UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Tarcísio Barbosa da Costa

Sistema de Suporte à Criação de Modelos de Classificação para a Previsão de Evasão no Ensino Superior

Recife

2024

Tarcísio Barbosa da Costa

Sistema de Suporte à Criação de Modelos de Classificação para a Previsão de Evasão no Ensino Superior

Trabalho apresentado ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Andrêza Leite de Alencar

Coorientador: Gabriel Alves de Albuquerque Junior

Recife

2024

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

B238s Costa, Tarcísio Barbosa
Sistema de Suporte à Criação de Modelos de Classificação para a Previsão de Evasão no Ensino Superior / Tarcísio Barbosa Costa. - 2024.
45 f. : il.

Orientadora: Andreza Leite de Alencar.
Coorientador: Gabriel Alves de Albuquerque Junior.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2024.

1. Evasão. 2. Mineração de Dados. 3. Analytics. I. Alencar, Andreza Leite de, orient. II. Junior, Gabriel Alves de Albuquerque, coorient. III. Título

CDD 004

Às memórias de minha avó, Alice Barbosa da Silva Carmo, meus avôs, Ailton Vieira da Costa e José do Carmo da Silva Filho, e todos os familiares que fizeram parte do meu crescimento e hoje deixam saudades.

Às famílias Barbosa e Costa: pais, irmã, tios, primos e mais, por sempre torcerem pelo meu sucesso e terem moldado meu caráter.

À minha avó, Maria do Carmo Vieira da Costa, a quem devo respeito imensurável pela criação, pelo carinho e pelos conselhos de vida.

Agradecimentos

Agradeço …

À Prof.ª Andrêza Leite, pelas orientações, revisões, conteúdo ensinado e o apoio por toda minha jornada no curso, desde as primeiras aulas até este trabalho.

Ao Prof. Gabriel Alves, pelas reuniões, mentorias, dicas, conselhos e acompanhamentos não só no desenvolvimento deste trabalho, mas também nos estágios.

A meus pais, Austregezilo Vieira da Costa Sobrinho e Maria de Fátima Barbosa, por terem me ensinado a valorizar a educação, terem me dado esta oportunidade de formação e me encorajarem a seguir em frente nos momentos de insegurança.

À minha irmã, Tássia, que, mesmo longe, me ajudou diretamente na elaboração deste trabalho e continua me inspirando a ser mais do que sou hoje.

Aos professores do Departamento de Computação, pelo louvável trabalho como docentes e por terem feito parte não só da minha formação, mas a de incontáveis outros estudantes da Universidade.

À minha colega Lhaíslla Cavalcanti, por todos os dias que passamos nos ajudando no desenvolvimento dos respectivos trabalhos.

Aos amigos que fiz durante minha formação escolar e na universidade, por tudo que compartilhamos e ainda vamos compartilhar.

À equipe de gestores da UFRPE como um todo, por manterem a instituição em funcionamento mesmo em épocas de dificuldade.

…e a todos que fizeram toda a diferença ao participarem, direta ou indiretamente, desta etapa em minha vida.

“A curiosidade nos conduz a novos caminhos. Siga em frente.”
(Walt Disney)

Resumo

A evasão estudantil é um dos maiores desafios a serem enfrentados por Instituições de Ensino Superior. A fim de mitigá-la, as instituições elaboram ferramentas para monitoramento e análise deste fenômeno. Uma das metodologias existentes para tal é a identificação de características de estudantes que levam à evasão, e uma das ferramentas construídas é o SABIA: um dashboard virtual responsável por dar suporte à gestão baseada em evidências, aliado a conceitos de *Learning/Academic Analytics* e *Business Intelligence*. Este trabalho expande o SABIA através de uma nova página capaz criar modelos de aprendizado supervisionado personalizáveis pelo usuário, oferecendo análises de características estudantis e realizando previsões da situação final do discente baseadas nas mesmas. As informações obtidas pelos modelos proporcionam a identificação de fatores de risco em perfis discentes e auxiliam os gestores da instituição no desenvolvimento de diretrizes para a adoção de medidas contra a evasão.

Palavras-chave: Evasão, Mineração de Dados, *Analytics*.

Abstract

Student dropout is one of the greatest challenges faced by university degree institutions. In order to mitigate it, those institutions develop moitoring and analysis tools regarding this phenomenon. One of many existing methodologies to do so is the recognition of student characteristics that leads to dropout, and ond of many existing tools is SABIA: a virtual dashboard responsible for supporting evidence-based management allied to concepts like Learning/Academic Analytics and Business Intelligence. This work expands SABIA through a new page able to create user-customizable supervisioned learning models, offering feature analysis from students and predicting their final status based on those features. Information obtained through those models enables the recognition of risk features on student profiles and assists managers on providin guidelines for applying countermeasures against dropout.

Keywords: *Dropout, Data mining, Analytics.*

Lista de ilustrações

Figura 1 – Filtros e dados disponibilizados no painel principal do SABIA. 12
Figura 2 – Representação de um modelo *Naive Bayes* (A) e uma Rede Bayesiana (B). 19
Figura 3 – Exemplo simples de uma árvore de decisão voltada para diagnósticos. 19
Figura 4 – Exemplo genérico de uma *Random Forest* para Classificação/Regressão. 20
Figura 5 – Aplicação de uma função de mapeamento $\phi(x)$ em uma SVM Não-Linear. 21
Figura 6 – Exemplo da classificação de um elemento num algoritmo KNN de $k = 3$. Neste exemplo, identifica-se o elemento como Classe B. . . . 21
Figura 7 – Fluxograma com as etapas da página até a exibição dos resultados. 25
Figura 8 – Ilustração de um *dataset* genérico e transformações baseadas em técnicas de balanceamento. 27
Figura 9 – Exemplo de gráfico de *feature importance* para um modelo com um pequeno conjunto de características. 32
Figura 10 – *Frontpage* do SABIA. 33
Figura 11 – Seção reservada para montagem do modelo no painel do SABIA. . 34
Figura 12 – Seção do painel do SABIA contendo informações sobre um modelo criado pelo usuário. 34
Figura 13 – Seção do painel do SABIA dedicada à previsão da situação final do estudante. 35
Figura 14 – Gráfico de *feature importance* do modelo definido. 38

Lista de tabelas

Tabela 1 – Comparação entre métricas de cada classificador. 37
Tabela 2 – Comparação entre métricas de cada algoritmo de balanceamento para o KNN. 37
Tabela 3 – Estimativas de probabilidade de evasão para estudantes baseadas em curso e duração de vínculo. 38
Tabela 4 – Estimativas de evasão para estudantes de um curso específico, baseadas em duração de vínculo e período de ingresso. 39

Lista de abreviaturas e siglas

IES	Instituição de Ensino Superior
SABIA	*System of Academic Business Intelligence and Analytics*
LA	*Learning Analytics*
AA	*Academic Analytics*
BI	*Business Intelligence*
ODG	Observatório de Dados da Graduação
UFRPE	Universidade Federal Rural de Pernambuco
BPM	*Business Performance Management*
DAG	*Directed Acyclic Graph*
LGN	Lei dos Grandes Números
SVM	*Support Vector Machines*
KNN	*K-Nearest Neighbors*
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
ETL	*Extract, Transform and Load*
SMOTE	*Synthetic Minority Over-sampling Technique*
ENN	*Edited Nearest Neighbors*
XAI	Inteligência Artificial Explicável

Sumário

Lista de ilustrações **7**

1 INTRODUÇÃO **11**
1.1 Motivação **12**
1.2 Objetivos **13**
1.3 Seções do Trabalho **13**

2 REFERENCIAL TEÓRICO **15**
2.1 Evasão Estudantil **15**
2.2 *Business Intelligence* **16**
2.3 Mineração de Dados Educacionais **16**
2.4 Modelos de Aprendizado de Máquina **17**
2.4.1 Modelos Bayesianos 17
2.4.2 *Random Forest* 18
2.4.3 Máquinas de Vetores de Suporte 20
2.4.4 *K-Nearest Neighbors* 21

3 TRABALHOS RELACIONADOS **23**

4 PROPOSTA **25**
4.1 Ingestão e Tratamento dos Dados **25**
4.2 Criação do Modelo **26**
4.3 Aplicação do Modelo **28**

5 MÉTODO E FERRAMENTAS **29**
5.1 Modelos e Balanceamento **29**
5.2 Apresentação dos Dados **30**

6 RESULTADOS ALCANÇADOS **33**
6.1 Interface e Lógica do Sistema **33**
6.2 Modelos e Previsões **35**
6.2.1 Criação do Modelo 36
6.2.2 Análise do Modelo 38

7 CONCLUSÕES FINAIS **40**

REFERÊNCIAS **41**

1 Introdução

No processo de transição do ensino médio ao superior, os estudantes são expostos a uma realidade completamente diferente à que estavam acostumados, visto que são introduzidos a uma diversidade de novos conteúdos e dinâmicas de ensino ao longo desta nova etapa de suas vidas. Nesta nova realidade, a compreensão e o domínio destes conteúdos variam de acordo com o contexto educacional prévio ao ingresso do discente nas Instituições de Ensino Superior (IES).

Estes diferentes níveis de contexto educacional podem causar um impacto significativo no vínculo entre certos perfis de estudantes e as instituições, levando à evasão: o fenômeno em que um estudante abandona um curso ou instituição de ensino antes de sua devida conclusão (VIANA; SANTANA; RABÊLO, 2022).

Impactos negativos, como a redução da possibilidade de ascensão social, comprometem o desenvolvimento social e econômico dos estudantes que abandonam sua formação acadêmica (LOBO, 2012). Além dos impactos individuais, que vão dos estudantes até seus familiares (que não tiveram o sonho da conclusão realizado) e professores (que não alcançaram sua meta como educador), a evasão acarreta problemas para a sociedade como um todo devido às perdas sociais e econômicas (FILHO et al., 2007).

No contexto das IES, os prejuízos causados pela evasão também são graves. Dentre eles, em universidades públicas, destaca-se a existência de vagas ociosas e prejuízos ao orçamento dessas instituições, uma vez que a perda de estudantes implica em menos recursos financeiros destinados aos programas e infraestrutura (FILHO et al., 2007), resultando num aumento de custo por aluno formado e reduzindo o impacto social causado pela instituição.

Como resposta à evasão, IES desenvolvem diferentes ferramentas com abordagens distintas em busca de mitigar os danos causados por este fenômeno. Dentre elas, este trabalho contempla o *System of Academic Business Intelligence and Analytics* (SABIA)[1], um *dashboard* que dá suporte à gestão baseada em evidências agregando conceitos de *Learning Analytics* (LA), *Academic Analytics* (AA) e *Business Intelligence* (BI) (MARQUES et al., 2023).

