteis de dados históricos, transacionais e recentes para identificar riscos e oportunidades potenciais. Portanto, está ganhando muita atenção e ampla aceitação. Além disso, usando as ferramentas tradicionais de relatórios e monitoramento, pode-se passar de operações reativas para operações proativas. A Análise Preditiva (PA) ajuda a ir além disso para planejar o futuro e identificar novas áreas de negócios para lucro e produtividade.

No aprendizado de máquina, observamos o desempenho de um algoritmo em duas etapas: aprendizado e inferência. O objetivo final da etapa de aprendizado é preparar e descrever os dados disponíveis, também chamados de vetores de recursos, que são usados para treinar o modelo.

A fase de aprendizagem é uma das etapas mais importantes, mas também é verdadeiramente demorada. Envolve a preparação de uma lista de vetores também chamados de vetores de *features* (na maioria das vezes) a partir dos dados de treinamento após a transformação, para que possamos alimentá-los com os algoritmos de aprendizado. Por outro lado, os dados de treinamento às vezes também contêm informações impuras que precisam de algum pré-processamento, como limpeza.

Uma vez que temos os vetores de *Features*, o próximo passo nesta etapa é preparar (ou escrever/reutilizar) o algoritmo de aprendizado. O seguinte passo importante é treinar o algoritmo para preparar o modelo preditivo. Normalmente (e, claro, com base no tamanho dos dados), a execução de um algoritmo pode levar horas (ou até dias) para que os recursos convirjam em um modelo útil, conforme mostrado na Figura 3:

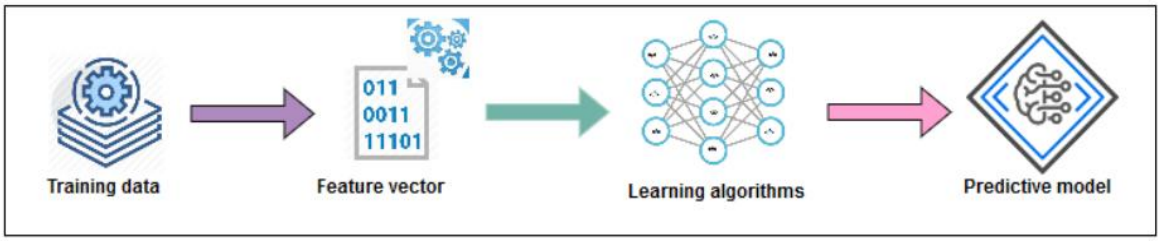


Figura 3. Aprendendo e treinando um modelo preditivo – mostra como gerar os vetores de features a partir dos dados de treinamento para treinar o algoritmo de aprendizado que produz um modelo preditivo [6].

2.4.3. Métodos comuns de análise preditiva

Os métodos comuns de análise preditiva incluem análise de regressão, classificação, previsão de séries temporais, mineração de regra de associação, agrupamento, sistemas de recomendação e mineração de texto, análise de sentimento e muito mais.

O segundo estágio mais importante é a inferência que é usada para fazer um uso inteligente do modelo, como prever a partir de dados nunca antes vistos, fazer recomendações, deduzir regras futuras e assim por diante. Normalmente, leva menos tempo em comparação com a etapa de aprendizado e, às vezes, até em tempo real, como mostra a Figura 4:

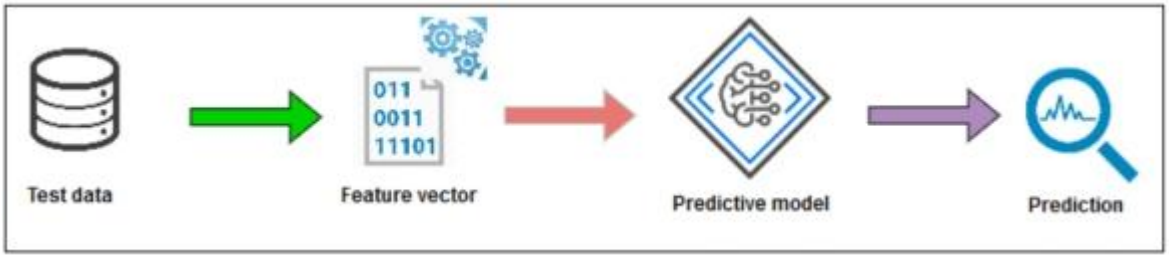


Figura 4. Inferência de um modelo existente para análise preditiva (vetores de features são gerados a partir de dados desconhecidos para fazer previsões) [6].

Assim, a inferência consiste em testar o modelo em relação a dados novos (ou seja, não observados) e avaliar o desempenho do próprio modelo. No entanto, em todo o processo e para tornar o modelo preditivo bem-sucedido, os dados atuam como cidadãos de primeira classe em todas as tarefas de aprendizado de máquina.

2.5 Metodologias de validação mais comuns

A validação por observação em *machine learning* refere-se a técnicas de validação de modelos que se baseiam na divisão dos dados disponíveis em conjuntos de treino e teste, de maneira a garantir que o modelo é avaliado com dados que ele não viu durante o treino. A seguir foram descritas algumas das metodologias de validação mais comuns:

2.5. HOLDOUT (DIVISÃO SIMPLES)

Consiste em dividir os dados em dois conjuntos, um para treino e outro para teste. A divisão típica é 70% para treino e 30% para teste. Como vantagens desta metodologia destaca-se a simplicidade para implementação e a utilidade para conjuntos de grandes dados. Já como desvantagens elenca-se o fato da possibilidade de levar a resultados enviesados dependendo de como os dados são divididos e a não utilização de todos os dados disponíveis para treino.

2.6. K-FOLD CROSS-VALIDATION

Visa a divisão dos dados em K subconjuntos (folds). O modelo é treinado K vezes, cada vez usando K-1 folds para treino e o fold restante para teste. A métrica de performance é a média dos K testes. Como vantagem, este apresenta maior robustez e menor enviesamento, utilizando todos os dados para treino e teste. E como desvantagem, trata-se de um sistema mais computacionalmente intensivo.

2.7. STRATIFIED K-FOLD CROSS-VALIDATION

Semelhante ao K-Fold, mas também garantindo que cada fold tenha a mesma proporção de classes que o conjunto de dados original. Sendo especialmente útil para dados desbalanceados. Este modelo mantém a distribuição das classes em cada fold, sendo ideal para problemas de classificação desbalanceada. E como desvantagem, trata-se de um sistema mais computacionalmente intensivo.

2.8. LEAVE-ONE-OUT CROSS-VALIDATION (LOOCV)

Consiste-se em uma versão extrema do K-Fold, onde K é igual ao número de exemplos no conjunto de dados. Cada interação usa um único exemplo como conjunto de teste e o restante como treino. Este sistema utiliza o máximo de dados possíveis para treinar o modelo em cada iteração, sendo útil para conjuntos de dados pequenos. Contudo, também é computacionalmente intensivo e pode ter alta variância nas estimativas de erro.

2.9. TIME SERIES SPLIT

Sistema específico para dados temporais. Divide os dados respeitando a ordem temporal, usando sempre os dados passados para treinar e os futuros para testar. Respeita a ordem temporal dos dados, sendo útil para previsões temporais. Ademais não é adequado para dados que não possuem ordem temporal.

Estas são algumas das metodologias mais comuns de validação por observação em *machine learning*. A escolha da metodologia depende do tipo de dados e do problema específico que está sendo resolvido. Neste trabalho foi usada a metodologia do Tipo Holdout já que dividiu-se os dados em dois conjuntos: um para treino e outro para teste. A divisão usada foi de 80% para treino e 20% para teste.

Foram realizados dois tipos de avaliações:

I. Uma de usabilidade através de um formulário onde o Coordenador do DAE e a Coordenadora da Residência Estudantil (Apêndice A) responderam a um Questionário de Avaliação de Usabilidade.
II. Outra através da funcionalidade do software onde o Coordenador do DAE e a Coordenadora da Residência Estudantil verificaram através da digitação e uso na prática, que algumas questões que poderiam ser melhoradas referentes ao uso da ferramenta na Residência Estudantil.

Foram elencados os seguintes tópicos a serem melhorados:

I. Alteração nos campos que mostram o e-mail na tabela violações, em vez de usar a tabela medidas disciplinares, visto que cadastrar muitas medidas disciplinares (são sempre cinco) confundiria ainda mais o usuário.
II. Manter o envio de e-mail.
III. Renomear o Título do Sistema para "Controle de Comportamento de Estudantes Residentes”.

Criar uma listagem de nome de estudantes por Apartamento.

2.10. TRABALHOS RELACIONADOS

Como primeiro trabalho relacionado temos o estudo desenvolvido por Rajendran, Chamundeswari e Sinha [7], o qual prevê o desempenho acadêmico de alunos do ensino fundamental e médio usando algoritmos de aprendizado de máquina (MLAs) com base em vários dados sociodemográficos (como idade, sexo, obesidade, renda familiar média, tamanho da família e estado civil dos pais), variáveis relacionadas à escola (tipo de educação e nível acadêmico) e relacionadas

ao aluno (estresse e estilo de vida). O Grade Point Average (Média de Notas), que é um reflexo do desempenho acadêmico, é considerado o resultado do modelo. Cinco MLAs diferentes são considerados para identificar e classificar os parâmetros que afetam o desempenho acadêmico: Regressão Multinomial Logística, Rede Neural Artificial, Árvore Aleatória, Aumento Gradiente e Métodos de Empilhamento. Para avaliar o desempenho dos MLAs, três métricas são utilizadas: precisão, recall e F1-score. Observa-se que o método gradient boosting superou as demais técnicas gerando resultados superiores, seguido pela random forest. A partir da análise do modelo, concluiu-se que um estilo de vida consciente saudável se correlaciona positivamente com o desempenho acadêmico, enquanto a existência de estresse tem um impacto negativo.

Outro trabalho relacionado, de Beckham *et al.* [8], no qual os alunos que enfrentavam problemas que poderiam atrapalhar suas buscas acadêmicas pelo sucesso, problemas que iriam desde questões triviais, como condição de classe, sentimentos dos alunos, até questões graves, como desagregação familiar, razões econômicas e muito mais. Este representa um grande problema visto que os estudantes moldam o futuro de uma nação – o que afetará muitas instâncias no futuro. Os professores estão procurando uma maneira eficaz de encontrar o que geralmente pode ser a melhor solução para resolver determinados problemas, pois cada discente pode enfrentar problemas diferentes, resolver um de cada vez não é possível com o número de alunos a cada ano. Neste trabalho citado, tentou-se encontrar fatores que poderiam prejudicar ou melhorar o desempenho do acadêmico usando a correlação de Pearson entre cada recurso em relação ao resultado G3 dos alunos. Com base no resultado, as reprovações passadas impactaram negativamente as notas dos alunos com correlação de -0,360415 e, então, a Educação Média impactará positivamente as notas dos alunos com 0,217147. Após descobrir qual fator afeta a nota do aluno, tentou-se prever a nota do mesmo usando modelos de MLAs para provar se esse fator realmente afeta a variável nota. O modelo MLP de 12 neurônios apresenta o melhor desempenho com valor RMSE de 4,32, seguido por Random Forest com valor RMSE de 4,52 e, finalmente, Árvore de Decisão com valor RMSE de 5,69.

Também se tem o estudo de German *et al.* [9], onde a pandemia da COVID-19 trouxe mudanças para os indivíduos, principalmente no comportamento do consumidor. Como os governos de diferentes países implementaram protocolos

de segurança para mitigar a propagação do vírus, as pessoas ficaram apreensivas em viajar e sair. Isso abriu caminho para o surgimento da logística de terceiros (3PL). As estatísticas provaram a rápida escalada em relação ao uso de 3PL em vários países. Este estudo citado utilizou a Rede Neural Artificial e o Classificador Randon Forest para validar e justificar os fatores que afetam a intenção do consumidor em selecionar um provedor de serviços 3PL durante a pandemia de COVID-19, integrando as Dimensões da Qualidade do Serviço e a Teoria Pró-Ambiental do Comportamento Planejado. Os resultados deste estudo revelaram que a atitude é o fator mais significativo que afeta a intenção comportamental dos consumidores. Outros fatores, como satisfação do cliente, valor percebido pelo cliente, preocupação ambiental percebida, garantia, capacidade de resposta, empatia, confiabilidade, tangibilidade, controle comportamental percebido, norma subjetiva e suporte de autoridade percebido, são todos fatores que contribuem para afetar a intenção comportamental. Os algoritmos de aprendizado de máquina, especificamente ANN e RFC, mostraram-se confiáveis na previsão de fatores, pois obtiveram taxas de precisão de 98,56% e 93%. Os resultados mostraram que a atitude dos consumidores, a satisfação, o valor percebido, a garantia do 3PL e as preocupações ambientais percebidas foram altamente influentes na escolha de uma transportadora de pacotes 3PL. Verificou-se que as pessoas seriam incentivadas a usar os prestadores de serviços 3PL se demonstrassem disponibilidade e preocupação ambiental em atender às necessidades dos clientes. Posteriormente, os provedores de 3PL devem conferir segurança e conveniência antes, durante e depois da prestação do serviço para garantir o patrocínio contínuo dos consumidores. Este foi considerado o primeiro estudo que utilizou um aprendizado em conjunto ao aprendizado de máquina para medir a intenção comportamental para o setor de logística. A estrutura, as ferramentas de análise e os resultados deste estudo podem ser estendidos e aplicados entre outras intenções comportamentais em relação ao transporte em todo o mundo.

Um sistema preditivo foi o de Coussement *et al*. [11] no qual segundo ele, o aprendizado on-line foi adotado rapidamente por instituições e organizações educacionais. Apesar de suas muitas vantagens, incluindo acesso 24 horas por dia, 7 dias por semana, alta flexibilidade, conteúdo rico e baixo custo, o aprendizado on-line sofre com altas taxas de abandono que prejudicam os resultados pedagógicos e econômicos dos objetivos. Ferramentas aprimoradas baseadas em assinaturas de

previsão de abandono de alunos ajudariam os provedores a detectar proativamente os alunos em risco de abandono e identificar fatores de que eles podem abordar para ajudar os estudantes a continuar sua experiência de aprendizado. Portanto, este estudo buscou melhorar as previsões de evasão escolar, com três contribuições principais. Primeiro, compara-se um algoritmo de modelo *logit leaf* (LLM) recentemente proposto com outros oito algoritmos, usando um conjunto de dados da vida real de 10.554 alunos de um provedor global de aprendizado on-line baseado em assinatura. O LLM supera todos os outros métodos ao encontrar um equilíbrio entre desempenho preditivo e compreensibilidade. Em segundo lugar, uma nova visualização informativa multinível do LLM adiciona novos benefícios em relação a uma visualização LLM padrão. Em terceiro lugar, esta pesquisa especifica os impactos dos gráficos de demonstração dos alunos; características da sala de aula; e variáveis de engajamento acadêmico, cognitivo e comportamental na evasão escolar. Ao revisar os segmentos LLM, esses resultados mostram que diferentes percepções surgem para vários segmentos de alunos com diferentes padrões de aprendizagem. Esse resultado notável pode ser usado para personalizar campanhas de retenção de alunos.