Desenvolvido pela equipe do Observatório de Dados da Graduação da Universidade Federal Rural de Pernambuco (ODG-UFRPE), o SABIA disponibiliza páginas públicas, com dados voltados à graduação, e páginas restritas a professores, gestores e técnicos da instituição. Cada uma destas páginas apresenta informações específicas,

[1] https://ufrpe-odg.github.io/site/sabia.html

de indicadores de qualidade da instituição e número de discentes ativos/inativos até tabelas voltadas para taxa de sucesso de disciplinas e cálculo de reoferta de vagas. A Figura 1 apresenta uma das páginas disponibilizadas pelo SABIA.

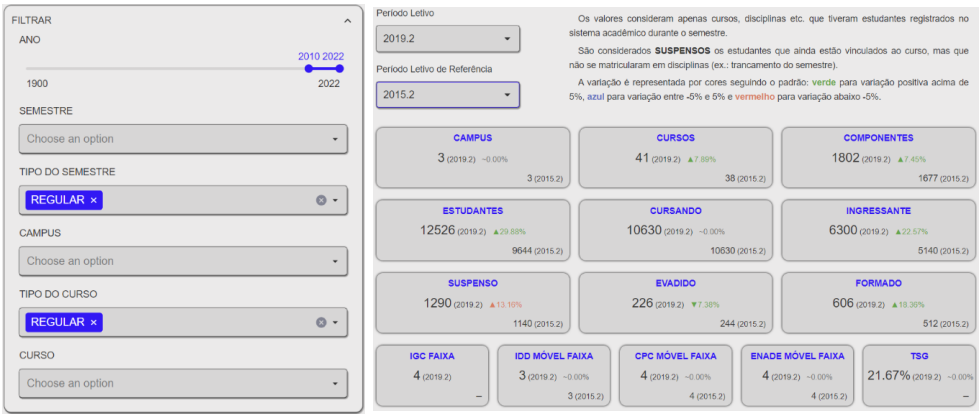


Figura 1 – Filtros e dados disponibilizados no painel principal do SABIA.
Fonte: (MARQUES et al., 2023).

1.1 Motivação

Nos últimos anos, percebe-se um aumento na taxa de evasão em IES. Um dos fatores para este fenômeno se dá pela quantidade desproporcionalmente maior de políticas de incentivo ao acesso às instituições, quando comparadas ao número de ações voltadas à permanência e melhoria de chances de sucesso dos estudantes (BRITO; MELLO; ALVES, 2020).

Em estudos anteriores, características dos perfis de estudantes foram consideradas um dos fatores determinantes para sua permanência nas instituições (NUNES, 2021). Diante desse cenário, é fundamental que os gestores identifiquem os fatores que influenciam a evasão de acordo com os perfis estudantis da sua instituição e implementem medidas e estratégias para que seus estudantes mantenham o vínculo, garantindo seu sucesso acadêmico.

Para facilitar a identificação de fatores que levam à evasão estudantil, foram aplicadas pelas instituições diferentes metodologias, como entrevistas e pesquisas (NUNES, 2021). Nos últimos anos, entretanto, evidencia-se um aumento na tendência de aderirem a soluções tecnológicas como sistemas e algoritmos para automatizar o processo de abordagens matemáticas e estatísticas (GONÇALVES; SILVA; CORTES, 2018; FILHO; VINUTO; LEAL, 2020).

Dentre as ferramentas tecnológicas criadas pelas IES, o SABIA apresenta-se como uma abordagem eficiente e distinta ao aplicar conceitos de IA, AA e BI na mineração e análise de dados. O sistema tem como objetivo, dentre outros, auxiliar direta-

mente coordenadores de cursos e gestores da UFRPE a identificar fatores-chave de desempenho e, através de diferentes resultados (filtros, tabelas, gráficos, etc.), elaborar políticas institucionais contra a evasão (MARQUES et al., 2023). Dada a infraestrutura já disponibilizada pelo sistema e metas alinhadas com as deste trabalho, torna-se apropriada a sua utilização para desenvolvimento de uma ferramenta para identificação de perfis estudantis propensos à evasão.

1.2 Objetivos

Como principal objetivo deste trabalho, busca-se criar um novo painel no SABIA, com finalidade de aferir riscos de evasão estudantil baseado em características dos estudantes, bem como avaliar o impacto de mudanças em variáveis relacionadas ao discente em sua trajetória no curso.

Foram determinados os seguintes objetivos específicos, a fim de traçar metas para assegurar que esse destino final seja alcançado:

1. Compreender o problema da evasão no ensino superior e formas de estudá-lo.
2. Procurar, avaliar e selecionar modelos de aprendizado supervisionado para o sistema.
3. Construir um sistema de criação de modelos ajustáveis para análise de características de estudantes.
4. Fornecer gráficos e tabelas com métricas referentes aos modelos criados pelo usuário.
5. Fornecer previsões, baseadas nos modelos construídos, sobre a situação final do perfil de estudante analisado.
6. Entregar a versão final do sistema em forma de uma página nova no SABIA.

1.3 Seções do Trabalho

Desta forma, o trabalho divide-se nos seguintes capítulos: O Capítulo 2 apresenta o Referencial Teórico, que introduz e descreve os fundamentos-base para compreensão e execução deste trabalho. O Capítulo 3 contém os Trabalhos Relacionados, onde é feita uma análise e revisão de suas contribuições. O Capítulo 4 apresenta a Proposta para o desenvolvimento do sistema, do tratamento dos dados até a sua apresentação na tela. O Capítulo 5 apresenta Método e Ferramentas, descrevendo os métodos utilizados para o funcionamento do painel de acordo com a proposta. O Capítulo 6

descreve os Resultados Alcançados pelo painel e levanta discussões sobre alguns deles. O Capítulo 7 finaliza o trabalho com Conclusões, observações e reflexões sobre futuras implementações no sistema.

2 Referencial Teórico

Nesta seção, são apresentados e introduzidos os conceitos em que este trabalho se baseou para chegar aos seus resultados.

2.1 Evasão Estudantil

Um dos maiores problemas quando se trata da compreensão da evasão é como ela é definida: existem diferentes formas de interpretar o problema, trazendo inconsistências de terminologia. Essas inconsistências são problemáticas, considerando que a definição de evasão é o que vai determinar medidas, abordagens e pesquisas sobre o fenômeno (XAVIER; MENESES, 2020). A comissão especial para estudo da evasão composta pelo ANDIFES, ABRUEM, SESu e MEC, por exemplo, classifica a evasão em três graus diferentes (MEC, 2014 apud UNIVERSIDADES; ESPECIAL; BORDAS, 1996):

1. **Evasão de Curso:** a saída definitiva do aluno de seu curso de origem, sem concluí-lo.
2. **Evasão de Instituição:** o desligamento da instituição na qual o estudante estava matriculado.
3. **Evasão do Sistema:** quando o aluno abandona o ensino de modo geral.

Adicionalmente, observa-se que nem todos os trabalhos sobre evasão levam o período de tempo em consideração, às vezes incluindo dados de estudantes que estão apenas em uma pausa temporária, como em casos de trancamento do curso.

Como forma de definir concretamente o modo que este trabalho irá abordar o fenômeno da evasão, foi escolhida a definição para cursos de graduação baseada na comissão especial: “a saída definitiva do aluno de seu curso de origem, sem concluí-lo” (UNIVERSIDADES; ESPECIAL; BORDAS, 1996). A Comissão também evidencia, dentre outras, a necessidade da identificação de causas internas e externas para que se alcance um conhecimento mais complexo e confiável do fenômeno.

Imprecisões também existem em termos relacionados à evasão (XAVIER; MENESES, 2020), tornando ainda mais difícil a compreensão de conceitos-chave para o estudo do problema. Um fenômeno contrário ao da evasão é a retenção estudantil, que pode tanto ser definido como a permanência do estudante na IES (SANTINI; GUIMARÃES; SEVERO, 2014) quanto o prolongamento da estadia do estudante no

curso, por um tempo maior do que o esperado (LAMERS; SANTOS; TOASSI, 2017). Este trabalho abordará o conceito de retenção de acordo com a segunda definição.

Os fenômenos da evasão e retenção, apesar de conceitualmente opostos, são relacionados. De acordo com o MEC (2014), a retenção estudantil é um dos principais fatores que propiciam a evasão e, além disso, compartilha com a evasão vários fatores individuais, internos e externos que as ocasionam. Esta relação evidencia a necessidade da compreensão de conceitos relacionados à evasão, como a retenção, para um maior domínio sobre a compreensão do problema como um todo.

2.2 *Business Intelligence*

O primeiro conceito-chave referete à elaboração do SABIA é o *Business Intelligence*. BI é um termo que abrange um conjunto de aplicações, plataformas, ferramentas e tecnologias que dão apoio direto a organizações, proporcionando o acesso interativo a dados e a sua manipulação, transformando-os em informação. (TURBAN et al., 2009; RAISINGHANI, 2004). Desta forma, o BI desempenha o papel de um facilitador, permitindo que a organização trabalhe de maneira mais inteligente e tome melhores decisões através do uso das informações obtidas. (BIANCHI et al., 2022 apud LARSON; CHANG, 2016).

Segundo Turban et al. (2009), uma plataforma de BI consiste de quatro componentes-chave: um "armazém de dados" (*Data Warehouse*), um conjunto de ferramentas para a manipulação dos dados armazenados, *Business Performance Management* (BPM) para monitoramento e análise de desempenho e, por fim, uma interface de usuário para apresentar os resultados obtidos.

No contexto das IES, uma forma comum de se aplicar o BI é a criação de *dashboards* para monitoramento de dados e compreensão da informação. Para os gestores dessas instituições, a adoção do BI é fundamental para proporcionar ferramentas de auxílio à gestão acadêmica baseada em evidências (BIANCHI et al., 2022).

2.3 Mineração de Dados Educacionais

Se o BI é o responsável por proporcionar o armazenamento e a interface de dados e informação, o *Analytics* é responsável pela análise dos dados e sintetização do conhecimento. Das técnicas de mineração utilizadas no contexto educacional, o SABIA baseia-se principalmente no *Learning Analytics* e no *Academic Analytics*.

O LA é uma estratégia de mineração de dados voltada ao desenvolvimento do discente. Seu processo pode ser interpretado como um ciclo que se inicia na introdução dos estudantes ao ambiente, parte para a geração e captura de dados de/pelos

estudantes ao longo do processo de aprendizagem, alcança a fase de métricas e análises dos dados obtidos e conclui-se ao serem implementadas intervenções baseadas nas conclusões de análise (CLOW, 2012). O processo se repete *ad infinitum*, considerando que estudantes constantemente entram e saem do ambiente educacional, recontextualizando tanto os perfis estudantis quanto o ambiente.

Enquanto o LA é mais voltado para o discente e seu processo de aprendizagem, o AA é mais abrangente, visando analisar dados dentro de todo o escopo da IES através de conceitos como Educação Corporativa e Gestão Educacional orientada a dados (CAMPOS; FONSECA, 2023).

De acordo com Andrade e Ferreira (2016), pode-se identificar cinco etapas principais no desenvolvimento de um sistema utilizando AA:

- **Acesso:** Disponibilização dos dados da instituição (internos e externos).
- **Transformação:** Preparação e tratamento dos dados.
- **Analytics:** Análise dos dados tratados através de algoritmos preditivos e/ou descritivos.
- **Visualização:** Apresentação dos resultados dos algoritmos via ferramentas, como gráficos e tabelas.
- **Exploração:** Compartilhamento de resultados e tomada de decisão entre gestores.

O uso destas estratégias foi adotado no SABIA a fim de adaptar o estudo ao contexto educacional. Este trabalho estará mais voltado à abordagem do AA, lidando com dados institucionais direcionados ao corpo docente e aos responsáveis pela IES.

2.4 Modelos de Aprendizado de Máquina

Esta seção explica, de forma resumida, o funcionamento dos algoritmos identificados como promissores para a primeira versão do sistema desenvolvido.

2.4.1 Modelos Bayesianos

Dentro do contexto dos sistemas que agem racionalmente, duas abordagens principais podem ser utilizadas: raciocínio lógico e raciocínio probabilístico (MARQUES; DUTRA, 2002). Em situações onde há domínio prévio do escopo do conhecimento, o raciocínio lógico é uma ferramenta poderosa para cálculos, mas pode não ser eficiente em situações onde não se garante o total conhecimento sobre o escopo do problema.

No contexto do SABIA, que trata de situações de incerteza ao abordar o problema da evasão no ensino superior (i.e. fatores que levam os estudantes à evasão), pode-se afirmar que a abordagem do raciocínio probabilístico pode ser benéfica para o sistema.

Com este conhecimento, foi escolhido o pensamento bayesiano como base para análise da eficiência do raciocínio probabilístico no sistema. Para realizar os cálculos, os modelos baseiam-se no Teorema de Bayes [Equação 2.1, (DAVIES, 1988)], publicado pelo matemático Thomas Bayes em 1763, onde A e B são eventos, as funções P(A) e P(B) são as probabilidades a priori de A e B, respectivamente, e P(A|B) e P(B|A) são as probabilidades a posteriori de um evento condicional ao outro.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

Há ainda outra forma de se interpretar e escrever o teorema de Bayes, como mostra a Equação 2.2 (DAVIES, 1988).

$$P(A|B)P(B) = P(A \cap B) = P(B \cap A) = P(B|A)P(A) = P(A|B)P(B) \tag{2.2}$$

O modelo a ser selecionado varia de acordo com a relação de dependência entre as características (*features*) presentes no mesmo. Temos, por exemplo, as Redes Bayesianas, que realizam a abordagem do raciocínio probabilístico através da construção de modelos probabilísticos em forma de Grafos Acíclicos Direcionados (Do inglês, *Directed Acyclic Graph* - DAG) incluindo informação quantitativa (probabilidades) e qualitativa (relações de dependência) entre as variáveis (LADEIRA; VICARI; COELHO, 1999) apresentadas na rede.

Alternativamente, temos o classificador *Naive Bayes*: nomeado com o adjetivo em inglês *naive*, que significa ingênuo, este classificador apresenta a premissa de que não existe dependência entre as *features* a serem classificadas (BATISTA; BAGATINI; FROZZA, 2018). Sua simplicidade o torna altamente escalável e menos custoso para execução, sem comprometer a precisão das análises em modelos com *features* sem dependência concreta.

A Figura 2 ilustra ambas as abordagens em forma de grafos, e mostra suas principais diferenças.

2.4.2 *Random Forest*

No contexto do raciocínio lógico, onde conclusões são tomadas racionalmente através do conhecimento prévio sobre o problema (MARQUES; DUTRA, 2002) as Árvores de Decisão são exemplos simples e eficientes: elas são algoritmos que podem

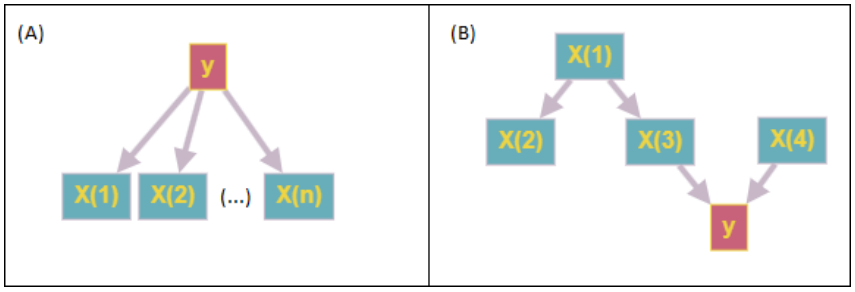


Figura 2 – Representação de um modelo *Naive Bayes* (A) e uma Rede Bayesiana (B).
Fonte: (SHAHINFAR et al., 2014) (Adaptado)

ser representados como fluxogramas, construindo um caminho lógico através de nós de decisão. Um exemplo de Árvore de Decisão está na Figura 3. A função do algoritmo é percorrer o fluxograma até encontrar uma resposta, ou seja, chegar num nó final da árvore e atribuir resultados de acordo com o que foi encontrado em sua busca.

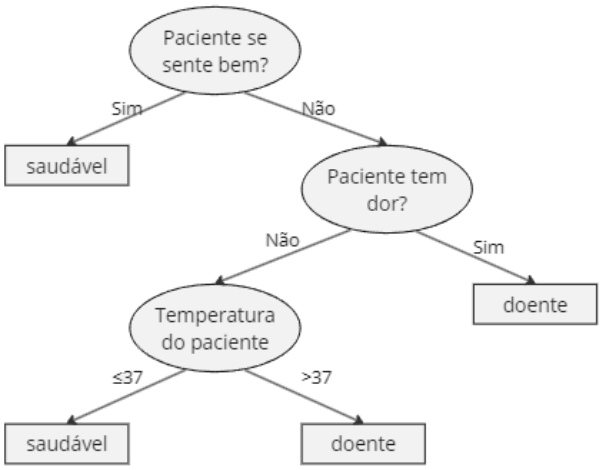


Figura 3 – Exemplo simples de uma árvore de decisão voltada para diagnósticos.
Fonte: (MONARD; BARANAUSKAS, 2003b) (Adaptado)

Um problema comum na aplicação de Árvores de Decisão é o *overfitting*, que ocorre quando o modelo apresenta resultados enganosamente altos ao ajustar-se em excesso ao conjunto de treinamento, mas falha em identificar corretamente situações diferentes em dados novos (MONARD; BARANAUSKAS, 2003a).

Uma forma de combater o *overfitting* encontra-se num algoritmo introduzido por BREIMAN (2001) como uma extensão do conceito das Árvores de Decisão, chamado de *Random Forest*. Como um refino dos modelos de árvores para classificação e regressão, o algoritmo consiste na criação aleatória de múltiplas árvores distintas baseadas em dados, que são reunidas posteriormente a fim de alcançar uma previsão mais precisa do resultado.

Segundo BREIMAN (2001), um número N de árvores distintas trabalham de forma mais eficiente em conjunto do que independentes, visto que erros de árvores individuais influenciam menos numa “floresta” onde a maioria estará correta. Em mode-

los de classificação, são realizados “votos” por cada árvore e então classificam o dado de acordo com a maioria dos votos. Em modelos de regressão, o *Random Forest* simplesmente tira a média entre os *outputs* das árvores. Uma ilustração do funcionamento do *Random Forest* pode ser visto na Figura 4.

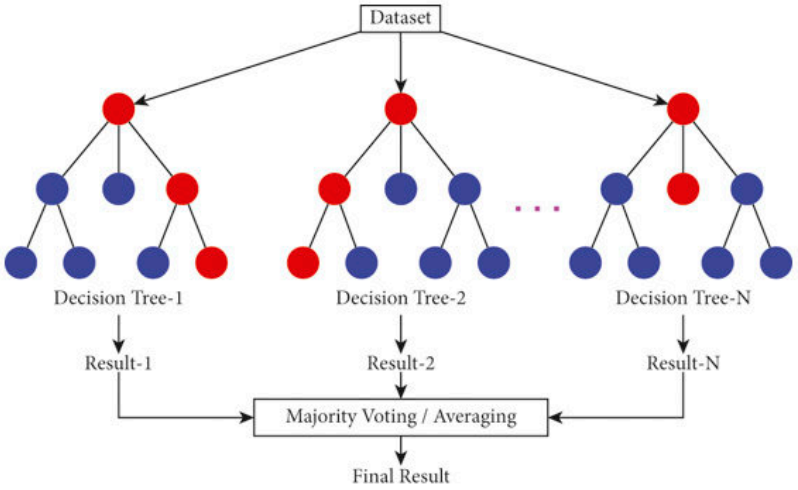


Figura 4 – Exemplo genérico de uma *Random Forest* para Classificação/Regressão. Fonte: (KHAN et al., 2021)

Dada a natureza do *Random Forest*, BREIMAN (2001) afirma que o algoritmo tende a evitar o problema de *overfitting* devido à Lei dos Grandes Números (LGN). No caso do *Random Forest*, a LGN diz que quanto mais se adicionam árvores à floresta, mais a média dos resultados (votos) obtidos tende ao resultado real.

2.4.3 Máquinas de Vetores de Suporte

Outra técnica de aprendizado de máquina cada vez mais utilizada é a aplicação de Máquinas de Vetores de Suporte (do inglês, *Support Vector Machines* ou SVM), que se baseiam na Teoria do Aprendizado Estatístico para traçar fronteiras entre dados num gráfico, a fim de separá-los em classes diferentes (LORENA; CARVALHO, 2007). Esta separação se baseia em elementos destes dados mais próximos de classes diferentes, nomeados vetores de suporte, que traçam retas conhecidas como hiperplanos separadores.

Existem duas abordagens na aplicação das SVMs: Linear e Não-Linear. Quando os dados possuem classes linearmente separáveis, a criação de fronteiras é simples e direta em SVMs Lineares; mas quando não há uma distribuição tão simples nos dados, as fronteiras se apresentam curvas e difíceis de serem representadas. as SVMs Não-Lineares surgem para transformar estes dados através de um mapeamento φ, transformando o espaço em um de maior dimensão onde é possivel a representação linear do hiperplano (LORENA; CARVALHO, 2007).

Na Figura 5, ilustra-se um gráfico contendo um hiperplano não linear e uma função de mapeamento ϕ, que o transforma em um gráfico com hiperplano linear. Os vetores de suporte da ilustração são aqueles que estão formando o hiperplano.