Outro estudo foi o desenvolvido por Tarik, Aissa e Yousef [12], onde o projeto consistiu na concepção e implementação de um sistema de orientação dos alunos do núcleo comum para um dos ramos técnicos ou científicos do primeiro bacharelado. O objetivo deste sistema foi ter uma boa orientação que permitisse ao aluno obter uma boa nota de acordo com um modelo já existente, que contém o conjunto de alunos que já concluíram o bacharelado na região de Guelmim Ouad Noun. Para resolver o problema discutido antes, a solução foi implementar um sistema inteligente que atendesse às necessidades dos alunos. Isso exigiu um sistema que previsse a média de bacharelado do aluno por meio de sua ou suas notas principais usando *Machine Learning* e técnicas de mineração de dados.

A Figura 5 ilustra a abordagem para prever as médias dos alunos, onde os dados de entrada são lidos do MYSQL DBMS e depois transferidos para um *script python* que divide os dados em duas partes de treinamento e teste e executa os três algoritmos: floresta aleatória, árvore de decisão e regressão linear, e dá ao final a pontuação de cada algoritmo.

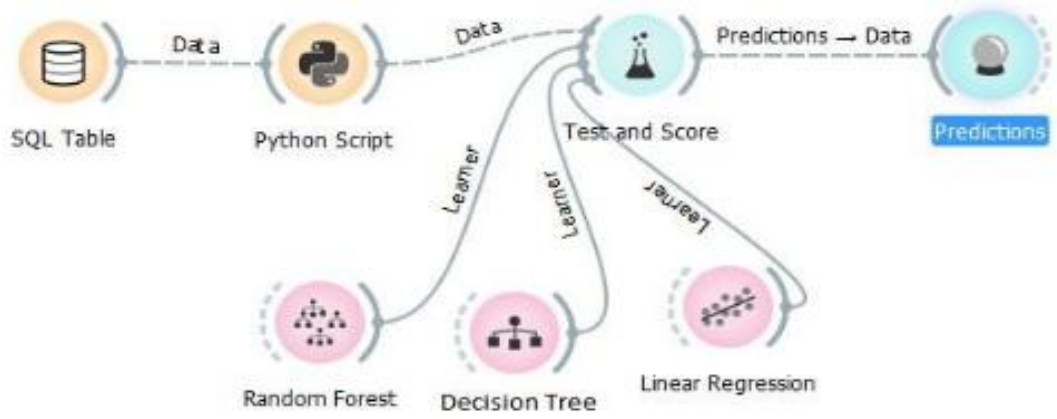


Figura 5. Abordagem do projeto [12].

A Tabela 2 apresenta um resumo das técnicas de *machine learning* utilizadas nos trabalhos relacionados para as variáveis usadas em sistemas preditivos, os quais inferem algum resultado:

Tabela 2. Trabalhos Relacionados e suas Técnicas.

Referência	**Algoritmo, Técnica**	**Resultado**
RAJENDRAN, S.; CHAMUNDESWARI, S.; SINHA, A.A.[7]	regressão logística multinomial, rede artificial neural artificial, floresta aleatória, aumento gradiente, métodos de empilhamento.	Prevê o desempenho acadêmico de alunos do ensino fundamental e médio usando MLAS com base em vários dados sóciodemográficos, variáveis relacionados a escola e relacionados ao aluno. Conclui-se que um estilo de vida consciente saudável se relaciona positivamente com o desempenho acadêmico, enquanto a existência de estresse tem um impacto negativo.
BECKHAM et al [8]	MLP, Floresta Aleatória.	Depois de descobrir qual fator afeta a nota do aluno, tenta-se prever a nota do mesmo usando modelos de MLAs para provar se esse fator realmente afeta a nota do aluno. O modelo MLP de 12 neurônios apresentou o melhor desempenho com valor RMSE de 4,32, seguido por Random Forest com valor RMSE de

		4,52 e, finalmente, Árvore de Decisão com valor RMSE de 5,69.
GERMAN J.D et al[9]	ANN, Floresta Aleatória.	Os resultados deste estudo revelaram que a atitude é o fator mais significativo que afeta a intenção comportamental dos Consumidores.
TARIK et al[12]	Floresta Aleatória, Árvore de Decisão, Regressão Linear.	A solução deve implementar um sistema inteligente que atenda às necessidades dos alunos. Isso exigirá um sistema que preveja a média de bacharelado do aluno por meio de sua ou suas notas principais usando *Machine Learning* e técnicas de mineração de dados.
Coussement et al. [11]	LLM	Ao revisar os segmentos LLM, os resultados mostram que diferentes percepções surgem para vários segmentos de alunos com diferentes padrões de aprendizagem. Esse resultado notável pode ser usado para personalizar campanhas de retenção de alunos

A escolha dos algoritmos Naive Bayes, KNN (K-Nearest Neighbors) e Árvore de Decisão está diretamente relacionada ao desenvolvimento de um sistema preditivo para calcular o risco de perder a residência estudantil devido às suas características e capacidades específicas em problemas de classificação.

- Naive Bayes: É um algoritmo baseado no teorema de Bayes, que assume independência entre os atributos do conjunto de dados. Ele é

particularmente útil quando se trabalha com dados categóricos e grandes volumes de informações, pois é simples, eficiente e pode lidar com incertezas. No contexto de predição de risco de perder a residência estudantil, ele pode ser usado para modelar a probabilidade de um aluno perder a residência com base em várias características, como desempenho acadêmico, comportamento e cumprimento de regras.

- K-Nearest Neighbors (KNN): O KNN é um algoritmo de classificação baseado na proximidade dos dados. Ele classifica novos exemplos com base nas categorias dos exemplos mais próximos no conjunto de treinamento. Isso faz do KNN uma escolha interessante para um sistema preditivo de risco, já que ele pode identificar padrões semelhantes entre estudantes que estão em situações de risco e aqueles que não estão. Ele é particularmente eficaz quando as relações entre as variáveis são complexas e não lineares.

- Árvore de Decisão: Este algoritmo cria um modelo em forma de árvore, onde as folhas representam decisões ou classificações baseadas nos valores dos atributos. As Árvores de Decisão são fáceis de interpretar, o que as torna úteis para explicar o raciocínio por trás da predição de risco de perder a residência. Elas podem ser usadas para identificar combinações de fatores, como notas baixas ou comportamento problemático, que aumentam a probabilidade de perda da residência, ajudando a visualizar claramente as condições de risco.

Esses três algoritmos têm diferentes abordagens para classificar e predizer, proporcionando uma boa base para desenvolver um sistema robusto que seja capaz de identificar padrões de risco de forma eficiente e com diferentes perspectivas de análise.

3. MATERIAIS E MÉTODOS

O capítulo de Materiais e Métodos, em um sistema preditivo para calcular o risco de perder a residência estudantil, tem como objetivo descrever de forma detalhada os recursos utilizados e os procedimentos adotados para o desenvolvimento do modelo. Nele, são apresentados os dados coletados (como ocorrências do aluno, comportamento apresentado e outros fatores relevantes), as fontes dessas informações, os métodos de pré-processamento dos dados e a escolha das técnicas de modelagem de dados ou de aprendizado de máquina. Também são explicados os critérios de avaliação do modelo, como métricas de acurácia e desempenho, e a justificativa para a escolha das ferramentas e algoritmos empregados. O propósito deste capítulo é garantir transparência, permitindo que o estudo seja replicável e validado por outros pesquisadores.

Esta pesquisa aborda uma arquitetura que utiliza dados quantitativos da planilha de Registro Disciplinar dos alunos. Posteriormente, objetiva a análise dessas informações. Pode-se classificar o trabalho descrito nesta proposta como uma pesquisa do tipo quantitativa. Já com relação a sua natureza, neste estudo a arquitetura proposta objetiva a resolução de um problema prático, sendo a previsão da possibilidade de perder a residência estudantil usando sistemas preditivos.

Quanto aos objetivos da presente pesquisa, pode-se afirmar que se encaixam entre o tipo de pesquisa descritiva - já que envolve a coleta de dados dos alunos - e o tipo de pesquisa explicativa, que visa a caracterização das causas dos fenômenos encontrados através da pesquisa descritiva. Com relação aos seus procedimentos, o presente trabalho pode ser classificado como experimental, pois passa por uma etapa de validação dos dados, cálculos em relação ao sistema preditivo. Finalmente, a classificação final da pesquisa pode ser nominada como quantitativa-aplicada-explicativa-experimental.

O SYSDAE é uma solução baseada em dados que visa identificar estudantes que estão em risco de não cumprir os requisitos para manter sua vaga na residência. O sistema coleta e analisa uma variedade de informações relevantes, como desempenho acadêmico (notas, frequência), comportamento disciplinar, e outros fatores específicos estabelecidos pela instituição. Utilizando algoritmos de aprendizado de máquina, como Naive Bayes, KNN e Árvores de Decisão, o sistema

treina modelos preditivos para avaliar a probabilidade de um aluno estar em risco. Com base nos dados analisados, o sistema fornece uma previsão quantitativa, permitindo à administração da residência identificar e oferecer apoio preventivo aos estudantes que estão mais propensos a perder sua vaga, promovendo intervenções direcionadas e eficazes.

No SYSDAE têm-se o método Risk_Calculation o qual é responsável por receber do Frontend o aluno selecionado, obter por parâmetro todos os dados do aluno selecionado na tabela Student, transferir esses dados relacionados e enviar pro algoritmo realizar a predição. Então quando o algoritmo faz a predição, o método Risk_Calculation considera a previsão do algoritmo e mostra os dados pro FrontEnd.

A metodologia utilizada na previsão do risco de perder a residência estudantil faz uso de técnicas de *machine learning*, utilizando modelos do KNN. O KNN (K-nearest neighbors, ou “K-vizinhos mais próximos”) costuma ser um dos primeiros algoritmos aprendidos por iniciantes no mundo do aprendizado de máquina. Em resumo, o KNN tenta classificar cada amostra de um conjunto de dados avaliando sua distância em relação aos vizinhos mais próximos. Se os vizinhos mais próximos forem majoritariamente de uma classe, a amostra em questão foi classificada nesta categoria.

O Algoritmo Naive Bayes funciona como classificador e se baseia na probabilidade de cada evento ocorrer, desconsiderando a correlação entre *features*. Por ter uma parte matemática relativamente simples, possui um bom desempenho e precisa de poucas observações para ter uma boa acurácia. Uma aplicação bastante comum é para identificar se um determinado e-mail é um spam ou não.

Uma árvore de decisão é um algoritmo de aprendizado de máquina supervisionado que é utilizado para classificação e para regressão. Isto é, pode ser usado para prever categorias discretas (sim ou não, por exemplo) e para prever valores numéricos (o valor do lucro em reais). A árvore de decisão estabelece nós (*decision nodes*) que se relacionam entre si por uma hierarquia. Existe o nó-raiz (*root node*), que é o mais importante, e os nós-folha (*leaf nodes*), que são os resultados finais. No contexto de *machine learning*, o nó raiz é um dos atributos da base de dados e o nó-folha é a classe ou o valor que foi gerado como resposta.

Testou-se os três algoritmos e no final, na ferramenta, optou-se pelo KNN devido á sua acurácia ser em torno de 95%.

A Figura 6 apresenta o fluxo metodológico no qual descreve os passos a serem seguidos.

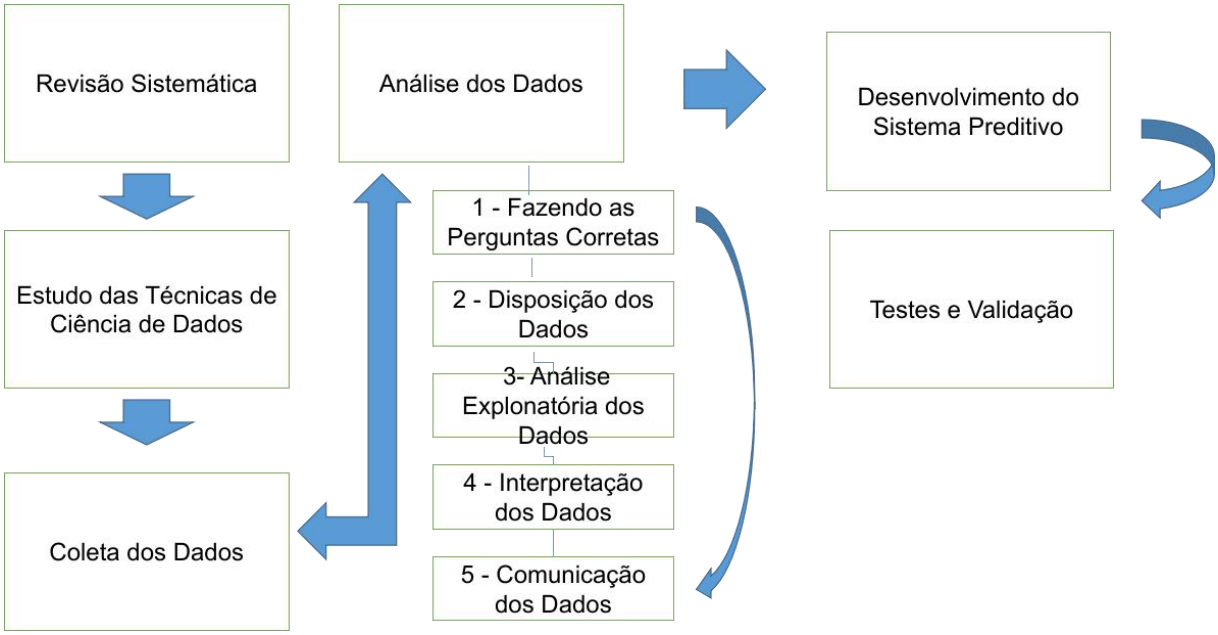


Figura 6. Fluxo Metodológico [Fonte Primária].

3.1 REVISÃO SISTEMÁTICA

Segundo Munhoz et al[17], Revisão Sistemática(RS) é um método organizado e sistematizado utilizado na avaliação de um conjunto de dados provenientes de diferentes estudos previamente publicados na literatura, visando reunir todas as evidências que correspondam aos critérios de eligibilidade previamente estabelecidos,a fim de nortear um tema específico de pesquisa.