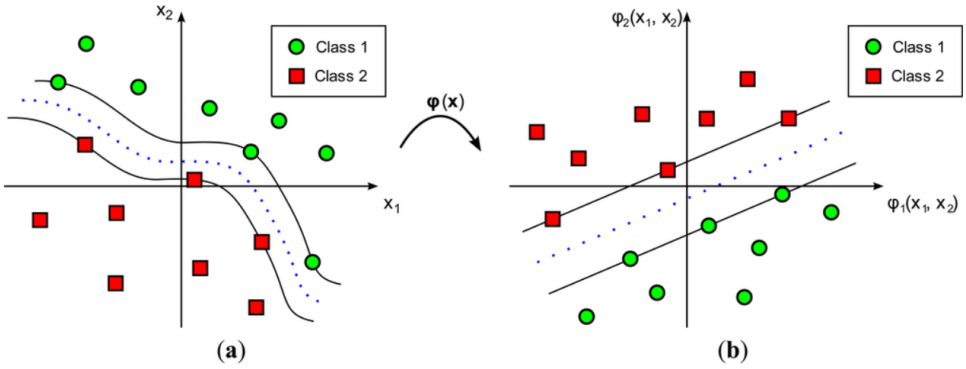


Figura 5 – Aplicação de uma função de mapeamento $\phi(x)$ em uma SVM Não-Linear.
Fonte: (RUIZ-GONZALEZ et al., 2014)

2.4.4 *K-Nearest Neighbors*

Outra técnica comumente utilizada para mineração de dados é o *K-Nearest Neighbors* (KNN), cuja classificação é baseada nas características dos k-elementos mais próximos ao elemento a ser analisado. Uma vez identificados, cada um deles irá "votar" na classificação do elemento-alvo, sendo então decidido pela maioria (BIJALWAN et al., 2014). É uma abordagem de decisão semelhante à do *Random Forest* para classificação, mas substituindo a exploração das Árvores de Decisão por votos de elementos próximos com sua classe já determinada. Um exemplo em forma de gráfico encontra-se na Figura 6.

Figura 6 – Exemplo da classificação de um elemento num algoritmo KNN de $k = 3$. Neste exemplo, identifica-se o elemento como Classe B.
Fonte: IBM.

São distintos os parâmetros para a definição do valor de de k (ZHANG et al., 2017), e a performance dos modelos varia de acordo com o valor selecionado. Por exemplo, um algoritmo com $k = 3$ irá analisar as características dos três elementos mais próximos do elemento analisado, mas o mesmo algoritmo com um $k = 5$ pode

tornar uma classe minoritária como majoritária, revelando circunstâncias cuja classificação do elemento mostre-se diferente. Apesar de simples e relativamente antigo, o KNN ainda se mostra um algoritmo acurado, sendo utilizado até os dias de hoje, inclusive no contexto da análise da evasão em IES (FILHO; VINUTO; LEAL, 2020).

3 Trabalhos Relacionados

Neste capítulo, encontram-se os trabalhos de outros estudiosos que se tornaram o fundamento para a elaboração deste, com uma breve descrição de seu conteúdo e relevância.

A fim de fundamentar o uso do AA no sistema, é usado de inspiração o estudo de Mesquita et al. (2021), uma revisão sistemática da literatura do AA com objetivo de facilitar a identificação do perfil discente e de características relevantes para seu sucesso ou insucesso acadêmico. Os resultados obtidos pelos autores identificam uma grande quantidade de soluções tecnológicas em formato de modelos matemáticos ou de aprendizado de máquina para o processo de identificação, mas quase não foram encontrados trabalhos voltados à criação de um ambiente de auxílio à compreensão dos perfis estudantis.

A ferramenta criada por Zapparolli et al. (2017) aplica técnicas de BI e LA num ambiente virtual de aprendizagem, utilizando os dados disponibilizados pelo ambiente para realizar apoio à gestão de uma IES. O trabalho fundamenta a aplicação das técnicas na implementação da ferramenta com intenção de auxiliar professores e gestores a combater a evasão através do acompanhamento de salas virtuais, e menciona a aplicação do método em diferentes ambientes e contextos. Este trabalho então se propõe a auxiliar o corpo docente e a gestão da IES através da identificação de perfis estudantis com altos riscos de evasão.

Visando a aplicação do BI, LA e AA num ambiente online para compreensão de fatores que levam à evasão, considera-se a adoção do SABIA, apresentado e descrito por Marques et al. (2023), como ambiente para desenvolvimento da proposta. Capaz de fornecer indicadores de desempenho através de *cards*, gráficos, tabelas e mapas, o sistema alinha-se às metas deste trabalho, mas não disponibiliza um painel capaz de realizar a análise de riscos de evasão baseado em características presentes nos dados disponibilizados pelo sistema. Logo, este trabalho busca incrementar uma nova página no SABIA com as funcionalidades propostas.

No contexto de algoritmos de aprendizado de máquina, Brito, Mello e Alves (2020) evidenciam a eficácia da utilização do algoritmo *Random Forest* no contexto educacional através de referências a trabalhos anteriores, que utilizaram o algoritmo com sucesso, e uma aplicação prática utilizando dados acadêmicos e demográficos para identificação de características relevantes para a análise da evasão, trazendo *insights* notáveis nos resultados obtidos. Este trabalho busca expandir o escopo da análise de características que levam à evasão, implementando diferentes algoritmos

em um único sistema para comparação pelo usuário.

Os trabalhos de Gonçalves, Silva e Cortes (2018) e Filho, Vinuto e Leal (2020) tratam casos de diferentes IES como objetos de estudo para análise de evasão utilizando múltiplas técnicas de mineração de dados, dentre elas: *Naive Bayes*, Árvores de Decisão, J48, KNN e SVM. Ao fim dos trabalhos, são utilizadas diferentes técnicas de análise de performance e precisão das técnicas selecionadas, evidenciando vantagens e desvantagens de cada abordagem. Ao aplicar diferentes algoritmos no sistema, este trabalho propõe que o usuário possa visualizar e configurar interativamente os modelos disponibilizados de acordo com suas necessidades de estudo, ao invés de analisarem apenas modelos pré-definidos.

Resumidamente, este trabalho se inspira em ideias dos trabalhos citados, oferecendo uma experiência robusta ao entregar um sistema provido dos conceitos de AA, LA e BI para oferecer modelos detalhados, apresentando previsões sobre a situação final e a jornada de formação de estudantes com características específicas. O trabalho proporciona uma experiência distinta ao oferecer aos usuários a capacidade de montarem e ajustarem seus próprios modelos de acordo com as especificidades dos dados analisados, além de permitir uma simulação de perfis específicos de estudantes com finalidade de estimar sua chance de sucesso na IES.

4 Proposta

Este capítulo apresenta a proposta para o desenvolvimento do sistema, descrevendo seu funcionamento, da ingestão dos dados até a apresentação dos resultados na tela para o usuário. Para isto, foi elaborado um fluxograma (Figura 7) ilustrando o processo utilizado no sistema, onde pode-se traçar paralelos com os processos de BI e AA citados anteriormente.

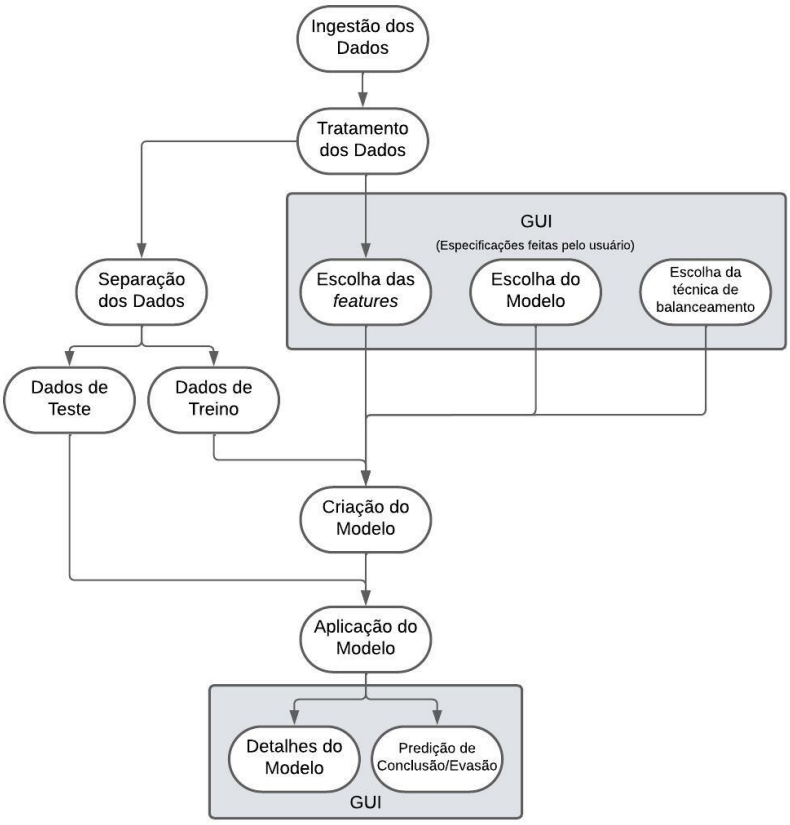


Figura 7 – Fluxograma com as etapas da página até a exibição dos resultados.

4.1 Ingestão e Tratamento dos Dados

Inicialmente, são realizadas as etapas de ingestão e tratamento dos dados através de *Data Marts*, bases de dados menores e refinadas para casos de uso mais específicos, de acordo com os requerimentos do usuário, quando comparados a um *Data Warehouse*, que é um repositório centralizado de dados (GHEZZI, 2001). Os dados utilizados no trabalho foram disponibilizados pela UFRPE, oriundos de seu próprio *Warehouse*, em junção a microdados disponibilizados pelo Instituto Nacional de Estudos

e Pesquisas Educacionais Anísio Teixeira (INEP)[1] voltados à avaliação institucional. No SABIA, cada página e seus respectivos painéis possuem seu próprio *Data Mart* orientado aos objetivos do painel.

A criação destes *Data Marts* se baseou nos princípios do *pipeline de dados ETL*, que significa Extrair, Transformar e Carregar (do inglês, *Extract, Transform and Load*). As principais etapas para a elaboração dos *Data Marts* são a ingestão, limpeza, transformação, carregamento e compartilhamento dos dados (MARQUES et al., 2023 apud RAJ et al., 2020), seguindo o mesmo caminho das etapas intermediárias do processo de AA. O produto final deste processo são arquivos *parquet*, que são disponibilizados para as páginas.

Dado que a situação do discente só é atualizada no final de cada período letivo e os dados do INEP levam tempo para serem recolhidos e disponibilizados, não há necessidade de atualizar os *Data Marts* em tempo real, sendo o processo de ETL realizado apenas a cada semestre (MARQUES et al., 2023).

Dentre as sub-etapas realizadas no processo de tratamento dos dados, cita-se a remoção e filtragem de dados duplicados/inconsistentes da base de dados, a modificação de campos (incluindo *merges* e criação de colunas novas no *parquet*) (MARQUES et al., 2023) e a limitação do escopo para os anos de 2010 a 2023.

4.2 Criação do Modelo

Após a inserção dos dados na página, eles são divididos em dois conjuntos de diferentes tamanhos, rotulados para treino e teste. A primeira e maior distribuição de dados, a de treino, será responsável pelo aprendizado do classificador, etapa na qual o modelo estará devidamente pronto para uso. Enquanto isso, a distribuição separada para teste será utilizada posteriormente para obter os resultados finais do classificador.

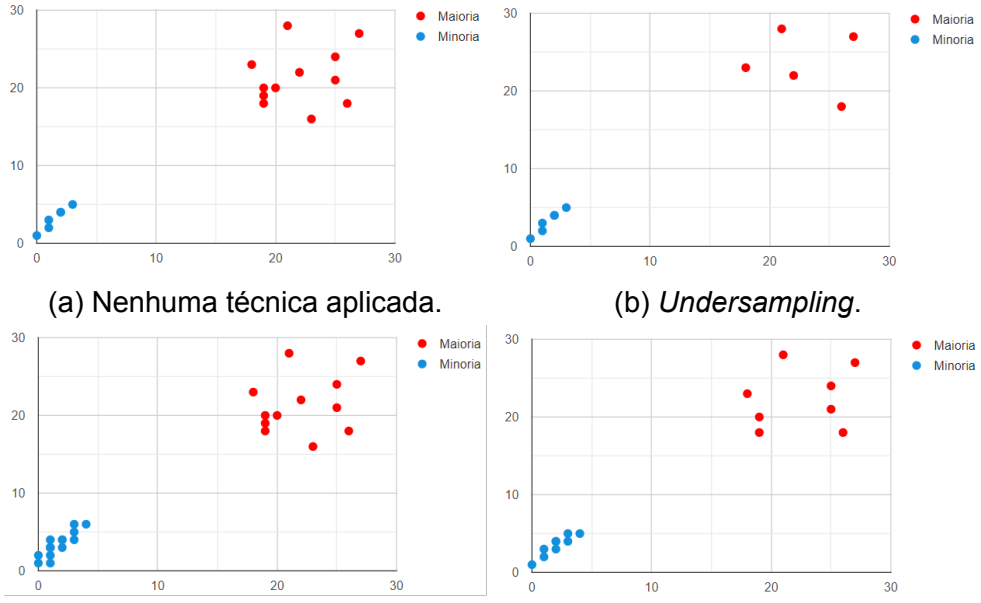
Para que o modelo seja criado, é necessário que o usuário especifique três atributos: os primeiros atributos são as *features* que estarão inclusas no modelo, baseadas em dados presentes nas colunas dos *parquets* disponibilizados para a página de previsão. O segundo atributo se trata de qual algoritmo de *machine learning* será utilizado pelo modelo. Por fim, o usuário seleciona uma técnica de balanceamento de dados.

A página de previsão utiliza diferentes classificadores de dados para realizar a previsão da situação acadêmica. Contudo, classificadores funcionam de forma mais eficiente quando alimentados por uma distribuição equilibrada de amostras de dados, visto que dados onde não há uma proporção justa de classes distintas, também cha-

[1] https://www.gov.br/inep/pt-br

mados de dados desbalanceados, levam os classificadores a interpretarem erroneamente os exemplos dados pela classe minoritária (BARBOSA et al., 2019 apud HULSE; KHOSHGOFTAAR; NAPOLITANO, 2007). Em outras palavras: ao utilizar classificadores com dados desbalanceados, o desempenho dos modelos tende a cair. Para evitar a queda de desempenho dos modelos no sistema, que lida com diferenças na proporção entre formados e evadidos, é necessária a implementação de técnicas de balanceamento de dados.

Dentre as estratégias de balanceamento existentes, duas estratégias tradicionais são o *oversampling* e o *undersampling*. Os algoritmos de *oversampling* lidam com dados desbalanceados aumentando o número de elementos na classe minoritária, através da criação de dados sintéticos baseados nos elementos existentes, enquanto os de *undersampling* cortam elementos da classe majoritária a fim de aproximar a quantidade de elementos entre as classes (BARBOSA et al., 2019). Vale frisar que estas estratégias são aplicadas exclusivamente nos dados de treino, visto que os de teste serão utilizados no sistema para análise real do problema. Adicionalmente, as duas estratégias podem ser usadas simultaneamente a fim de realizar uma abordagem híbrida ao problema de dados desbalanceados. Ilustrações da aplicação de cada técnica de balanceamento são mostradas na Figura 8.



(a) Nenhuma técnica aplicada. (b) *Undersampling*.

(c) *Oversampling*. (d) Abordagem híbrida.

Figura 8 – Ilustração de um *dataset* genérico e transformações baseadas em técnicas de balanceamento.

Com os atributos selecionados, o modelo é devidamente montado e treinado, podendo-se enfim inserir nele os dados separados para teste, concluindo esta etapa do processo.

4.3 Aplicação do Modelo

Após a inserção dos dados de teste no modelo, é possivel realizar a análise das informações obtidas por ele, como métricas de desempenho do modelo, a comparação entre dados reais e dados previstos pelo modelo e um gráfico de previsão da situação final de um estudante baseado nas características selecionadas pelo usuário.

Observa-se que, a fim de evitar confusões por parte do usuário quanto aos resultados apresentados pelo sistema, é importante a aplicação de uma Inteligência Artificial Explicável, ou *Explainable Artificial Intelligence* (XAI). O propósito de um sistema XAI é “tornar o seu comportamento mais inteligível para humanos ao providenciar explicações sobre o mesmo” (GUNNING et al., 2019, tradução nossa). Com finalidade de aplicar XAI ao sistema, busca-se a implementação de um gráfico com métricas de importância das características presentes no modelo (*feature importance*, em inglês) capaz de contextualizar os resultados obtidos pelo algoritmo selecionado.

Ao final do processo de obtenção de informações através dos modelos, espera-se que o sistema forneça contexto necessário para que os objetivos de auxílio à tomada de decisão sejam alcançados.

5 Método e Ferramentas

Provido do contexto e conhecimento necessários para o desenvolvimento do objetivo e da proposta, este capítulo descreve o processo de execução do método e pesquisa do trabalho.

Para todas as etapas do processo, algumas ferramentas em comum foram utilizadas: o sistema predominantemente utiliza a linguagem de programação Python[1] em seu desenvolvimento, e o tratamento de dados foi efetuado com o Pandas[2], uma biblioteca popular para *data science* devido à sua simplicidade e eficiência, proporcionando flexibilidade em relação a manipulação e limpeza de dados.

Em relação aos dados presentes no *parquet*, vale mencionar que eles são categóricos baseados em colunas presentes nos *dataframes*. Colunas relevantes a serem mencionadas, além das selecionadas para as *features* dos modelos, contêm a situação de vínculo atual e final dos estudantes.

5.1 Modelos e Balanceamento

Enquanto outras páginas do sistema compartilham das bibliotecas citadas nesta seção, a página de previsão apresenta necessidades específicas que precisam ser atendidas por outras bibliotecas disponibilizadas pelo Python. Os classificadores serão montados baseados em modelos obtidos pelo Scikit-Learn [3], uma biblioteca que oferece modelos diversos para realizar, dentre outras operações de dados, classificação, regressão e *clustering* (PEDREGOSA et al., 2011).

O novo painel desenvolvido para o sistema foi elaborado de forma facilmente escalável para modelos baseados em aprendizagem supervisionada da biblioteca Scikit-Learn (PEDREGOSA et al., 2011). Nele, o usuário pode livremente selecionar método de balanceamento, classificador e *features*, baseadas em colunas presentes nos *parquets* processados antes do processo de criação do modelo.

Para este trabalho, uma quantidade limitada dos modelos do Scikit-Learn foi selecionada, baseando-se na pesquisa realizada ao longo da trajetória do estudo. Nomeadamente, os modelos selecionados foram o *Random Forest*, o SVM, o KNN e o *Naive Bayes*.

O modelo escolhido para o *Random Forest* foi o *Random Forest Classifier* do

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

sklearn.ensemble, devido aos resultados positivos alcançados pelo algoritmo nos trabalhos encontrados ao longo do estudo e sua aplicação da abordagem lógica. Para o SVM, foi escolhido o SVC, do *sklearn.svm*, devido à simplicidade e ao parâmetro *kernel*, capaz de determinar se a abordagem será linear ou não-linear (PEDREGOSA et al., 2011). Para o KNN, utiliza-se o *KNeighbors Classifier* do *sklearn.neighbors* com o valor padrão $k = 5$. A fim de aplicar o raciocínio probabilístico, foi selecionado o GaussianNB do *sklearn.naive-bayes*, que utiliza distribuição normal (ou gaussiana) para estimativas de elementos e *features*.

Antes de aplicar no modelo os dados de treino e teste presentes no *parquet* da página, uma última filtragem é realizada nos dados, onde apenas serão considerados para cálculo os estudantes que já encerraram o vínculo com a IES: ou seja, as únicas classes de situação acadêmica final presentes serão $FORMADO$ e $EVADIDO$. Adicionalmente, é necessário codificar numericamente as colunas das *features* presentes no modelo para leitura do algoritmo selecionado. Este *encoding* é realizado com apoio da classe $LabelEncoder$ do *sklearn.preprocessing*, com seus valores originais (geralmente *strings*) armazenados em dicionários para consulta posterior.

Para aplicação das estratégias de balanceamento dos dados, foi utilizada a biblioteca Imbalanced-learn[4], que fornece algoritmos para lidar com classes desbalanceadas em casos de classificação. Para o *undersampling*, o Imbalanced-learn proporciona *Edited Nearest Neighbors* (ENN), reduzindo a classe majoritária a elementos que generalizam outros elementos próximos a eles. Para o *oversampling*, temos o *Synthetic Minority Over-sampling Technique* (SMOTE), criando novos elementos da classe minoritária com características em comum de outros elementos já existentes na classe. Adicionalmente, temos um algoritmo que combina as duas técnicas através de SMOTE e ENN, chamado de SMOTEENN. Ele realiza o SMOTE como primeira etapa e, após a sintetização de dados na classe minoritária, realiza o ENN para remover elementos em ambas as classes, a fim de reduzir os elementos mal classificados.

5.2 Apresentação dos Dados

Para as etapas de interface e visualização do BI e do AA, o sistema utiliza a biblioteca Streamlit[5], responsável pela criação de uma aplicação web com suporte a ferramentas de aprendizado de máquina e ciência de dados. Nas páginas montadas, os dados (devidamente processados) são inseridos em gráficos, tabelas e infográficos criados por funções de bibliotecas como o Matplotlib[6] para criação de *plots* estáticos

[4] https://imbalanced-learn.org/stable/
[5] https://streamlit.io/
[6] https://matplotlib.org/

simples e o Plotly[7] para visualizações dinâmicas e interativas (MARQUES et al., 2023).