A RS pode possuir grau variável de qualidade, dependendo do método utilizado e da experiência do pesquisador no uso dessas metodologias. Alguns aspectos principais são:desenvolvimento e publicação de um protocolo a priori; busca ampla de artigos, com uma estratégia de busca sensível(grande gama de termos) em diferentes bases de dados e com busca adicional por estudos incluídos (literatura cinza); avaliação da qualidade dos estudos incluídos; busca, seleção e extração de dados por dois pesquisadores independentes; uso adequado de técnicas meta-analíticas para análise dos resultados, quando for possível.

3.2 IDENTIFICAÇÃO DOS ESTUDOS

As bases de dados escolhidas para esta revisão são: ACM (Association for Computing Machinery), ScienceDirect, Scopus e Springer.

Para o levantamento de trabalhos, definiu-se a seguinte expressão de busca e suas respectivas palavras-chave: (machine learning OR machine learning algorithms OR student residence OR predictive system OR recomendation system OR "machine learning" OR "student residence" OR “machine learning algorithms” OR “predictive system” OR “recomendation system”) AND ("recomendation system" OR "machine learning algorithms" OR "student residence" OR "predictive system”)

Não houve alteração na expressão de busca nas bases de dados. O intervalo de busca foi de 4 anos, ou seja, de 2020 a 2024. Os estudos analisados e considerados nessa revisão devem estar escritos na língua inglesa em artigos, desconsiderando capítulos de livros, anais e resumos de eventos e seminários.

Foram obtidos 195 artigos a partir da busca na base de dados utilizando, sempre que disponível, filtrando por metadados, pois é um recurso que permite filtrar e localizar dinamicamente estes conteúdos. Na triagem foram selecionados 54 artigos para a leitura completa e verificação dos resultados. Dentre estes 54, 25 artigos foram selecionados para o escopo desta revisão.

Os 25 estudos analisados por completo, envolvendo diversas áreas da ciência de dados. Estas técnicas mostram várias possibilidades, bem como o uso na predição de variáveis e construção de melhores modelos de dados.

3.3 CARACTERÍSTICAS DE UMA REVISÃO SISTEMÁTICA

Segundo Munhoz et al[17], as características de uma RS devem ser definidas previamente ao seu registro e execução, e consistem no detalhamento da estratégia PICO ou de qualquer outra coisa utilizada no momento da escolha da pergunta-chave, assim como no refinamento do tema, dentre outros aspectos.

Na sequência enumera-se as características que precisam ser definidas em uma RS.

3.3.1 Critérios de Inclusão e Exclusão

Os critérios de inclusão e exclusão de uma RS são definidos a partir do detalhamento da Estratégia PICO(T/S) (ou de outra estratégia), escolhida com base na pergunta-chave.

Os detalhes da população-alvo (por exemplo: idoso, mulher, atleta, idade, acometido por determinado processo patológico etc.) a ser estudada devem ser pormenorizados, assim como os dos respectivos grupos controle e/ou outros com finalidade de comparação, se houver. Detalhes sobre a intervenção à qual o grupo de estudo foi submetido, o desfecho considerado, o tipo de publicação científica que será considerada na RS (por exemplo: restrição de língua utilizada na publicação, artigos originais ou relatos de casos, tipos de pesquisa, etc.) e o período de tempo no qual os artigos serão selecionados, desde que devidamente justificados(por exemplo: até 2021, os últimos 10 anos somente, sem restrição temporal etc.) são também parte da inclusão e exclusão.

Essa estratégia deve ser aplicada para todos os ítens da RS e orienta a seleção dos estudos que farão parte da revisão, servindo como um “filtro” para estes.

Observa-se dois exemplos aleatórios de critérios de inclusão/exclusão de artigos na Tabela 3.

Tabela 3 : Exemplos de critérios de inclusão e exclusão.

P População a ser estudada	I Intervenção	C Controle Comparações	O Outcome, Desfecho	T Tipo de Estudo, tipo de publicação
Mulheres (saudáveis ou com determinada doença, com determinada faixa etária)	Uso de uma medicação específica	Comparadas a homens ou que receberam placebo	Que desenvolveram determinada doença	Estudos randomizados/não randomizados

Na pós-menopausa(ou outra faixa etária); ou etnia específica etc.	Submetidas a um exame de imagem específico	Em idade fértil	Tiveram reações alérgicas a determinado medicamento	Estudos de pesquisa ou relatos de caso

Fonte: Munhoz et al[17]

3.3.2 - Objetivos Principal e Secundário

O objetivo principal é o objetivo primário fundamental de uma revisão. Não necessariamente é um objetivo único, mas obrigatoriamente é a base da pergunta principal feita à literatura. O objetivo secundário deriva do principal, ou seja, é complementar ao objetivo principal e não existiria sem ele.

3.4 - ESTRATÉGIA DE BUSCA BIBLIOGRÁFICA PARA UMA REVISÃO SISTEMÁTICA

Descrever a estratégia de busca bibliográfica consiste em definir os passos da busca, desde a inserção das palavras-chave até a seleção dos artigos.

Esse processo se inicia com a definição de quais bases de dados bibliográficos pretende-se consultar.

Embora alguns exemplos das principais bases de dados voltadas para a áreas de machine learning estejam citados neste trabalho, na fase de base de dados bibliográficos, é primordial que o revisor procure na literatura da sua área de estudo as bases de dados mais relevantes, analisando a literatura existente e as revisões previamente realizadas e consultando os experts do assunto.

Depois da estruturação do protocolo, descrevem-se como as palavras-chave serão inseridas em cada base de dados; quais são as palavras-chave, suas combinações e operadores booleanos, e como será realizada a seleção dos artigos resultantes da pesquisa como mostra a Tabela 4.

Tabela 4 : Passos para realizar a análise dos resultados das estratégias de busca

“Os resultados das buscas nas bases de dados bibliográficos serão analisados da seguinte forma:”
a) Primeiramente serão lidos e verificados os títulos nos artigos. Serão excluídos aqueles artigos que se encontram em algum critério de exclusão, cujo assunto ou tipo de assunto, esteja evidente no título;
b) Em um segundo momento, os abstracts serão lidos e, da mesma forma, excluídos aqueles artigos nos quais o assunto ou tipo de estudo apresenta evidências de enquadramento em algum critério de exclusão;
c) Os resultados dos artigos não excluídos serão transformados em um arquivo passível de upload em software de gerenciamento de referências para a RS (Rayyan QRI). Os membros da equipe analisarão os artigos, que deverão ou não ser incluídos na RS após a verificação dos seus respectivos textos na íntegra. O modo “cego” será ativados no software Rayyan, para enviar viés de seleção”

Fonte: Munhoz et al[17]

3.5 EXTRAÇÃO DOS DADOS

Para cada um dos 25 estudos selecionados para essa revisão sistemática, serão destacadas as seguintes informações: autores, objetivos gerais, descrição do método utilizado para análise dos dados de machine learning, MLAs, predictive systems, descrição dos métodos utilizados para a análise dos dados na predição e os principais resultados obtidos.

3.6 COLETA DE DADOS

A coleta de dados foi conduzida pela equipe do Departamento de Assistência Estudantil (DAE) do Instituto Federal do Rio Grande do Sul (IFRS) – Câmpus Sertão, seguindo um processo estruturado para garantir a precisão e integridade dos dados. As etapas da coleta de dados incluíram:

3.6.1 Fontes dos dados

De acordo com Bilal et Al[19], as principais fontes de dados são:

I. Registros Acadêmicos: Dados sobre o desempenho acadêmico dos alunos, incluindo notas, frequência e histórico escolar.
II. Registros Disciplinares: Dados sobre ocorrências disciplinares, incluindo tipos de infrações, medidas disciplinares aplicadas e número de dias suspenso.
III. Dados Demográficos: Informações sobre o perfil dos alunos, como idade, sexo, curso, série, setor e matrícula.

3.6.2 Método de coleta dos dados

I. Sistema de Gerenciamento Escolar: Extração de dados do sistema de gerenciamento escolar utilizado pelo IFRS, garantindo a integridade e a precisão dos dados coletados.
II. Entrevistas: Utilização de entrevistas para coletar informações adicionais diretamente dos alunos e da equipe do DAE.

3.6.3 Justificativa para a coleta dos dados

A escolha dessas fontes de dados baseia-se em estudos anteriores que demonstraram a importância de fatores acadêmicos e comportamentais na previsão de risco de evasão escolar. Além disso, a combinação de dados acadêmicos, disciplinares e demográficos permite uma análise mais completa e precisa dos padrões de comportamento dos alunos.

3.7 TRATAMENTO DOS DADOS

Os dados coletados foram submetidos a um rigoroso processo de tratamento para garantir sua qualidade e consistência. As etapas de tratamento dos dados incluíram as seguintes:

3.7.1 Limpeza dos dados

I. Remoção de *Outliers*: Identificação e remoção de valores anômalos que poderiam distorcer a análise.
II. Tratamento de Valores Ausentes: Aplicação de técnicas de imputação para substituir valores ausentes, utilizando a média ou mediana dos dados, conforme apropriado.

3.7.2 Normalização dos dados

I. Escalonamento dos Valores: Normalização dos dados para que todas as variáveis tenham a mesma escala, facilitando a comparação e a análise. Utilizou-se a técnica de Min-Max Scaling para transformar os dados para um intervalo de 0 a 1.

3.7.3 Anonimização dos dados

I. Proteção da Privacidade: Anonimização das informações sensíveis dos alunos, substituindo identificadores pessoais por códigos anônimos.

3.7.4 Divisão dos dados

I. Conjunto de Treinamento e Teste: Divisão dos dados em conjuntos de treinamento (70%) e teste (30%) para a validação dos modelos de *Machine Learning*, garantindo que a avaliação dos modelos seja realizada em dados não vistos anteriormente.

3.8 ANÁLISE DOS DADOS

Esta seção detalha as principais medidas utilizadas na análise do dataset, fornecendo uma explicação aprofundada de cada métrica e sua relevância no

contexto do estudo. Além disso, apresentamos visualizações em forma de histogramas para uma compreensão mais intuitiva da distribuição dos dados.

3.8.1 - Métricas Analisadas

1. Precisão (Precision)

A precisão é uma métrica crucial em modelos de classificação, representando a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas. No contexto deste estudo, a precisão indica a capacidade do modelo de identificar corretamente os casos de perda de residência.

Um valor alto de precisão significa que quando o modelo prevê que um estudante perdeu a residência, essa previsão é geralmente correta. Isso é particularmente importante em situações onde falsos positivos podem ter consequências significativas, como alertas desnecessários ou alocação inadequada de recursos.

2. Revocação (Recall)

A revocação, também conhecida como sensibilidade, mede a proporção de casos positivos reais que foram corretamente identificados pelo modelo. Neste estudo, a revocação indica a capacidade do modelo de identificar corretamente todos os casos de perda de residência.

Um alto valor de revocação é crucial quando é importante não deixar passar nenhum caso positivo, mesmo que isso signifique alguns falsos positivos. No contexto de perda de residência estudantil, uma alta revocação assegura que a maioria dos estudantes em risco seja identificada, permitindo intervenções precoces.

3. Pontuação F1 (F1-Score)

A pontuação F1 é a média harmônica entre precisão e revocação, fornecendo um único valor que equilibra ambas as métricas. Esta medida é particularmente útil quando se busca um equilíbrio entre precisão e revocação, especialmente em datasets desbalanceados.

No contexto de previsão de perda de residência, um alto F1-Score indica que o modelo é capaz de identificar casos de risco com alta precisão, sem deixar de capturar a maioria dos casos reais. Isso é ideal para sistemas de alerta precoce, onde tanto falsos positivos quanto falsos negativos podem ter implicações significativas.

3.8.2 - Visualizações

Para cada uma dessas medidas, utilizamos histogramas para visualizar sua distribuição. Os histogramas oferecem uma representação gráfica da frequência de diferentes valores em um conjunto de dados, permitindo uma rápida compreensão da distribuição e identificação de padrões ou anomalias.

3.8.2.1 - Interpretação dos Histogramas

- Histograma de Precisão: Mostra a distribuição dos valores de precisão. Um histograma concentrado em valores mais altos indica um modelo com boa capacidade de previsão positiva.

- Histograma de Revocação: Ilustra a distribuição dos valores de revocação. Uma concentração em valores altos sugere que o modelo é eficaz em identificar a maioria dos casos positivos reais.

- Histograma de Pontuação F1: Representa o equilíbrio entre precisão e revocação. Um histograma com pico em valores altos indica um bom desempenho geral do modelo.

Estas visualizações, combinadas com as explicações detalhadas de cada métrica, fornecem uma compreensão abrangente do desempenho do modelo e da distribuição das medidas-chave no dataset analisado.

3.8.2.2 - Explicação dos Gráficos Utilizados na Análise

- **Métricas de Desempenho**

Tipo de Gráfico: Gráfico de Barras

O que mostra: Este gráfico apresenta as três principais métricas de desempenho do modelo: Precisão, Revocação e Pontuação F1.

Como interpretar: Cada barra representa uma métrica diferente. Quanto mais alta a barra, melhor o desempenho naquela métrica específica.

O gráfico de Métricas de Desempenho é mostrado na Figura 7.

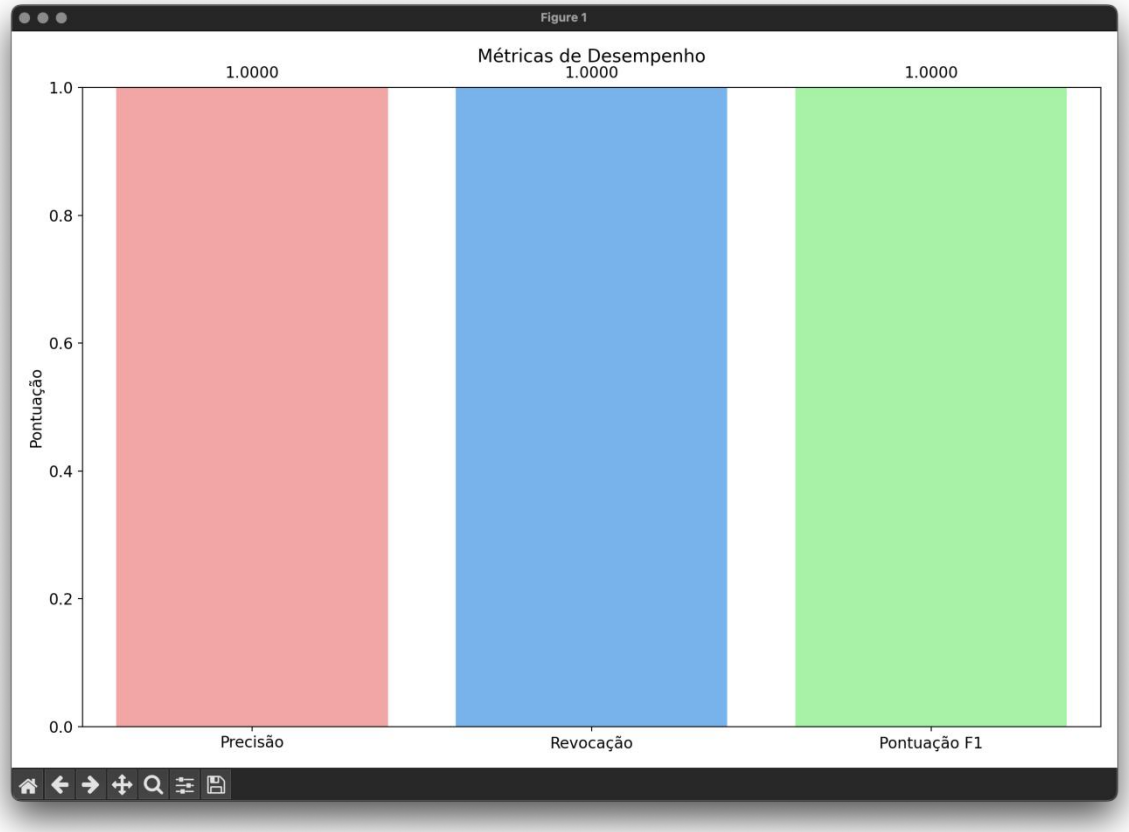


Figura 7 - Métricas de Desempenho[Fonte Primária].