A interface do sistema possui uma seção dedicada para a análise do modelo criado pelo usuário. Esta seção possui um conjunto de métricas importantes para a compreensão dos resultados, sendo elas: *accuracy*, *precision*, *support*, *recall* e *f1-score*.

A *accuracy*, ou acurácia, indica o quão próxima a análise está do objeto a ser observado; em outras palavras, o quanto os dados previstos condizem com os dados reais. A precisão (*precision*) se difere da acurácia sendo uma métrica que demonstra, intuitivamente, a habilidade do classificador de não rotular um elemento negativo como positivo. No caso do SABIA, os valores positivos são os formados, enquanto os negativos são os evadidos. O *support* se trata simplesmente do número de ocorrências daquela classe nos dados disponibilizados. Já a recordação (*recall*) se trata, especificamente, da capacidade do classificador de encontrar os elementos positivos, sem considerar os negativos. O *f-beta score*, por fim, é uma métrica de média harmônica ponderada entre precisão e *recall*, onde a função encontra seu melhor resultado quando *beta* for igual a 1 (logo, calcula-se o $F1$) e seu pior resultado quando *beta* for 0 (PEDREGOSA et al., 2011).

As Equações 5.1, 5.2, 5.3 e 5.4 mostram como os valores dessas métricas são alcançados. Nelas, TP se refere a quantidade dos valores positivos reais, TN são valores negativos reais, FP são os falsos positivos (ou seja, valores previstos imprecisos) e FN os falsos negativos. O cálculo de *f1-score* já está adaptado para os valores de *beta* igual a 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.4}$$

Os gráficos e tabelas criados para a página são montados dentro de *containers* do Streamlit para visualização, como por exemplo o gráfico de feature importance na Figura 9 implementado com finalidade de aplicar XAI ao sistema. Para o desenvolvimento deste gráfico, foi utilizado o algoritmo de *feature importance* por permutação presente no *sklearn.inspection* (PEDREGOSA et al., 2011), dada sua disponibilidade

[7] https://plotly.com/

a todos os modelos escolhidos para o trabalho. O gráfico consolida informações como o impacto daquela característica no modelo atual e a sua variação, e é crucial para refinar o modelo através da inclusão de *features* novas e a poda daquelas redundantes ou prejudiciais ao mesmo.

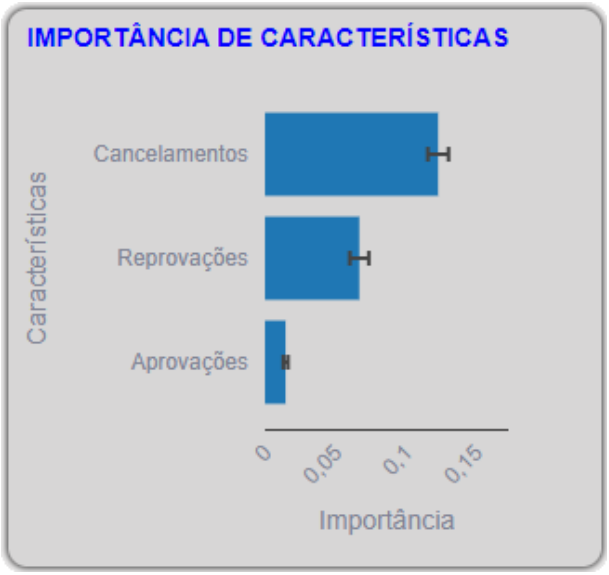


Figura 9 – Exemplo de gráfico de *feature importance* para um modelo com um pequeno conjunto de características.

6 Resultados Alcançados

Neste capítulo será apresentada a primeira versão do sistema, exemplificado o processo de criação e ajustes de um modelo utilizando dados fornecidos para o SABIA. Ao longo do processo haverão análises do sistema e dos resultados obtidos pelo modelo, com algumas conclusões e observações baseadas nestas análises.

6.1 Interface e Lógica do Sistema

Ao acessar o sistema, o usuário será recebido pela *frontpage* do SABIA (Figura 10), onde terá acesso ao painel de previsão através da barra lateral, caso tenha as permissões necessárias.

Figura 10 – *Frontpage* do SABIA.

O painel criado neste trabalho se divide em três seções: um menu para criação do modelo, uma seção para informações detalhadas do modelo criado e uma seção para a previsão da situação final de perfis de estudantes, baseada no modelo treinado.

A primeira seção consiste em um conjunto de filtros e seletores para *features*, classificador e técnica de balanceamento, onde serão disponibilizadas as opções para criação do modelo (Figura 11).

Os classificadores oferecidos para esta versão da página são *Naive Bayes*, *Random Forest*, *K-Nearest Neighbors* e *Support Vector Machine*; e os algoritmos de balanceamento presentes são o *oversampling* com SMOTE, o *undersampling* com ENN e a abordagem híbrida do SMOTEENN. As *features* disponibilizadas para análise são dados referentes a curso, duração do vínculo, forma de ingresso, turno, período de

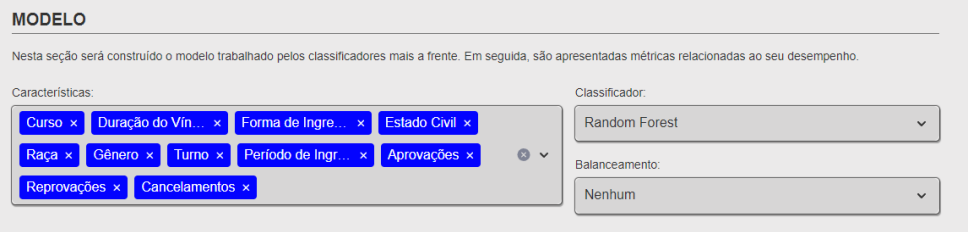


Figura 11 – Seção reservada para montagem do modelo no painel do SABIA.

ingresso, quantidade de aprovações, reprovações e cancelamentos; estado civil, raça e gênero.

Quanto ao uso de *features* de cunho social, enquanto pode-se afirmar sua inclusão no sistema, torna-se importante iterar a responsabilidade de seu uso por quem estiver interpretando as informações obtidas. Enquanto é possível que as informações sejam utilizadas para realizar a exclusão de discentes com características desfavoráveis para a IES, reafirma-se que o sistema foi desenvolvido para fins inclusivos, através da adoção de medidas para mitigar o fenômeno da evasão estudantil.

Uma vez selecionados, as escolhas presentes nos seletores são armazenadas e passadas como argumentos no código para montagem do modelo e, posteriormente, apresentação das informações.

A seção de informações do modelo possui conteúdo crucial para que os usuários (professores e gestores) analisem o quão eficiente o modelo é em sua previsão, e como ele pode ser melhorado. Como exemplo, a Figura 12 mostra a visão de um modelo através de tabelas com detalhes do modelo e matrizes de confusão.

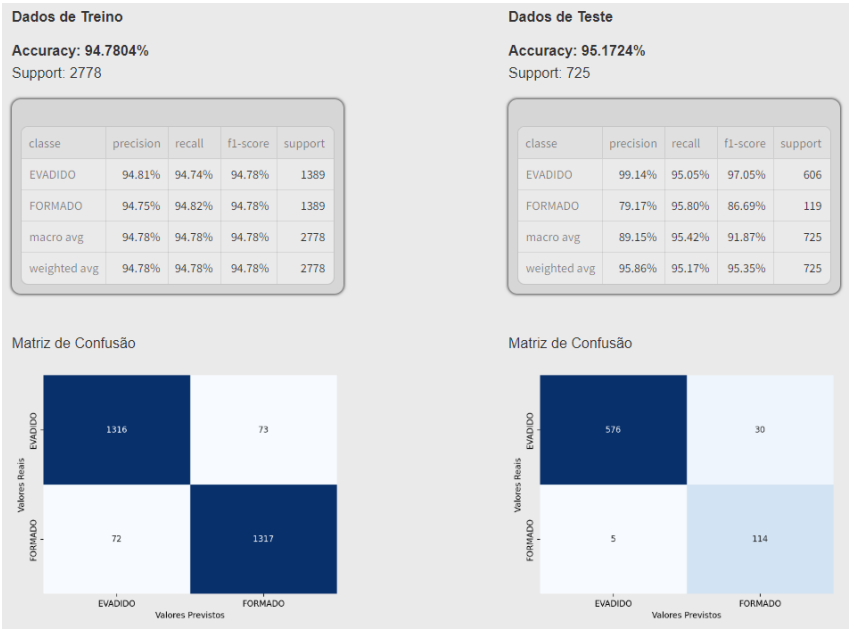


Figura 12 – Seção do painel do SABIA contendo informações sobre um modelo criado pelo usuário.

Esta seção também contém o gráfico de *feature importance* apresentado na Figura 9. No contexto de análise de características, Pedregosa et al. (2011) trazem uma consideração importante para certos cenários de *feature importance*: os algoritmos não refletem o valor de previsão intrínseco às características, e sim o quão importantes elas são para modelos em particular; ou seja, *features* tratadas como de baixa importância em modelos ineficientes podem se mostrar importantes em modelos com boas métricas. Por isso, é sempre importante avaliar informações até mesmo em modelos de baixa eficiência, visto que neles pode-se encontrar informação importante para a elaboração de um bom modelo.

Por fim, a seção final da página é reservada para as previsões do modelo treinado (Figura 13) utilizando os dados de teste. Nela, o usuário pode selecionar cada uma das características que ele inseriu no modelo para simular um estudante com uma situação específica e observar sua chance de sucesso quando comparada à de falha, e chegar a conclusões baseadas nas informações obtidas. Por exemplo, ao observar as variações de comportamento do classificador enquanto modifica a duração do vínculo, um período por vez, o usuário simula a trajetória daquele perfil de estudante na instituição.

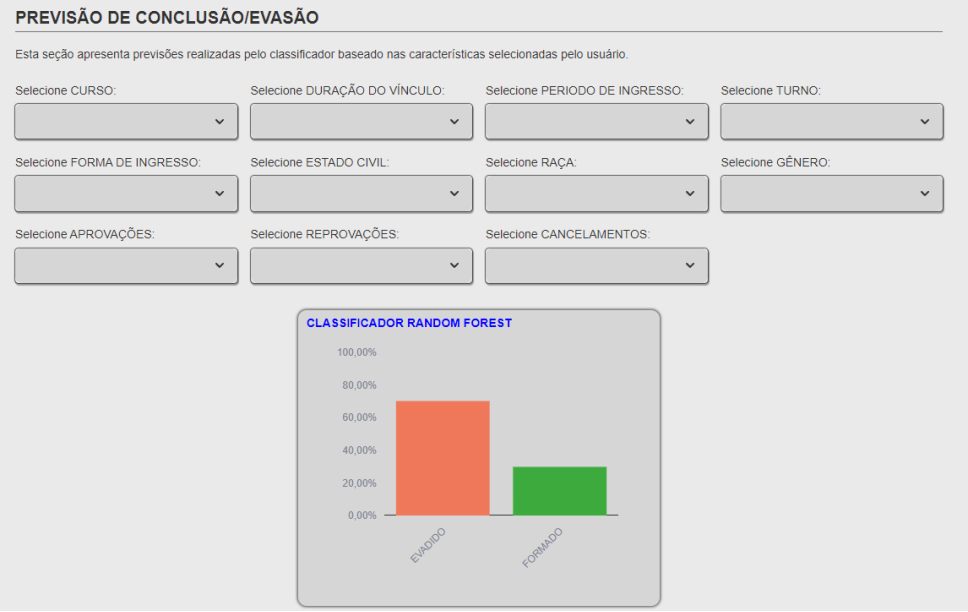


Figura 13 – Seção do painel do SABIA dedicada à previsão da situação final do estudante.

6.2 Modelos e Previsões

Nesta seção, é demonstrado o processo de criação e refino de um modelo; além de possíveis análises que o usuário pode conseguir a partir dos resultados obtidos.

6.2.1 Criação do Modelo

O primeiro parâmetro a ser definido é o classificador. Para isso, foram selecionadas algumas *features* e, temporariamente, a técnica de *undersampling* através de ENN. Este exemplo utiliza como *features* apenas aquelas voltadas para a situação acadêmica: o nome do curso, a duração do vínculo (em períodos), a forma de ingresso na IES, o turno dos horários do curso e o período de ingresso (primeiro ou segundo semestre).

Com essas características selecionadas, é possivel fazer um comparativo entre métricas presentes nos classificadores selecionados. Na Tabela 1 encontra-se os resultados obtidos nos dados de teste para estudantes evadidos e formados e as médias aritmética e ponderada, respectivamente.

Preliminarmente, pode-se ver que os classificadores demonstram eficiência na previsão de evasão de estudantes, com o KNN e o *Random Forest* se sobressaindo em comparação aos outros dois modelos. Entretanto, todos apresentam quedas drásticas na precisão e *f1-score* ao tentar prever a situação de estudantes formados, demonstrando dificuldades em identificar características da classe minoritária. Este comportamento pode indicar problemas de *overfitting* com os dados de treino utilizados.

Para a próxima etapa, o KNN foi escolhido como classificador por apresentar considerável vantagem sobre o *Random Forest* em acurácia e *f1-score*. Numa tentativa de aprimorar o modelo, a técnica de balanceamento de dados foi ajustada para confererir o impacto nas métricas, registrado na Tabela 2.

As principais métricas a serem analisadas nessa etapa foram acurácia e precisão. Percebe-se métricas estáveis e altas na classificação de evadidos para todos os algoritmos de balanceamento. Entretanto, há uma melhora considerável na acurácia (0.07%) e precisão de formados (3.08%) do modelo quando se utiliza *oversampling* através de SMOTE nos dados, ao invés de *undersampling* com ENN. A vantagem do ENN está na precisão de evadidos (3.17%) e a precisão da média ponderada (2.02%), visto que evadidos são a classe majoritária. Num geral, o SMOTEEN demonstrou métricas em sua maioria inferiores às demais técnicas que, por sua vez, demonstram precisões médias praticamente equivalentes.

Visto que o trabalho pretende identificar fatores que levam à evasão, o ENN será o algoritmo selecionado para o modelo final deste estudo, apesar do SMOTE ter apresentado métricas melhores em certos aspectos.

	Naive Bayes			Random Forest			KNN			SVM			
	Precision	***Recall***	***F1-Score***	***Precision***	***Recall***	***F1-Score***	***Precision***	***Recall***	***F1-Score***	***Precision***	***Recall***	***F1-Score***	***Support***
Evadidos	95.34%	81.25%	87.73%	**97.34%**	82.46%	89.28%	96.15%	**84.90%**	**90.18%**	96.69%	81.13%	88.23%	8618
Formados	67.41%	90.72%	77.35%	69.78%	**94.74%**	80.36%	**72.28%**	92.05%	**80.97%**	67.94%	93.51%	78.70%	3685
Média	81.38%	85.98%	82.54%	83.56%	**88.60%**	84.82%	**84.21%**	88.48%	**85.58%**	82.32%	87.32%	83.47%	12303
Média Ponderada	86.98%	84.09%	84.62%	**89.09%**	86.13%	86.61%	89.00%	**87.04%**	**87.42%**	88.08%	84.84%	85.38%	12303
Acurácia	84.08%			86.13%			**87.04**%			84.84%			

Tabela 1 – Comparação entre métricas de cada classificador.

	ENN			SMOTE			SMOTEENN			
	Precision	***Recall***	***F1-Score***	***Precision***	***Recall***	***F1-Score***	***Precision***	***Recall***	***F1-Score***	***Support***
Evadidos	**96.15%**	84.90%	90.18%	93.28%	**88.16%**	**90.55%**	93.28%	86.66%	89.85%	8618
Formados	72.28%	**92.05%**	**80.97%**	**75.36%**	84.67%	79.74%	73.24%	85.40%	78.85%	3685
Média	84.21%	**88.48%**	**85.58%**	**84.22%**	86.42%	85.15%	83.26%	86.03%	84.35%	12303
Média Ponderada	**89.00%**	87.04%	**87.42%**	87.77%	**87.12%**	87.32%	87.28%	86.28%	86.55%	12303
Acurácia	87.04%			**87.11%**			86.27%			

Tabela 2 – Comparação entre métricas de cada algoritmo de balanceamento para o KNN.

6.2.2 Análise do Modelo

Em seguida, foi realizada uma análise de *feature importance* por permutação do modelo, que pode ser vista na Figura 14. Nela, evidencia-se a duração de vínculo como principal característica do modelo, e o curso como a segunda mais importante.

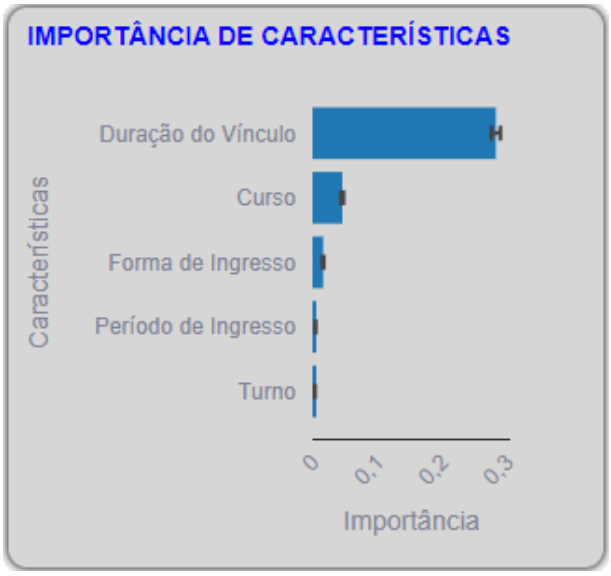


Figura 14 – Gráfico de *feature importance* do modelo definido.

Dada a importância dessas características, torna-se essencial dar atenção à forma em que o classificador prevê as chances de evasão e conclusão com o passar dos períodos em diferentes cursos. A Tabela 3 consolida previsões obtidas para cada período em toda a IES (nenhum curso selecionado) e previsões de três cursos distintos da área de computação.

Curso/Duração	1∼4	5	6	7	8	9	10	11	12∼15
-	100%	100%	100%	0%	0%	0%	0%	0%	0%
Curso I	100%	60%	100%	100%	40%	100%	0%	0%	0%
Curso II	100%	100%	100%	100%	100%	100%	100%	0%	0%
Curso III	100%	100%	100%	100%	60%	0%	0%	100%	0%

Tabela 3 – Estimativas de probabilidade de evasão para estudantes baseadas em curso e duração de vínculo.

Com a tabela, pode-se tirar conclusões sobre o impacto da duração de vínculo na permanência do estudante na instituição. A formação é altamente improvável nos primeiros períodos, visto que ele ainda está nas etapas iniciais do curso. Certos perfis são exceções para estes casos, tais como alunos transferidos ou que passaram por processos de equivalência de disciplinas.

Ao passar dos períodos, a probabilidade da simulação classificar a situação do estudante como de sucesso eventualmente supera a probabilidade de falha e torna-se a situação esperada em períodos tardios. Neste exemplo, contudo, pode-se notar um

atraso nos cursos de computação como um todo para alcançar a mesma tendência da instituição num geral de obter sucesso a partir do sétimo período.

Como este gráfico não apresenta números reais e simplesmente probabilidades, não é possível mensurar a proporção deste fenômeno. Torna-se então importante a verificação de outras ferramentas do sistema como as de acompanhamento do progresso acadêmico ou análise de sobrevivência para uma compreensão mais fundamentada. Contudo, é possível analisar perfis específicos que levam o modelo a estimar essas chances ao especificar mais características no classificador.

Ao analisar a situação de um estudante baseado em curso, duração de vínculo e período de ingresso, como visto na Tabela 4, percebe-se uma leve mudança de comportamento do estimador para com os estudantes do Curso I: aqueles que ingressaram nos períodos referentes ao primeiro do ano letivo (20XX.1) demonstram tendências à conclusão acadêmica mais cedo, apesar de estarem dispersos em períodos distintos. Enquanto isso, ingressantes no segundo semestre letivo (20XX.2) apresentam na previsão uma tendência menos favorável à conclusão acadêmica antes do nono período.

Curso I									
Período Ingresso/Letivo	**1∼4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12∼15**
1º Semestre	100%	60%	100%	100%	40%	100%	0%	0%	0%
2º Semestre	100%	100%	100%	100%	100%	0%	0%	0%	0%

Tabela 4 – Estimativas de evasão para estudantes de um curso específico, baseadas em duração de vínculo e período de ingresso.

Observando outra *feature* do modelo, a forma de ingresso, também pode-se encontrar discrepâncias em estimativas entre as duas classes mais populosas na mesma: Vestibular e SISU. Num geral, ingressantes através do SISU demonstram uma chance de evasão consideravelmente maior que ingressantes através do vestibular tradicional, implicando num possível obstáculo para a permanência de alunos que entraram na IES através do sistema de seleção mais recente.

A ampla disponibilidade de combinações de modelos e *features* para a previsão da situação acadêmica torna esse apenas um dos vários exemplos possíveis para análise e a interatividade na criação do modelo permite que professores e gestores possam regular livremente o escopo de seu estudo, de casos mais gerais a casos específicos.

Vale relembrar que o sistema é apenas uma ferramenta, com suas imprecisões e defeitos, e suas informações não podem ser lidas como verdades absolutas. Usando como exemplo o modelo montado nesta seção, apesar de ter apresentado métricas melhores após ajustes, não foi capaz de solucionar o problema do *overfitting* para discentes na situação de formados, comprometendo sua precisão.