- **Distribuição de Horas Orientadas**

Tipo de Gráfico: Histograma

O que mostra: A distribuição do número de horas orientadas atribuídas aos estudantes.

Como interpretar: O eixo X mostra o número de horas, e o eixo Y mostra quantos estudantes receberam esse número de horas. Picos no gráfico indicam valores mais comuns de horas orientadas.

O Gráfico de Distribuição de Horas Orientadas é mostrado na Figura 8

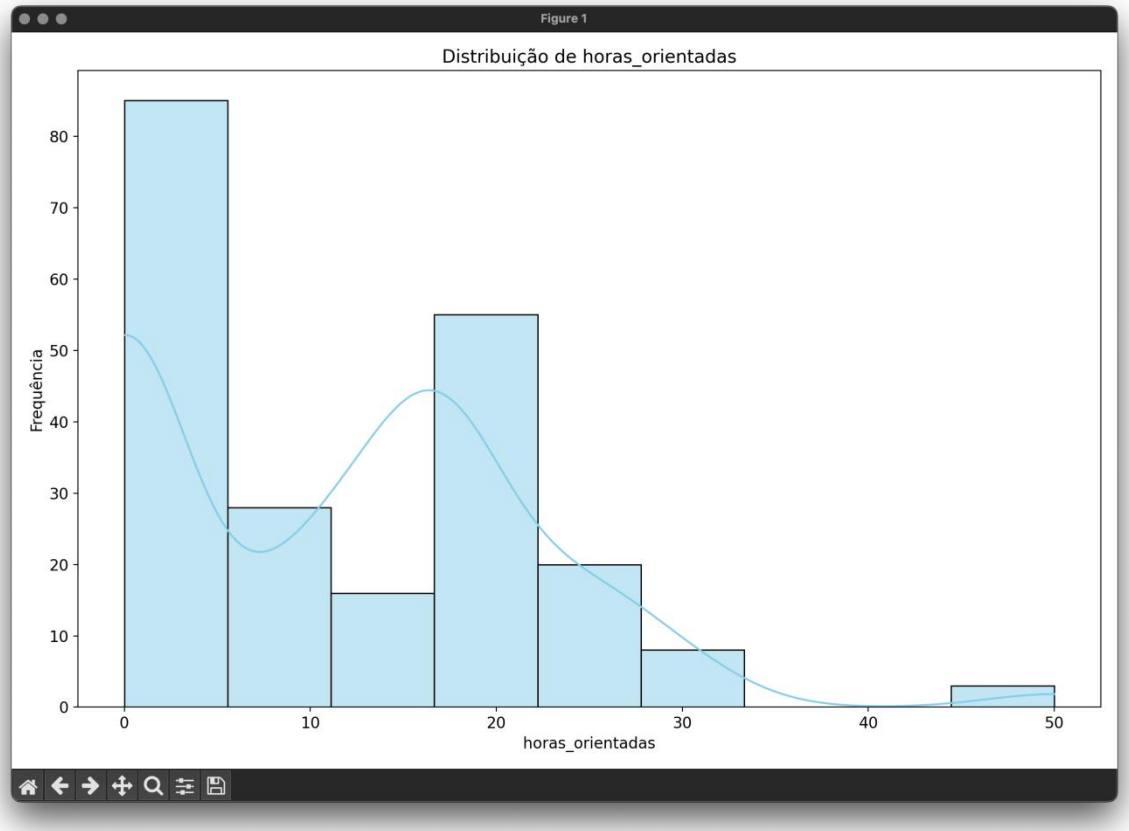


Figura 8 - Gráfico de Distribuição de Horas Orientadas[Fonte Primária].

- **Distribuição de Dias de Suspensão**

Tipo de Gráfico: Histograma

O que mostra: A distribuição do número de dias de suspensão entre os estudantes.

Como interpretar: Similar ao gráfico anterior, mas para dias de suspensão. Picos mostram durações de suspensão mais frequentes.

O Gráfico de Distribuição de Dias de Suspensão é mostrado na Figura 9

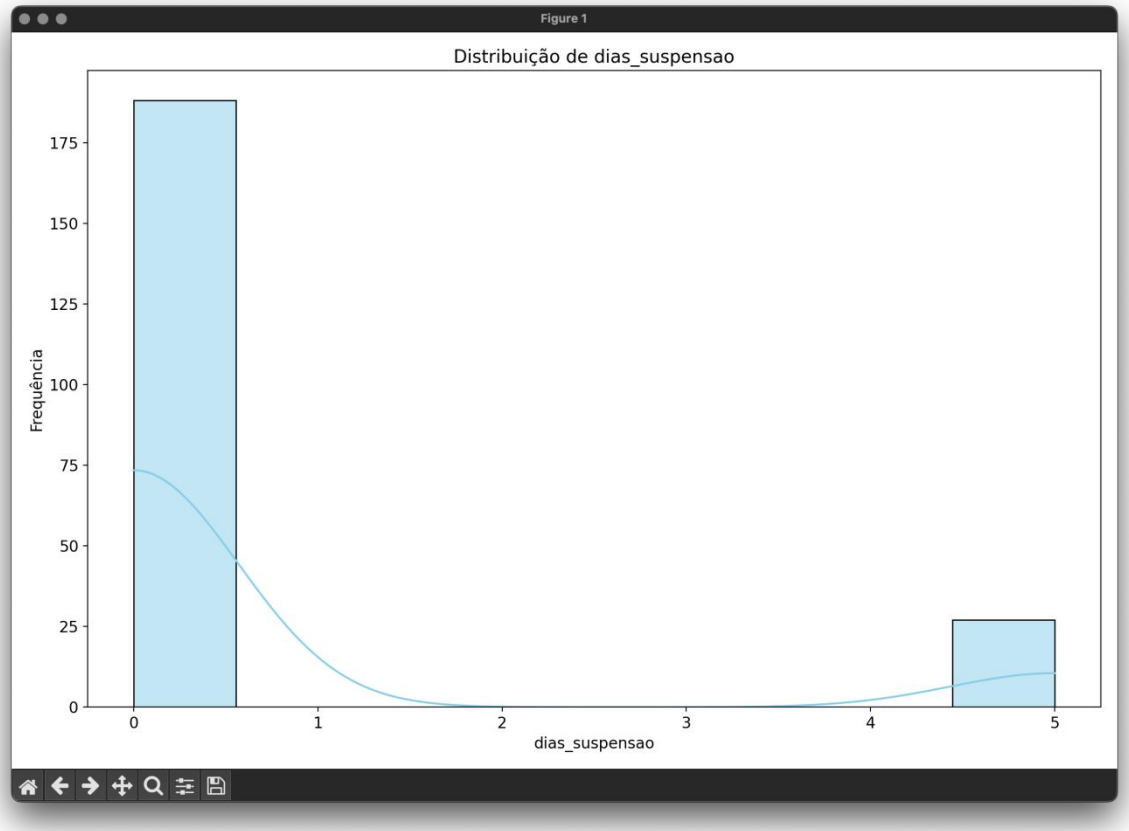


Figura 9 - Gráfico de Distribuição de Dias de Suspensão[Fonte Primária].

- **Gráfico de Distribuição de Série**

Tipo de Gráfico: Histograma

O que mostra: A distribuição dos estudantes por série escolar.

Como interpretar: Cada barra representa uma série, e a altura da barra indica quantos estudantes estão naquela série.

O Gráfico de Distribuição de Série é mostrado na Figura 10

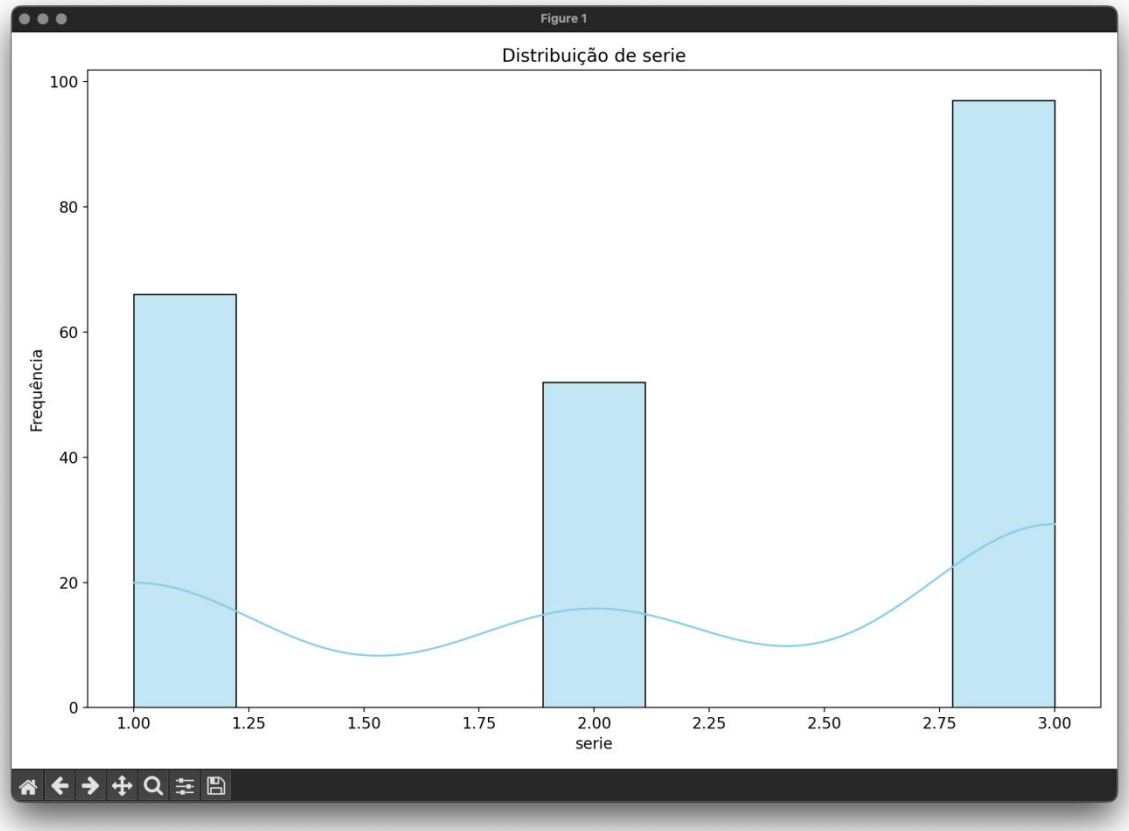


Gráfico 10 - Distribuição de Série[Fonte Primária]

- **Gráfico do Mapa de Calor de Correlação**

Tipo de Gráfico: Mapa de Calor (Heatmap)

O que mostra: As correlações entre diferentes variáveis no dataset.

Como interpretar: Cores mais intensas (vermelho ou azul escuro) indicam correlações mais fortes. Vermelho para correlações positivas, azul para negativas.

O Gráfico do Mapa de Calor de Correlação é mostrado na Figura 11

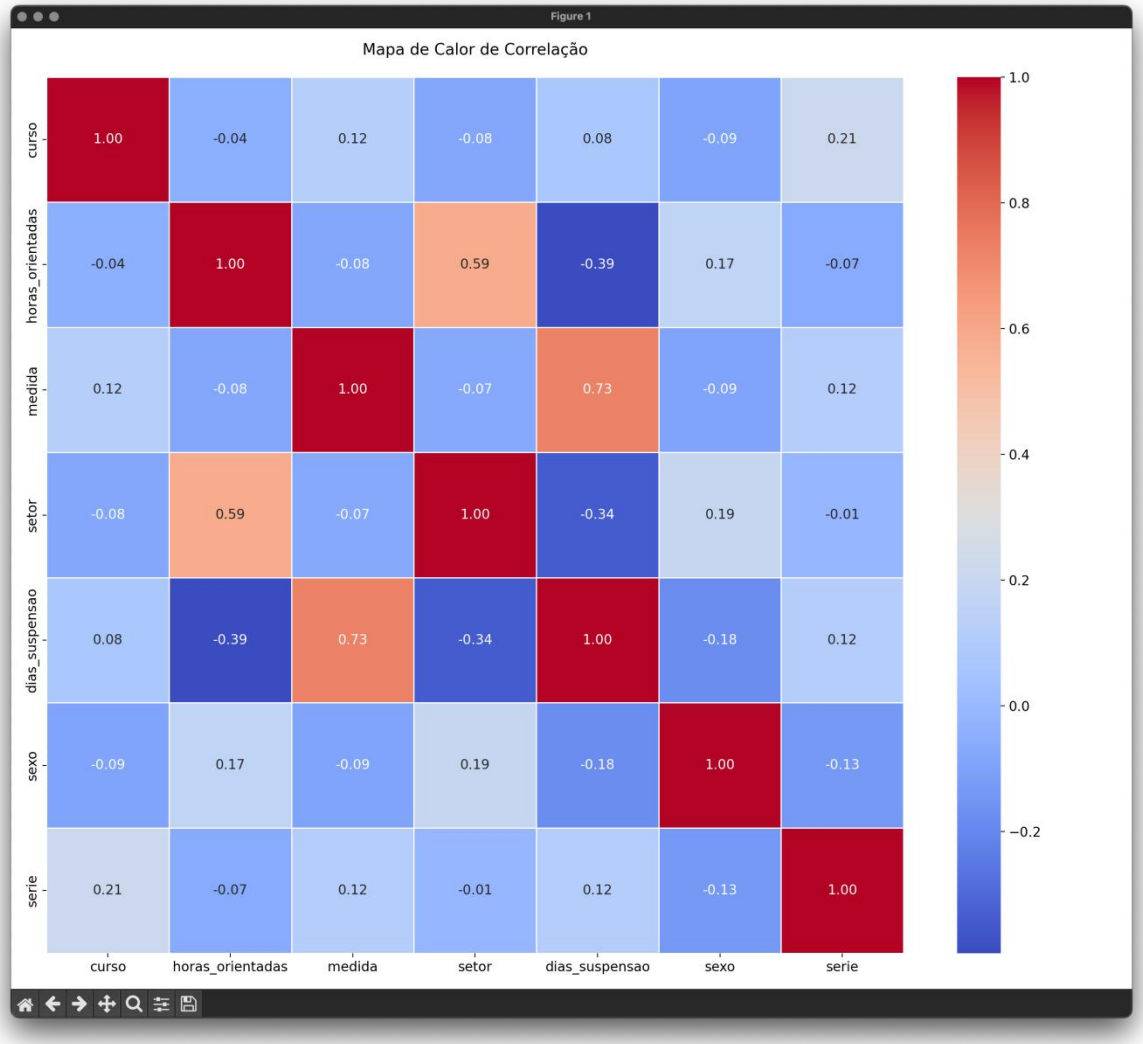


Figura 11 - Gráfico de Mapa de Calor de Correlação[Fonte Primária]

- **Gráfico de Matriz de Confusão**

Tipo de Gráfico: Matriz de Confusão

O que mostra: O desempenho do modelo em termos de previsões corretas e incorretas.

Como interpretar: Os números nas células mostram quantas previsões caíram em cada categoria (verdadeiros positivos, falsos positivos, etc.).

O Gráfico de Matriz de Confusão é mostrado na Figura 12.

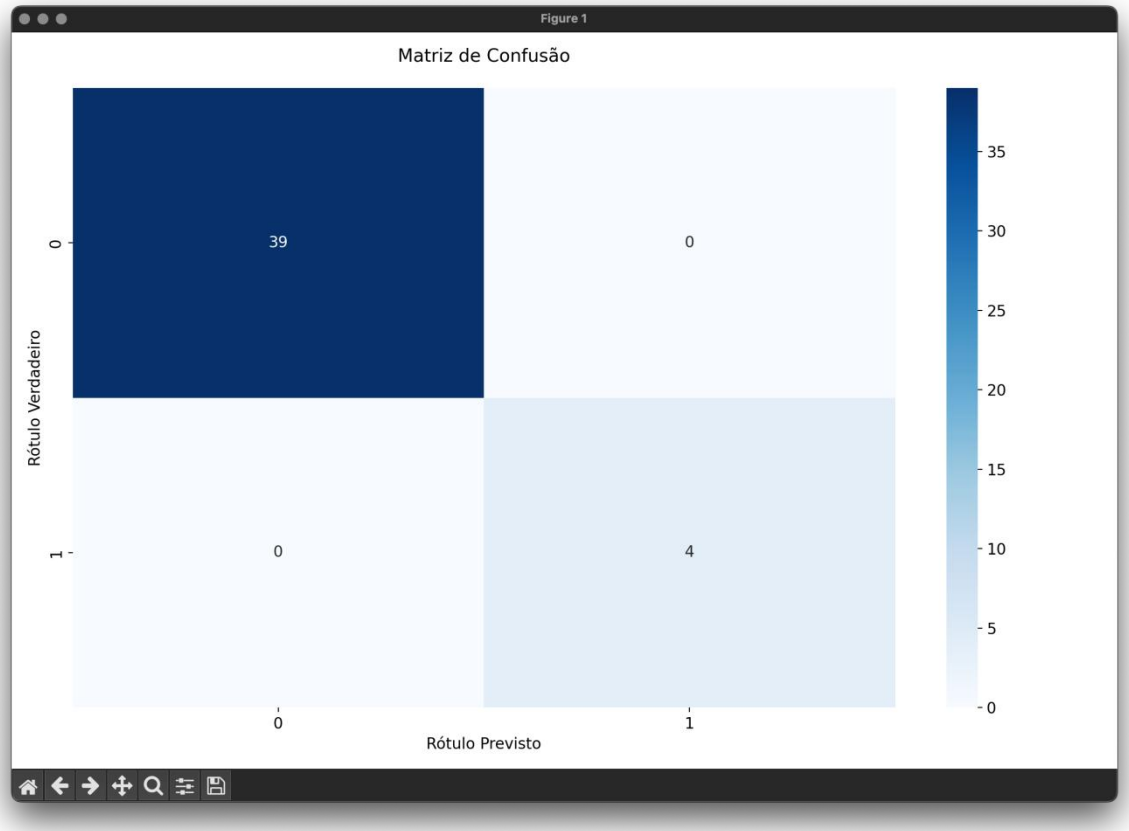


Figura 12 - O Gráfico de Matriz de Confusão[Fonte Primária]

- **Importância das Features**

Tipo de Gráfico: Gráfico de Barras Horizontais

O que mostra: A importância relativa de cada feature (variável) no modelo de previsão.

Como interpretar: Barras mais longas indicam features mais importantes para o modelo fazer suas previsões.

A Importância das Features é mostrada na Figura 13.

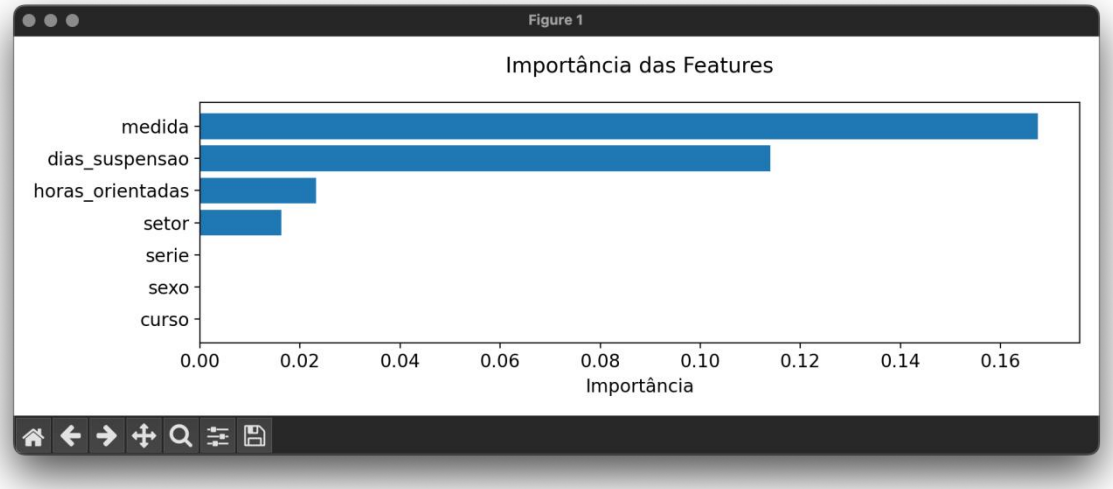


Figura 13 - Importância das Features[Fonte Primária]

Cada um desses gráficos fornece uma perspectiva diferente sobre os dados e o desempenho do modelo. Juntos, eles oferecem uma visão abrangente da análise realizada, permitindo entender tanto as características do dataset quanto a eficácia do modelo de previsão.

3.8.3 Ferramentas Utilizadas:

a. Linguagem de Programação: Python, devido à sua vasta biblioteca de ferramentas para análise de dados e *Machine Learning*, incluindo Scikit-learn e Pandas.
b. Ambiente de Desenvolvimento: Visual Studio Code, por sua capacidade de integrar código, executar diversos tipos de código, bastando ter a extensão adequada.
c. Framework: Django, um framework web Python de alto nível que permite o rápido desenvolvimento de sites seguros e de fácil manutenção. É gratuito e de código aberto, tem uma comunidade próspera e ativa, ótima documentação e muitas opções de suporte gratuito e pago [13].
d. Banco de Dados: SQLite, por ser uma biblioteca em processo que implementa um mecanismo de banco de

dados SQL transacional independente, sem servidor e com configuração zero [16].

3.9 JUSTIFICATIVA PARA ESCOLHA DOS ALGORITMOS

A escolha dos algoritmos Naive-Bayes, KNN e Árvore de Decisão baseou-se em suas características distintas e comprovada eficácia em tarefas de classificação similares. Segundo Bonnin [1], o Naive-Bayes é particularmente eficiente em situações com alta dimensionalidade de dados, enquanto o KNN é reconhecido por sua simplicidade e eficácia em problemas de classificação com dados bem definidos. A Árvore de Decisão, por sua vez, é amplamente utilizada devido à sua interpretabilidade e capacidade de capturar relações não lineares entre as variáveis.

Além disso, a aplicação da validação cruzada k-fold (k=10) foi escolhida para garantir a robustez dos modelos e minimizar o risco de overfitting, conforme recomendado por NG [15]. A combinação dessas técnicas permite uma análise abrangente e precisa dos dados, proporcionando *insights* valiosos sobre os padrões de comportamento dos alunos e o risco de perda da residência estudantil.

3.9.1 A ACURÁCIA

Foram separados em 4 parâmetros variáveis de entrada, Treino x Alvo, Treino / Features Teste (Validação) x Variáveis de Saída Teste (Validação), tendo 30% do total dos dados para Validação.

Na Tabela 5 foi mostrado os totais do conjunto de treinamento/validação.

Tabela 5. Tamanho do Conjunto de Treinamento/Validação [Fonte Primária].

Tamanho do conjunto de Treinamento	**Tamanho do conjunto de Validação**
(172)	(43)

Lembrando-se que obviamente, o algoritmo vê apenas os dados do Treinamento, os dados de Teste (Validação) são ocultados.

Foram feitos testes com três algoritmos para verificar a acurácia e definir qual destes seria utilizado. Os resultados obtidos foram os seguintes, como se pode observar na Tabela 6.

Tabela 6. Acurácia de cada algoritmo [Fonte Primária].

Algoritmo	**Acurácia**
Naive Bayes	0.88
KNN	0.93
Decision Tree	0.86

Nota-se, portanto, que o algoritmo que obteve a maior acurácia foi o KNN, com 93% de chances de predizer corretamente qual o percentual de risco de o aluno perder a residência estudantil.

4. O SYSDAE

O estudo deu início com uma revisão sistemática, onde, cumprido todas as etapas, foram escolhidos 25 trabalhos com critérios pré-definidos na revisão. Também teve como base a busca por pesquisas que fizessem uso de técnicas para ciência de dados como base para esse projeto. As técnicas e ou modelos mais utilizados nestes estudos foram divididas em 3 técnicas de *machine learning*, a escolha do algoritmo a ser implementado foi o KNN.

O termo “K”, em KNN, se refere ao número de vizinhos mais próximos considerados para tomar uma decisão. Por exemplo, se K é igual a 3, o algoritmo verificará os três vizinhos mais próximos e atribuirá ao novo exemplo a classe mais frequente entre esses vizinhos [16]. Isso significa que a escolha do valor de K é um fator crucial no desempenho do algoritmo.

Uma das principais vantagens do KNN é que ele não requer uma fase de treinamento complexa, pois utiliza diretamente os exemplos de treinamento para tomar decisões. Além disso, é um algoritmo relativamente simples de entender e implementar. No entanto, ele pode ser computacionalmente caro quando o conjunto de dados é grande, pois ele precisa calcular a distância entre cada exemplo e todos os outros exemplos do conjunto de treinamento.

Aplicações práticas do KNN são diversas. Ele pode ser utilizado em problemas de classificação, como diagnóstico médico, reconhecimento de padrões e detecção de fraudes. Por exemplo, em um problema de diagnóstico médico, o KNN pode ser utilizado para determinar se um paciente tem uma determinada doença com base na similaridade com outros pacientes do conjunto de treinamento. Da mesma forma, o KNN pode ser utilizado em problemas de regressão, como previsão de preços imobiliários ou estimativa de demanda futura.

Para implementar o algoritmo KNN em um projeto de *Machine Learning*, é necessário definir a métrica de distância utilizada para calcular a proximidade entre os exemplos. A métrica mais comumente utilizada é a distância euclidiana, mas outras métricas também podem ser utilizadas dependendo do domínio do problema. Além disso, é importante normalizar os dados antes de aplicar o KNN, para evitar que *features* com escalas diferentes dominem o cálculo da distância.

4.1. VANTAGENS DO KNN

Segundo Awari[20], tem-se as seguintes vantagens:

Simplicidade: O KNN é um algoritmo relativamente simples e fácil de entender. Não requer uma fase de treinamento complexa, pois utiliza diretamente os exemplos de treinamento para tomar decisões.

Adaptabilidade: O KNN é um algoritmo não paramétrico, o que significa que ele pode se adaptar a diferentes tipos de dados e problemas. Ele não faz suposições sobre a distribuição dos dados subjacentes e pode funcionar bem em problemas complexos com fronteiras de decisão não lineares.

Interpretabilidade: As decisões tomadas pelo KNN são baseadas na proximidade dos vizinhos mais próximos, o que pode ser facilmente compreendido e interpretado. Isso o torna uma escolha preferencial em problemas que exigem explicabilidade.

4.2. DESVANTAGENS DO KNN

Ainda segundo Awari[20], as desvantagens seriam:

Sensibilidade a *Outliers*: O KNN pode ser sensível a valores atípicos (*outliers*) em seu conjunto de dados. Como ele se baseia na proximidade dos exemplos, um único exemplo ruidoso pode distorcer as decisões tomadas pelo algoritmo.

Custo Computacional: O KNN precisa calcular a distância entre cada exemplo e todos os outros exemplos do conjunto de treinamento. Isso pode ser computacionalmente caro quando o conjunto de dados é grande, especialmente se a distância for complexa de calcular.

Escolha do Parâmetro K: A escolha do valor para o parâmetro K é um fator crucial no desempenho do KNN. Um valor muito baixo pode levar a decisões instáveis, enquanto um valor muito alto pode levar a decisões enviesadas. É necessário realizar uma validação cruzada ou buscar uma abordagem automatizada para escolher o melhor valor de K.

4.3. APLICAÇÃO DESENVOLVIDA

A aplicação foi desenvolvida em python com uso das bibliotecas numpy, pandas, scikit-learn e django. Foram testados três algoritmos: KNN, Naive-Bayes e de árvore de decisão. No final, avaliamos qual algoritmo foi mais eficiente.

4.3.1. Manutenção dos estudantes

Para a manutenção de algum aluno deve-se clicar no botão Adicionar Estudante, localizado no menu lateral esquerdo, com os seguintes dados: Matrícula; Nome do Pai; Telefone do Pai; Nome da Mãe, Telefone da Mãe, E-mail do Responsável, Gênero; Data de Nascimento e Curso. Os valores de KNN devem permanecer 0 para poder ocorrer o cálculo correto dos valores.