7 Conclusões Finais

O presente trabalho reconhece o problema da evasão estudantil em IES e os seus riscos, tanto para os discentes quanto para as instituições de ensino. Dado este contexto, o trabalho disponibiliza um sistema interativo de análise de características de discentes através de uma nova página do SABIA, que proporciona modelos de aprendizado supervisionado personalizáveis. Além disso, o sistema permite que se avalie o impacto de mudanças nas variáveis relacionadas aos discentes na sua situação acadêmica final.

Para auxiliar a criação do modelo, a página oferece gráficos e tabelas contendo métricas e informações sobre o mesmo. Os dados fornecidos podem orientar o usuário a refinar o modelo e realizar um estudo mais eficiente.

Através da identificação de fatores determinantes para a evasão, este sistema busca auxiliar a gestão e o corpo docente da IES na elaboração de dinâmicas de combate ao fenômeno, aumentando a chance de sucesso acadêmico dos estudantes.

Contudo, vale frisar que pelos resultados da ferramenta se tratarem de previsões e estimativas, eles funcionam melhor quando unidos a outros métodos de análise mais objetivos, alguns dos quais já se encontram disponíveis no SABIA, como acompanhamento de situação acadêmica e análise de sobrevivência. Adicionalmente, o sistema apresenta imperfeições, como o viés dos modelos voltados para as classes majoritárias e o *overfitting* de formados, que exigem um olhar crítico na interpretação de suas informações antes da tomada de decisão.

Como perspectiva futura, espera-se expandir o potencial de personalização dos modelos já existentes através da manipulação de variáveis específicas para cada algoritmo, como a quantidade de vizinhos do KNN; além da aplicação de outras técnicas como normalização e *scaling* de features, a fim de atender aos problemas mencionados anteriormente.

Ainda em perspectivas futuras, é de nosso interesse adicionar novos modelos e algoritmos ao sistema, ampliando seu potencial de análise, além de aprofundar a aplicação de XAI através de outras ferramentas, como a biblioteca SHAP [1] para facilitação da compreensão dos resultados. Por fim, vale conferir a viabilidade de uma abordagem sistemática da análise, comparando os resultados obtidos pelo modelo com uma lista de estudantes reais inserida pelo usuário e ordenando essa lista em níveis de criticidade.

[1] https://shap.readthedocs.io/en/latest/

Referências

ANDRADE, A.; FERREIRA, S. A. Aspetos morfológicos do tratamento de dados na gestão escolar. o potencial do analytics. *Revista Portuguesa de Investigação Educacional*, n. 16, p. 289–316, 2016. Citado na página 17.

BARBOSA, G. et al. Sequenciamento de algoritmos de amostragem para aumentar o desempenho de classificadores em conjuntos de dados desequilibrados. In: SBC. *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2019. p. 413–423. Citado na página 27.

BATISTA, R. de A.; BAGATINI, D. D.; FROZZA, R. Classificação automática de códigos ncm utilizando o algoritmo naïve bayes. *iSys-Brazilian Journal of Information Systems*, v. 11, n. 2, p. 4–29, 2018. Citado na página 18.

BIANCHI, I. S. et al. Business intelligence e dashboards na educação superior: uma revisão sistemática da literatura. *Encontro Internacional de Gestão, Desenvolvimento e Inovação (EIGEDIN)*, v. 6, n. 1, 2022. Citado na página 16.

BIJALWAN, V. et al. Knn based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, NADIA, v. 7, n. 1, p. 61–70, 2014. Citado na página 21.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001. Citado 2 vezes nas páginas 19 e 20.

BRITO, B. C. P. de; MELLO, R. F. L. de; ALVES, G. Identificação de atributos relevantes na evasão no ensino superior público brasileiro. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 1032–1041. Citado 2 vezes nas páginas 12 e 23.

CAMPOS, A. T.; FONSECA, R. C. B. da. Learning analytics e academic analytics nas instituições de ensino superior como paradigma para a gestão da educação corporativa. *Revista Novas Tecnologias na Educação*, v. 21, n. 2, p. 382–392, 2023. Citado na página 17.

CLOW, D. The learning analytics cycle: closing the loop effectively. In: *Proceedings of the 2nd international conference on learning analytics and knowledge*. [S.l.: s.n.], 2012. p. 134–138. Citado na página 17.

DAVIES, P. *Kendall's Advanced Theory of Statistics. Volume 1. Distribution Theory*. [S.l.]: Wiley Online Library, 1988. Citado na página 18.

FILHO, F. W. B. H.; VINUTO, T. S.; LEAL, B. C. Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 1132–1141. Citado 3 vezes nas páginas 12, 22 e 24.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, SciELO Brasil, v. 37, p. 641–659, 2007. Citado na página 11.

GHEZZI, C. Designing data marts for data warehouses. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, ACM New York, NY, USA, v. 10, n. 4, p. 452–483, 2001. Citado na página 25.

GONÇALVES, T. C.; SILVA, J. C. da; CORTES, O. A. C. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do instituto federal do maranhão. *Revista Brasileira de Computação Aplicada*, v. 10, n. 3, p. 11–20, 2018. Citado 2 vezes nas páginas 12 e 24.

GUNNING, D. et al. Xai—explainable artificial intelligence. *Science robotics*, American Association for the Advancement of Science, v. 4, n. 37, p. eaay7120, 2019. Citado na página 28.

HULSE, J. V.; KHOSHGOFTAAR, T. M.; NAPOLITANO, A. Experimental perspectives on learning from imbalanced data. In: *Proceedings of the 24th international conference on Machine learning*. [S.l.: s.n.], 2007. p. 935–942. Citado na página 27.

IBM. *What is the k-nearest neighbors algorithm?* n.d. <https://www.ibm.com/topics/knn>. Citado na página 21.

KHAN, M. Y. et al. Automated prediction of good dictionary examples (gdex): a comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques. *Complexity*, Hindawi Limited, v. 2021, p. 1–18, 2021. Citado na página 20.

LADEIRA, M.; VICARI, R. M.; COELHO, H. Redes bayesianas multiagentes. In: *CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, XIX*. [S.l.: s.n.], 1999. Citado na página 18.

LAMERS, J. M. d. S.; SANTOS, B. S. d.; TOASSI, R. F. C. Retenção e evasão no ensino superior público: estudo de caso em um curso noturno de odontologia. *Educação em Revista*, SciELO Brasil, v. 33, p. e154730, 2017. Citado na página 16.

LARSON, D.; CHANG, V. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, Elsevier, v. 36, n. 5, p. 700–710, 2016. Citado na página 16.

LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, v. 25, p. 14, 2012. Citado na página 11.

LORENA, A. C.; CARVALHO, A. C. D. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. Citado na página 20.

MARQUES, E. et al. Sabia: Uma plataforma para auxiliar a gestão baseada em evidências nas instituições de ensino superior. In: SBC. *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*. [S.l.], 2023. p. 71–80. Citado 6 vezes nas páginas 11, 12, 13, 23, 26 e 31.

MARQUES, R. L.; DUTRA, I. Redes bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. *Coppe Sistemas–Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil*, 2002. Citado 2 vezes nas páginas 17 e 18.

MEC. *Documento Orientador para a Superação da Evasão e Retenção na Rede Federal de Educação Federal de Educação Profissional, Científica e Tecnológica*. 2014. Citado 2 vezes nas páginas 15 e 16.

MESQUITA, J. L. de et al. Academic analytics como apoio ao sucesso na graduação: uma revisão sistemática da literatura academic analytics to support undergraduate success: a systematic review of the literature. *Brazilian Journal of Development*, v. 7, n. 10, p. 99882–99897, 2021. Citado na página 23.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003. Citado na página 19.

MONARD, M. C.; BARANAUSKAS, J. A. Indução de regras e árvores de decisão. *Sistemas Inteligentes-fundamentos e aplicações*, sn, v. 1, p. 115–139, 2003. Citado na página 19.

NUNES, R. C. Um olhar sobre a evasão de estudantes universitários durante os estudos remotos provocados pela pandemia do covid-19. *Research, Society and Development*, v. 10, n. 3, p. e1410313022–e1410313022, 2021. Citado na página 12.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 4 vezes nas páginas 29, 30, 31 e 35.

RAISINGHANI, M. S. *Business intelligence in the digital economy: opportunities, limitations and risks*. [S.l.]: Igi Global, 2004. Citado na página 16.

RAJ, A. et al. Modelling data pipelines. In: IEEE. *2020 46th Euromicro conference on software engineering and advanced applications (SEAA)*. [S.l.], 2020. p. 13–20. Citado na página 26.

RUIZ-GONZALEZ, R. et al. An svm-based classifier for estimating the state of various rotating components in agro-industrial machinery with a vibration signal acquired from a single point on the machine chassis. *Sensors*, MDPI, v. 14, n. 11, p. 20713–20735, 2014. Citado na página 21.

SANTINI, F. de O.; GUIMARÃES, J. C. F. de; SEVERO, E. A. Qualidade, comprometimento e confiança na retenção de alunos no ensino superior. *Revista Gestão Universitária na América Latina-GUAL*, Universidade Federal de Santa Catarina, v. 7, n. 1, p. 274–297, 2014. Citado na página 15.

SHAHINFAR, S. et al. Prediction of insemination outcomes in holstein dairy cattle using alternative machine learning algorithms. *Journal of dairy science*, Elsevier, v. 97, n. 2, p. 731–742, 2014. Citado na página 19.

TURBAN, E. et al. *Business intelligence: um enfoque gerencial para a inteligência do negócio*. [S.l.]: Bookman Editora, 2009. Citado na página 16.

UNIVERSIDADES, P. d. A. I. das; ESPECIAL, B. C.; BORDAS, M. C. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas: resumo do relatório apresentado a adifes, abruem e sesu/mec pela comissão especial. *Avaliação: revista da Rede de Avaliação Institucional da Educação Superior. Campinas, SP. Vol. 1, n. 2 (dez. 1996), p. 55-65*, 1996. Citado na página 15.

VIANA, F. S.; SANTANA, A. M.; RABÊLO, R. d. A. L. Avaliaçao de classificadores para prediçao de evasao no ensino superior utilizando janela semestral. In: SBC. *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2022. p. 908–919. Citado na página 11.

XAVIER, M.; MENESES, J. A literature review on the definitions of dropout in online higher education. In: . [S.l.: s.n.], 2020. Citado na página 15.

ZAPPAROLLI, L. et al. Aplicando técnicas de business intelligence e learning analytics em ambientes virtuais de aprendizagem. *Simpósio Brasileiro de Informática na Educação*, v. 28, n. 1, p. 536–545, 2017. Citado na página 23.

ZHANG, S. et al. Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM New York, NY, USA, v. 8, n. 3, p. 1–19, 2017. Citado na página 21.