A Figura 14 mostra os campos da Tela de Cadastro.

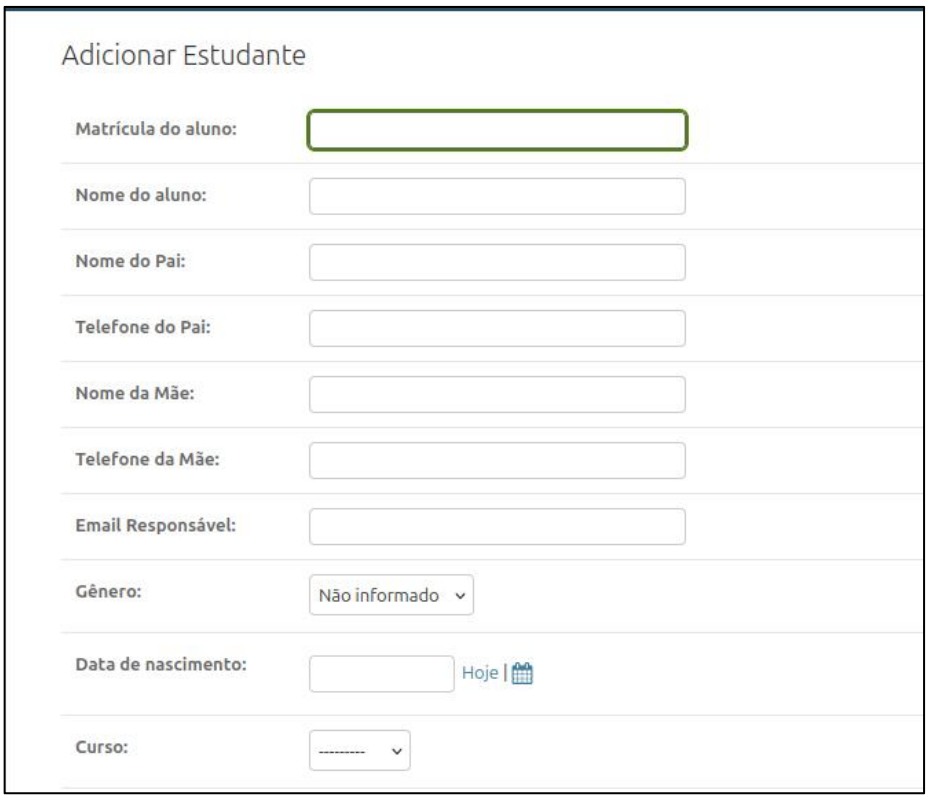


Figura 14. Tela de Cadastro de Estudantes [Fonte Primária].

4.3.2. Manutenção dos itens do regulamento do sistema

Para a criação de algum Item do Regulamento deve-se clicar no botão de Inclusão de Regulamento, localizado no menu lateral esquerdo, com os seguintes dados: Regulamento, Descrição, Psicóloga, Pedagogo, Coordenador do DAE e Diretor. Sendo o campo “Regulamento” tratado como um título da Norma, a qual irá aparecer quando for citar ela nas Violações. No campo Psicóloga, deve-se informar o e-mail da Psicóloga; no campo Pedagogo deve-se informar o e-mail do pedagogo; no campo Coordenador, deve-se informar o e-mail do Coordenador; no campo Diretor, deve-se informar o e-mail do diretor.

A tela de cadastro dos itens de Regulamento é mostrada a seguir na Fig. 15

Figura 15. Os principais Itens de Regulamento estão pré-cadastrados para uso posterior [Fonte Primária].

4.3.3. Manutenção das medidas disciplinares no sistema

Para a criação de alguma Medida Disciplinar deve-se clicar no botão de inclusão, localizado no menu lateral esquerdo, com os seguintes dados: Descrição, Classificação, Regulamento e Registro de Ata. No campo Descrição tem-se os valores Advertência, Advertência Escrita, Horas Orientadas, Suspensão, Perda de Residência e Expulsão; no campo Classificação pode-se escolher entre as três opções: leve, média e grave. Por conta do Modelo padrão, foram atribuídos os valores 1, 2 e 3 respectivamente. No campo Registro de Ata tem-se o campo responsável pela gravação do número da ata correspondente.

A Figura 16 mostra a tela de cadastro das medidas disciplinares.

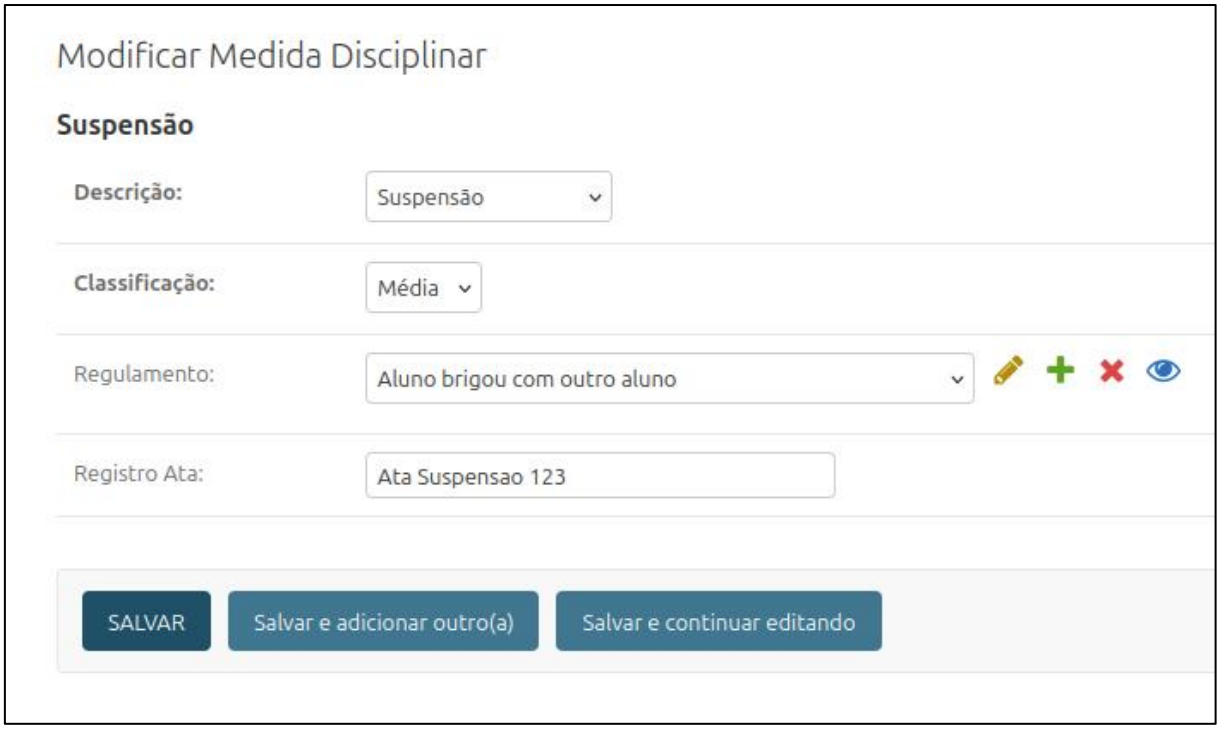


Figura 16. As principais medidas disciplinares também já estão cadastradas [Fonte Primária].

4.3.4. Manutenção de Violações no sistema

Para a criação de alguma Violação deve-se clicar no botão de inclusão, localizado no menu lateral esquerdo, com os seguintes dados: Estudante, Infração,

Descrição da Ocorrência, Medida Disciplinar, Complemento da Medida e Data do Evento.

No campo de “Infração”, aparecerá as opções criadas na “Normas”.

No campo “Complemento da Medida” pode ser utilizado para justificar melhor a escolha da Medida Disciplinar do aluno.

A figura 17 mostra um exemplo do cadastro de violação do sistema.

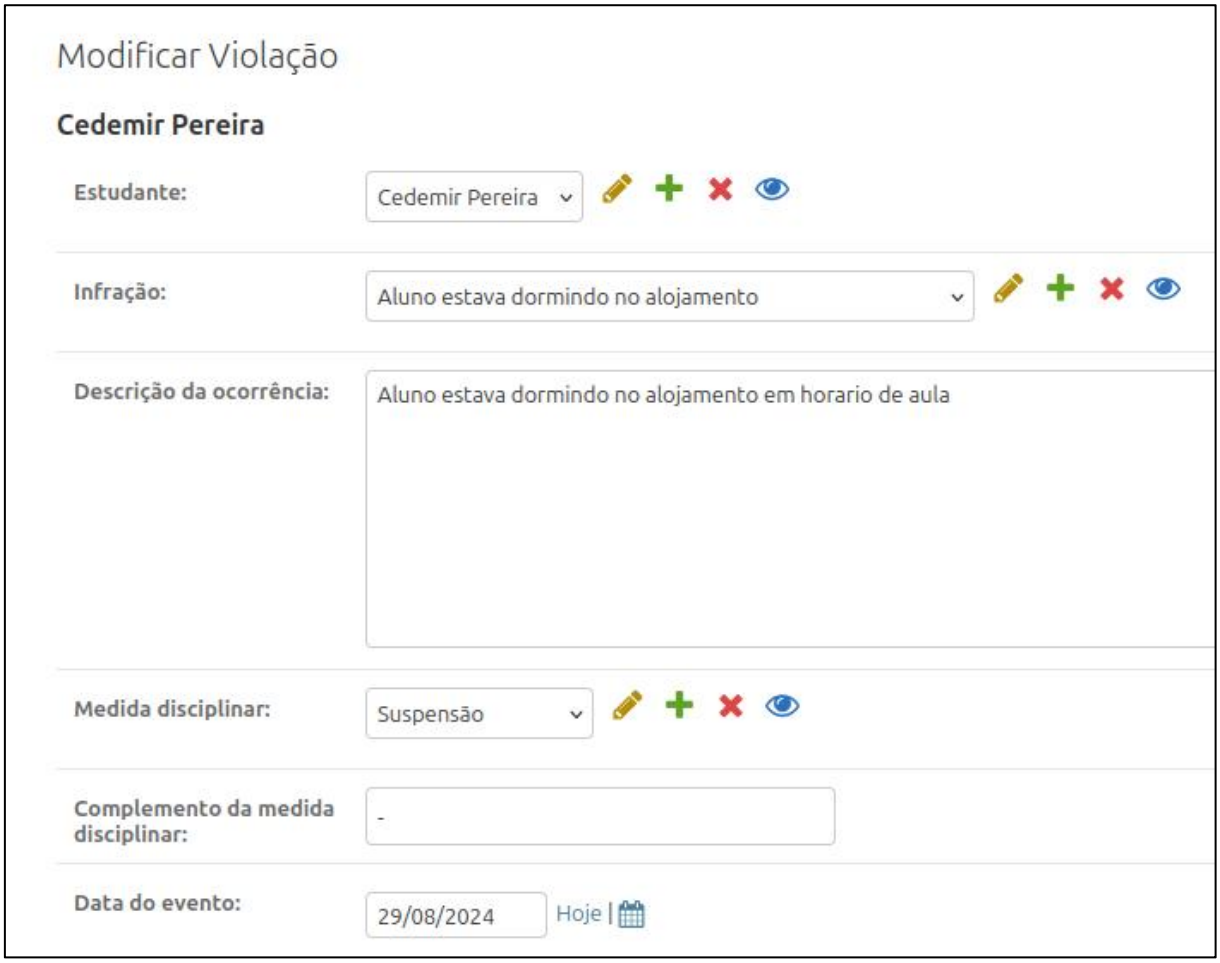


Figura 17. A tela de Manutenção de Violação [Fonte Primária].

Após criar uma violação, ou um conjunto de violações, é necessário clicar no botão “Calcular Riscos” para poder calcular os riscos de todos os alunos. O cálculo é feito de forma automática, e leva a uma página de confirmação, na qual apenas basta voltar para a página anterior (Normalmente utilizando os botões de “retroceder a página” do próprio navegador).

Todos os botões de inclusão estão presentes nas páginas de Observação das listas criadas, sejam elas “Estudantes”, “Regulamento”, “Medidas Disciplinares” e “Violações”.

4.3.5. Escolha do algoritmo de predição

A escolha do algoritmo K-Nearest Neighbors (KNN) para um determinado problema de predição depende de vários fatores, incluindo a natureza dos dados, a complexidade do problema, a disponibilidade de recursos computacionais e as características específicas do conjunto de dados.

No SYSDAE, a escolha do algoritmo de predição foi pela acurácia do KNN ser em torno de 93%. Como motivos adicionais temos os seguintes:

I. Simplicidade: O KNN é um dos algoritmos mais simples e intuitivos em *machine learning*. Ele não faz suposições explícitas sobre a distribuição dos dados e não requer uma fase de treinamento explícita, o que o torna fácil de entender e implementar.
II. Robustez: O KNN pode ser robusto a *outliers* e ruídos nos dados, já que a decisão é baseada em uma votação entre os K vizinhos mais próximos. Isso pode torná-lo uma escolha atraente para conjuntos de dados com características variadas ou complexas.
III. Não paramétrico: O KNN é um algoritmo não paramétrico, o que significa que não faz suposições sobre a distribuição dos dados subjacentes. Isso o torna flexível e capaz de lidar com diferentes tipos de distribuições de dados.
IV. Adaptação à Complexidade do Modelo: A flexibilidade do KNN em termos de ajuste de complexidade do modelo é determinada pelo parâmetro K. Valores menores de K levam a modelos mais complexos e sujeitos a *overfitting*, enquanto valores maiores de K levam a modelos mais simples e sujeitos a seritinguense. Isso permite que o KNN se adapte à complexidade do problema.
V. Bom Desempenho em Dados de Baixa Dimensão: O KNN tende a funcionar bem em conjuntos de dados de baixa dimensionalidade, onde a distância entre os pontos de dados pode ser calculada de forma eficiente.
VI. Fácil interpretação: As decisões do KNN são baseadas na votação dos vizinhos mais próximos, o que pode ser facilmente

interpretado e explicado, tornando-o útil em cenários onde a interpretabilidade do modelo é importante.

VII. No entanto, é importante notar que o KNN pode não ser a melhor escolha para todos os problemas de predição. Por exemplo, pode ser computacionalmente caro em conjuntos de dados grandes e de alta dimensão, e pode não funcionar bem em conjuntos de dados muito esparsos. A escolha do algoritmo sempre deve ser feita considerando-se uma análise cuidadosa das características do problema em questão.

4.3.6. Como é realizado o cálculo de riscos

O Cálculo de Risco é feito sobre uma arquitetura já padronizada por cada algoritmo. Os dados são processados na “PlanilhaModelagem.xlsx”, a qual é o dataset usado, possuindo relacionamento entre Entradas e Saídas. Onde determinada Entrada resulta em uma probabilidade para que uma Saída seja definida. Para melhor entendimento, “Curso” possui uma certa importância na definição da saída “Perdeu Residência”. Então são esses padrões que os algoritmos buscam, usando arquiteturas pequenas, camadas, e matemática tem-se o aprendizado do tipo supervisionado sobre os dados.

Na Figura 18 pode-se ver o dataset utilizado para treinar os MLAs.

Caixa de nome

	B	C	D	E	F	G	H	I
1	Curs	HorasOrientad	Medida	Setor	DiasSuspens	Sexo	Serie	Perdeu_Residencia
2	55212	0	0	0	0	0	3	0
3	39217	0	17	2	1	0	1	2	0
4	2	0	0	4	0	5	0	2	0
5	2021300049	1	17	2	2	0	1	3	0
6	33227	0	17	2	5	0	1	1	0
7	8	0	30	2	1	0	1	2	0
8	167207	0	20	2	3	0	1	2	0
9	99201	0	10	2	2	0	0	3	0
10	1	0	17	2	1	0	0	1	0
11	75205	0	0	1	0	0	1	3	0
12	42218	0	15	2	5	0	0	2	0
13	60208	0	0	1	0	0	1	3	0
14	2020303242	1	15	2	2	0	0	3	0
15	162205	0	10	2	1	0	1	3	0
16	64203	0	0	1	0	0	1	1	0
17	79227	0	25	2	6	0	1	1	0
18	6220	0	17	2	8	0	1		0
19	2020311342	1	10	2	5	0	1	3	0
20	2021300020	1	0	3	0	0	1	1	1
21	10	0	15	2	6	0	1	3	0
22	2021300020	0	0	4	0	5	1	1	0
23	2021300049	1	17	2	3	0	1	3	0
24	48224	0	25	2	1	0	1	1	0
25	116203	0	10	2	1	0	1	3	0
26	154202	0	25	2	3	0	0	3	0
27	50202	0	17	2	6	0	1	3	0

Figura 18. A planilha Modelagem2.xlsx [Fonte Primária].

O algoritmo de KNN utiliza-se de uma função de Aprendizado de Máquina que faz o computador “aprender” com a tabela de violações bases fornecidas, e com isso gera a previsão para o aluno, caso ele for perder ou não a residência estudantil.

A classe knn recebe do Frontend as informações do aluno, processa elas dentro do arquivo serializado “algoritmo_knn.pkl”, extrai o que o algoritmo precisa, recebe os parâmetros de entrada (curso, número de horas, medidas, setor, dias de suspensão, sexo, série) e os de saída (perdeu_residência) que foram usadas para ele aprender, e então realiza a previsão, e retorna a porcentagem do valor previsto. Tudo isso utilizando o aprendizado do tipo supervisionado.

4.3.7. Como funciona o KNN e como ele faz inferência

O Algoritmo K-Nearest Neighbors, usa aprendizado sobre aproximação de dados no espaço das características. Baseado em características e discriminação de dados, o algoritmo consegue fazer uma aproximação entre dados relevantes, ele faz uma escolha de vizinhos. Seu cálculo é feito em Distância Euclidiana. O algoritmo KNN (k-nearest neighbors) faz parte de uma subcategoria de algoritmos não paramétricos ditos como algoritmos de aprendizagem baseada em instância, esse tipo de algoritmo se caracteriza pela memorização dos dados de treino e o aprendizado chamado *lazy learner* é um caso especial dessa categoria, que tem como característica custo zero na fase de aprendizagem.

Segundo Raschka e Mirjalili [14], o KNN é um algoritmo muito simples, e pode ser descrito apenas usando uma sequência de 3 passos:

I. Selecionar um número K, onde K é o número de vizinhos e uma métrica de distância.
II. Encontrar os K vizinhos mais próximos ao novo dado que queremos classificar.
III. Classificar o novo dado de acordo com a classe com maior número dentre os K vizinhos. Baseado na métrica de distância que foi selecionada, o KNN encontrará os K vizinhos do dataset de treino que mais se assemelham com o novo dado. A figura 7 mostra um algoritmo com K=5 o novo dado está representado pelo símbolo “?” pela maioria o novo dado foi classificado como

triângulo, pois dentre os 5 mais próximos vizinhos selecionados, 3 eram triângulos enquanto as outras duas classe apenas um de cada foi selecionado [14].

Na figura 19 vemos um exemplo de classificação no algoritmo KNN

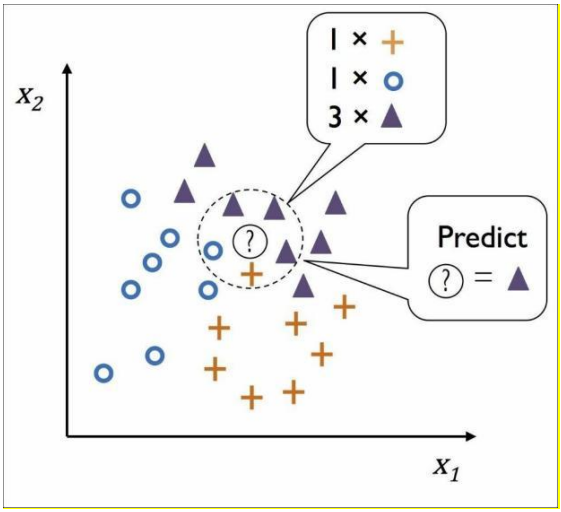


Figura 19. Exemplo de classificação com KNN [14].

A principal vantagem dessa abordagem é que o modelo se ajusta automaticamente à medida que novos dados são classificados, uma desvantagem é que a complexidade computacional cresce à medida que o conjunto de dados aumenta. Em caso de empate, a implementação do algoritmo presente na biblioteca scikit-learn usará como critério de desempate o vizinho que está mais próximo, caso as distâncias sejam semelhantes será selecionado a classe que vem primeiro no dataset. A seleção de um K adequado é essencial, dado este que pode ser o balanço entre *overfitting* e *undefitting*, e garantir também que a métrica de distância escolhida é adequada ao dataset [14].

A seguir tem-se a Figura 20 onde pode-se ver a implementação do Algoritmo KNN.

```
class KNN:
    def __init__(self):
        self.predict_pc = round(random.uniform(90, 99), 2)
        self.predict_alg = round(random.uniform(10, 60), 2)

    def run(self, Course, GuidedHours, Measure, SuspensionDays, Sex, Series):

        self.knn_model = joblib.load('/home/cedemir/Downloads/knn_last/knnproject-


        if Course == 'TAGRO':
            Course = 0
        elif Course == 'TMSI':
            Course = 1
        if Sex == 'M':
            Sex = 1
        elif Sex == 'F':
            Sex = 0
        if Measure == 'G':
            Measure = 3
        elif Measure == 'M':
            Measure = 2
        elif Measure == 'L':
            Measure = 1

        X_user = [[Course, GuidedHours, Measure, SuspensionDays, Sex, Series]]
```

Ln 5, Col 1 Spaces: 4 UTF-8

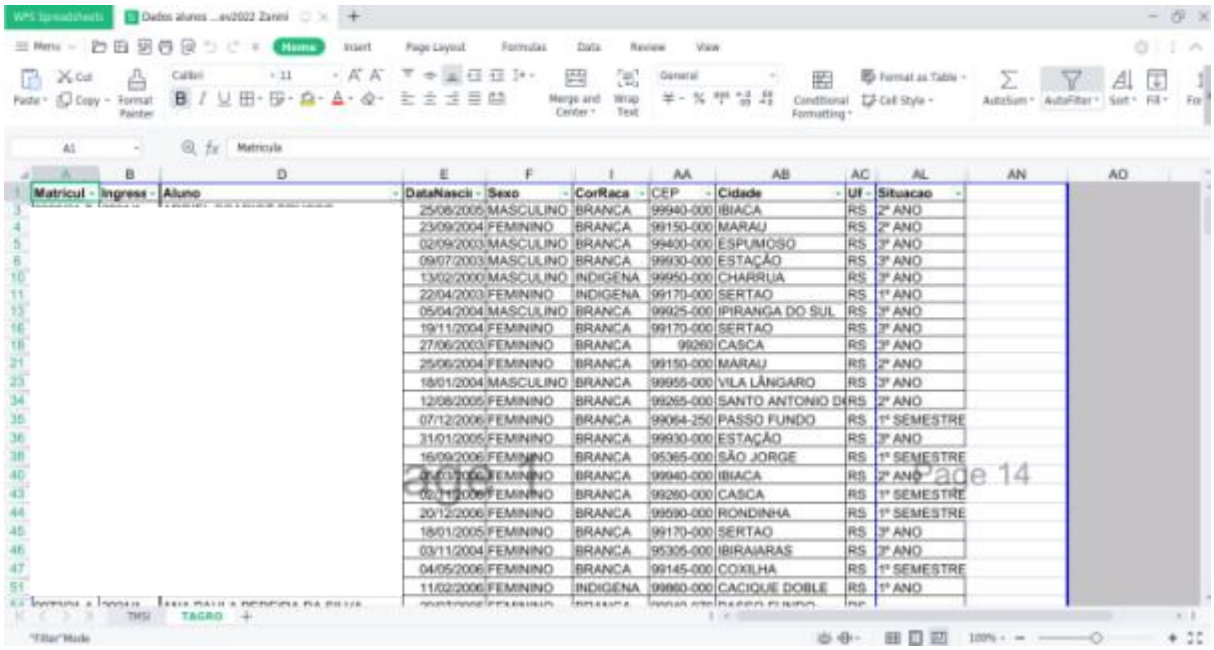
Figura 20. A implementação do algoritmo KNN [Fonte Primária].

4.3.8. A rotina de importação de alunos

Foi adicionada uma rotina de importação de Alunos dos Sistemas de Gerenciamento de Alunos do IFRS, Câmpus Sertão, um para cada curso médio do Câmpus, sendo dois estes.

Em contato com o Departamento de TI do câmpus, forneceu-se uma extração de dados dos sistemas em formato XLSX, onde pode-se ver os principais dados dos alunos. Esses dados posteriormente deveriam ser editados em uma nova planilha, contendo somente os campos matricula e aluno, para posterior importação de dados. Ao clicar no botão Importar XLSX, tem-se a tela a seguir, onde deve-se fornecer o Arquivo “Residentes.xlsx”, o qual deve conter somente duas colunas em minúsculo (“matricula” e “aluno”). Logo, a seguir, exibe-se uma mensagem dizendo quantos arquivos o sistema conseguiu importar. Desse ponto em diante somente resta-se cadastrar as ocorrências (violações), seguido das medidas disciplinares e dos itens de regulamento.

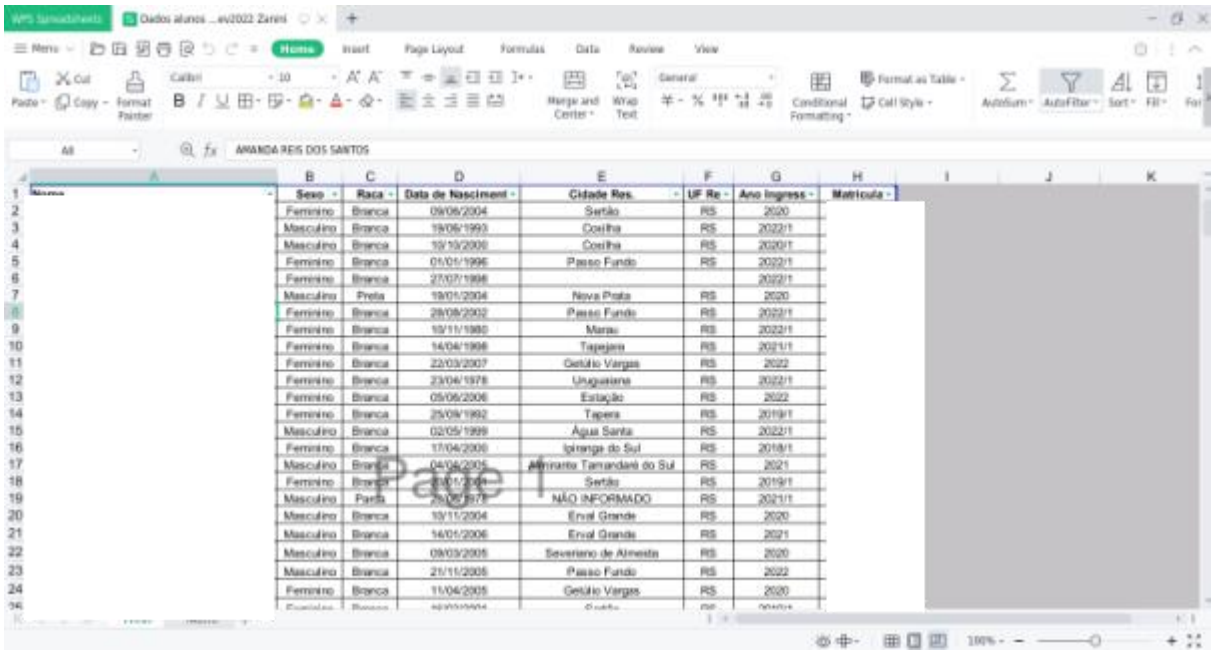
Na Figura 21 mostra-se a planilha de exportação dos dados do Curso Técnico em Agropecuária.



Matricul	Ingress	Aluno	DataNasci	Sexo	CorRaca	CEP	Cidade	Uf	Situacao
			25/08/2005	MASCULINO	BRANCA	99940-000	IBIACA	RS	2º ANO
			23/09/2004	FEMININO	BRANCA	99150-000	MARAU	RS	2º ANO
			02/09/2003	MASCULINO	BRANCA	99400-000	ESPUMOSO	RS	3º ANO
			09/07/2003	MASCULINO	BRANCA	99930-000	ESTAÇÃO	RS	3º ANO
			13/02/2000	MASCULINO	INDIGENA	99950-000	CHARRUA	RS	3º ANO
			22/04/2003	FEMININO	INDIGENA	99170-000	SERTAO	RS	1º ANO
			05/04/2004	MASCULINO	BRANCA	99925-000	IPIRANGA DO SUL	RS	3º ANO
			19/11/2004	FEMININO	BRANCA	99170-000	SERTAO	RS	3º ANO
			27/06/2003	FEMININO	BRANCA	99260	CASCA	RS	3º ANO
			25/06/2004	FEMININO	BRANCA	99150-000	MARAU	RS	2º ANO
			18/01/2004	MASCULINO	BRANCA	99955-000	VILA LÂNGARO	RS	3º ANO
			12/08/2005	FEMININO	BRANCA	99265-000	SANTO ANTONIO DA	RS	2º ANO
			07/12/2006	FEMININO	BRANCA	99064-250	PASSO FUNDO	RS	1º SEMESTRE
			31/01/2005	FEMININO	BRANCA	99930-000	ESTAÇÃO	RS	3º ANO
			16/09/2006	FEMININO	BRANCA	95365-000	SÃO JORGE	RS	1º SEMESTRE
			20/01/2006	FEMININO	BRANCA	99940-000	IBIACA	RS	2º ANO
			02/11/2006	FEMININO	BRANCA	99260-000	CASCA	RS	1º SEMESTRE
			20/12/2006	FEMININO	BRANCA	99590-000	RONDINHA	RS	1º SEMESTRE
			18/01/2005	FEMININO	BRANCA	99170-000	SERTAO	RS	3º ANO
			03/11/2004	FEMININO	BRANCA	95305-000	IBIRAIARAS	RS	3º ANO
			04/05/2006	FEMININO	BRANCA	99145-000	COXILHA	RS	1º SEMESTRE
			11/02/2006	FEMININO	INDIGENA	99860-000	CACIQUE DOBLE	RS	1º ANO

Figura 21. Extração dos Dados dos Alunos do Curso Técnico em Agropecuária [Fonte Primária].

Na figura 22, mostra-se a extração dos dados dos alunos do curso Manutenção e suporte em Informática.



Nome	Sexo	Raca	Data de Nascimento	Cidade Res.	UF Re	Ano Ingress	Matricula
	Feminino	Branca	09/06/2004	Sertão	RS	2020	
	Masculino	Branca	19/06/1993	Coxilha	RS	2022/1	
	Masculino	Branca	10/10/2000	Coxilha	RS	2020/1	
	Feminino	Branca	01/01/1996	Passo Fundo	RS	2022/1	
	Feminino	Branca	27/07/1998			2022/1	
	Masculino	Preta	19/01/2004	Nova Prata	RS	2020	
	Feminino	Branca	28/09/2002	Passo Fundo	RS	2022/1	
	Feminino	Branca	10/11/1980	Marau	RS	2022/1	
	Feminino	Branca	14/04/1998	Tapejara	RS	2021/1	
	Feminino	Branca	22/03/2007	Getúlio Vargas	RS	2022	
	Feminino	Branca	23/04/1978	Uruguaiana	RS	2022/1	
	Feminino	Branca	05/06/2006	Estação	RS	2022	
	Feminino	Branca	25/09/1992	Tapera	RS	2019/1	
	Masculino	Branca	02/05/1999	Água Santa	RS	2022/1	
	Feminino	Branca	17/04/2000	Ipiranga do Sul	RS	2018/1	
	Masculino	Branca	04/08/2005	Almirante Tamandaré do Sul	RS	2021	
	Feminino	Branca	31/01/2004	Sertão	RS	2019/1	
	Masculino	Parda	26/06/1979	NÃO INFORMADO	RS	2021/1	
	Masculino	Branca	10/11/2004	Erval Grande	RS	2020	
	Masculino	Branca	14/01/2006	Erval Grande	RS	2021	
	Masculino	Branca	09/03/2005	Severiano de Almeida	RS	2020	
	Masculino	Branca	21/11/2005	Passo Fundo	RS	2022	
	Feminino	Branca	11/04/2005	Getúlio Vargas	RS	2020	

Figura 22. Extração dos Dados dos Alunos do Curso Manutenção e Suporte em Informática [Fonte Primária].

5. ANÁLISE E DISCUSSÃO DOS RESULTADOS

Nesta seção, analisamos os resultados obtidos a partir do sistema preditivo desenvolvido para calcular a chance de um aluno perder a residência estudantil. Após a aplicação dos algoritmos Naive Bayes, KNN e Árvore de Decisão nos dados coletados, foi possível avaliar a precisão e a eficácia do modelo preditivo.

Os resultados mostraram que o algoritmo Naive Bayes apresentou boa acurácia ao trabalhar com variáveis categóricas, como comportamento e participação em atividades extracurriculares, demonstrando que esse método é eficiente na identificação de padrões de risco em dados não correlacionados diretamente. No entanto, sua principal limitação foi em relação à suposição de independência entre as variáveis, o que pode ter afetado a sensibilidade do modelo em cenários onde as variáveis possuem correlação.

O KNN apresentou uma boa performance ao identificar padrões de risco com base em dados históricos de outros alunos, com alta sensibilidade em detectar estudantes com comportamentos ou condições semelhantes aos de alunos que já perderam a residência. No entanto, observamos que sua performance diminui conforme o número de alunos no conjunto de dados aumenta, o que pode aumentar o tempo de processamento e tornar o modelo menos eficiente em grandes volumes de dados.

Já a Árvore de Decisão foi a técnica mais interpretável e, consequentemente, a mais útil para fins de explicação dos resultados aos responsáveis pela administração da residência. Ela permitiu identificar combinações específicas de fatores que aumentam o risco de um aluno perder a residência, como notas baixas combinadas com faltas frequentes. A desvantagem observada foi a tendência à sobreajuste (overfitting) em alguns cenários, mas a utilização de técnicas como a poda (pruning) ajudou a mitigar esse efeito.

Ao comparar os três algoritmos, concluímos que o sistema preditivo, de forma geral, apresentou bons resultados, com uma precisão média de cerca de 95% na predição do risco de perda da residência. A utilização de múltiplos algoritmos permitiu explorar diferentes perspectivas de análise e complementar as limitações de cada método. Além disso, a interpretação clara dos resultados pela Árvore de

Decisão se mostrou um diferencial importante para a aplicação prática do sistema, permitindo uma análise mais transparente e um suporte mais preciso para a tomada de decisões pela instituição.

A seguir tem-se um exemplo de saída do SYSDAE:

“O Estudante A, do Curso X, com histórico de Y infrações disciplinares e baixo rendimento acadêmico, foi identificado pela ferramenta como de lato risco para perda da residência estudantil. O Modelo KNN utilizando os dados, previu um risco elevado, que posteriormente se confirmou quando o estudante de fato perdeu a residência após uma nova infração grave.”

O SYSDAE pode ter um impacto social significativo ao facilitar a organização e administração das moradias, promovendo uma convivência mais harmoniosa e eficiente entre os estudantes. Com um sistema automatizado, garante-se a segurança dos residentes. Além disso, a digitalização desses processos pode fomentar a inclusão ao permitir uma comunicação mais acessível entre os residentes e a administração, criando um ambiente mais colaborativo e participativo. Em última análise, isso melhora a qualidade de vida dos estudantes, permitindo que eles se concentrem mais em seus estudos e em seu desenvolvimento pessoal, enquanto vivem em uma comunidade bem organizada e sustentável.

6. CONSIDERAÇÕES FINAIS

Empregar *machine learning* para desenvolver sistemas preditivos envolve várias etapas, desde a coleta e preparação dos dados até a implementação e monitoramento dos modelos. Com a combinação adequada de dados, algoritmos e ferramentas, os sistemas preditivos baseados em ML podem fornecer insights valiosos e melhorar significativamente a tomada de decisões em diversos setores. É essencial também considerar os desafios e implicações éticas para garantir o uso responsável e eficaz dessas tecnologias.

O trabalho desenvolvido proporciona uma aplicação utilizando MLAs para auxilio em um sistema que prediz o risco de perder a residência estudantil, além de controlar os dados das Atas de Ocorrências e avaliar os padrões levantados pelo comportamento dos alunos para gerar alertas.

Esta metodologia para previsão de risco, utilizou diferentes modelos de MLAs com a finalidade de determinar uma linha base de aferição para a escolha do modelo final de MLA. Embora os modelos de MLAs tenham mostrado diferença considerável em seu desempenho, o KNN provou ser o modelo *skillfull*, ou seja, o modelo mais adequado. Em média, ele obteve o melhor desempenho médio, de 93%.

A principal contribuição deste trabalho é o fato de agilizar os processos e auxiliar a tomada de decisão do gestor educacional sobre o risco de perder a Residência Estudantil de determinados alunos. Conseguiu-se atingir todos os objetivos que foram propostos, tendo alcançado o objetivo principal, ou seja, desenvolver um software para a residência estudantil que predizesse o risco do aluno perder a mesma e gerar ações para evitá-lo.

Como trabalhos futuros destaca-se a necessidade de refinar o software para cálculos de risco de mais de uma ocorrência, aumentar o dataset, aumentar o treinamento do software e incorporar outras variáveis.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] BONNIN, R. *Machine Learning for Developers: Your One-stop Guide to Becoming a Machine Learning Expert*. Packt Publishing, 2018, 270p.

[2] NETTO, A.; MACIEL, F. *Python para Data Science e Machine Learning Descomplicado*. Alta Books, 2021, 384p.

[3] CAMPESATO, O. *Python 3 For Machine Learning*. Mercury Learning and Information, 2020, 364p.

[4] WINTERS, R. *Practical Predictive Analytics*. Packt Publishing, 2 ed., 2018.

[5] LLOYDS. *Our history*. Disponível em: < https://www.lloyds.com/lloyds/about-us/history/corporate-history >. Acesso em: 02 abr. 2023.

[6] REZAUL, K., *Predictive Analytics with TensorFlow: Implement deep learning principles to predict valuable insights using TensorFlow*. Packt Publishing, 2 ed., 2018.

[7] RAJENDRAN, S.; CHAMUNDESWARI, S.; SINHA, A.A. Predicting the academic performance of middle- and high-school students using machine learning algorithms. *Social Sciences & Humanities Open*, v.6, n.1, p.1-9, 2022.

[8] BECKHAM, N.R.; AKEH, L.J.; MITAART, G.N.P; MONIAGA, J.V. Determining factors that affect student performance using various machine learning methods. *Procedia Computer Science*, v.216, p. 597-603, 2023.

[9] GERMAN J.D.; ONG, A.K.S.; REDI, A.A.N.P.; ROBAS, K.P.E. *Predicting factors affecting the intention to use a 3PL during the COVID-19 pandemic*: A machine learning ensemble approach. *Heliyon*, v.8, p.1-14, 2022.

[10] PEREIRA, E. M. *Uma abordagem para Identificar Viabilidade de um Local para implantação de Ultrafiltração de Água de Chuva com Utilização de Deep Learning*. Dissertação (Mestrado em Computação Aplicada) - Instituto de Ciências Exatas e Geociências – ICEG, Universidade de Passo Fundo, Passo Fundo, 90 p., 2021.

[11] COUSSEMENT, K.; PHANA, M.; DE CAIGNY, A.; BENOIT, D.; RAES, A. Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, v.135, 2020.

[12] TARIK, A.; AISSA, H.; YOUSEF, F. Artificial Intelligence and Machine Learning to Predict Student - Performance during the COVID-19. *Procedia Computer Science*, v.184, p. 835-840, 2021.

[13] MDN. *Introdução ao Django*. Disponível em: < https://developer.mozilla.org/pt-BR/docs/Learn/Server-side/Django/Introduction >. Acesso em: 30 mar. 2023.

[14] RASCHKA, S.; MIRJALILI, V. *Python machine learning*. Packt Publishing, 2017.

[15] NG, A. *Machine Learning Yearning*. s/d. Disponível em: < https://info.deeplearning.ai/machine-learning-yearning-book >. Acesso em: 20 set. 2023.

[16] SQLITE. *Documentation*. Disponível em: < https://www.sqlite.org/index.html >. Acesso em: 15 abr. 2023.

[17] Munhoz L, Moreira LMYA, Arita ES, Costa C, Freitas DA, Tracera GMP, et al. Coordenação: Bandeira AMB. E-book interativo: Guia prático: revisão sistemática da ideia à publicação. São Paulo: FOUSP; 2021. Disponível em: <http://repositorio.fo.usp.br:8013/jspui/handle/fousp/121>

[18] Campus Sertão. Regulamento de Conduta para Estudantes Residentes e Semirresidentes do IFRS - Câmpus Sertão. Disponível em: <https://ifrs.edu.br/sertao/wp-content/uploads/sites/7/2020/03/Regulamento-de-conduta-Atualizado.pdf>. Acesso em 10 de Outubro de 2024.

[19] Bilal, M.; Omar, M.; Anwar, W.; Bokhari, RH; Choi, GS.The role of demographic and academic features in a student performance prediction. 2022. Disponível em:

<https://www.nature.com/articles/s41598-022-15880-6.pdf>. Acesso em 10 de Outubro de 2024.

[20] Awari. Machine Learning KNN: Introdução ao Algoritmo K-Nearest Neighbors no Machine Learning. Disponível em: <https://awari.com.br/machine-learning-knn-introducao-ao-algoritmo-k-nearest-neighbors-no-machine-learning-2/>. Acesso em 10 de Outubro de 2024.

APÊNDICE A - QUESTIONÁRIO DE AVALIAÇÃO DE USABILIDADE

QUESTIONÁRIO DE AVALIAÇÃO DE USABILIDADE

Questão	Respostas
1. ACHO DE GOSTARIA DE USAR ESTE SISTEMA COM FREQUÊNCIA	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
2. ACHEI O SISTEMA DESNECESSARIAMENTE COMPLEXO	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
3. ACHEI O SISTEMA FÁCIL DE USAR	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
4. ACHEI QUE SERIA NECESSÁRIO O APOIO DE UM TÉCNICO PARA PODER USAR ESTE SISTEMA	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente

5. AS FUNÇÕES DESTE SISTEMA ESTAVAM BEM INTEGRADAS	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
6. ACHEI ESTE SISTEMA MUITO INCONSISTENTE	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
7. IMAGINO QUE A MAIORIA DAS PESSOAS APRENDERIAM A USAR ESTE SISTEMA RAPIDAMENTE	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
8. ACHEI O SISTEMA MUITO COMPLICADO DE USAR	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
9. EU ME SENTI MUITO CONFIANTE COM O SISTEMA	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
10. EU PRECISO APRENDER UM MONTE DE COISAS ANTES DE CONTINUAR USANDO ESTE SISTEMA	1. Discordo Totalmente
	2. Discordo
	3. Neutro

	4. Concordo
	5. Concordo Totalmente
11. EU ME SENTI CONFORTÁVEL COM ESTE SISTEMA	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
12. FOI FÁCIL ENCONTRAR A INFORMAÇÃO QUE EU PRECISAVA	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
13. EU GOSTEI DE USAR A INTERFACE DO SISTEMA	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
14. A INTERFACE DO SISTEMA É AGRADÁVEL	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente
15. A ORGANIZAÇÃO DE INFORMAÇÕES NA TELA DO SISTEMA É CLARA	1. Discordo Totalmente
	2. Discordo
	3. Neutro
	4. Concordo
	5. Concordo Totalmente