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

EBONY MARQUES RODRIGUES

CLASSIFICAÇÃO DO RISCO DE EVASÃO NO ENSINO SUPERIOR A PARTIR DO RENDIMENTO ACADÊMICO E DO AGRUPAMENTO DE CURSOS

RECIFE – PE

2024

EBONY MARQUES RODRIGUES

CLASSIFICAÇÃO DO RISCO DE EVASÃO NO ENSINO SUPERIOR A PARTIR DO RENDIMENTO ACADÊMICO E DO AGRUPAMENTO DE CURSOS

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

ORIENTADOR: Gabriel Alves de Albuquerque Junior

COORIENTADORA: Roberta Macêdo Marques Gouveia

RECIFE – PE

2024

Dados Internacionais de Catalogação na Publicação
Sistema Integrado de Bibliotecas da UFRPE
Bibliotecário(a): Suely Manzi – CRB-4 809

R696c Rodrigues, Ebony Marques.
Classificação do risco de evasão no ensino superior a partir do rendimento acadêmico e do agrupamento de cursos / Ebony Marques Rodrigues. – Recife, 2024.
142 f.; il.

Orientador(a): Gabriel Alves de Albuquerque Junior.
Co-orientador(a): Roberta Macêdo Marques Gouveia.

Dissertação (Mestrado) – Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, BR-PE, 2024.

Inclui referências.

1. Ensino superior. 2. Evasão universitária. 3. Aprendizado do computador. 4. Rendimento escolar 5. Informática. I. Junior, Gabriel Alves de Albuquerque, orient. II. Gouveia, Roberta Macêdo Marques, coorient. III. Título

CDD 004

EBONY MARQUES RODRIGUES

CLASSIFICAÇÃO DO RISCO DE EVASÃO NO ENSINO SUPERIOR A PARTIR DO RENDIMENTO ACADÊMICO E DO AGRUPAMENTO DE CURSOS

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

Aprovada em 27 de junho de 2024.

BANCA EXAMINADORA

Gabriel Alves de Albuquerque Junior (Orientador)
Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática

Roberta Macêdo Marques Gouveia (Coorientadora)
Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática

Péricles Barbosa Cunha de Miranda
Universidade Federal Rural de Pernambuco
Departamento de Computação

Cristian Cechinel
Universidade Federal de Santa Catarina
Centro de Ciências, Tecnologias e Saúde

Com amor e gratidão, dedico este trabalho aos meus pais, Ricardo e Rozivania, por seu apoio incondicional e ensinamentos; ao meu irmão, Kelvin, pela parceria e incentivo constantes; e à minha querida avó, Josefa (*in memorian*), que deixou em meu coração um legado de amor, sabedoria e inspiração.

Agradecimentos

Durante esta fase de minha trajetória acadêmica, várias vezes li agradecimentos em dissertações e teses de colegas da academia (muitos sequer conheço) e imaginei as dificuldades pelas quais eles passaram, buscando me convencer de que os meus desafios não eram exclusivos. Hoje com alegria vejo a conclusão desta fase como resultado da união de soluções ótimas locais, como se a trajetória fosse um algoritmo guloso com a esperança de encontrar uma solução ótima global diante das adversidades da vida. Eu agradeço a Deus por me conceder a força de vontade e as condições necessárias para chegar até aqui.

Eu agradeço profundamente ao meu pai, Ricardo, à minha mãe, Rozivania, e ao meu irmão, Kelvin, por viverem e por me darem tanto suporte e compreensão. Vocês foram meu alicerce. Passamos por muito juntos, e as palavras que vocês merecem não cabem nesta página. Fico verdadeiramente feliz por me verem concluindo mais uma etapa.

Um agradecimento especial à minha avó, Josefa, que já não está entre nós. Sua sabedoria, amor e ensinamentos continuam vivos em mim, como uma fonte de inspiração constante. Dedico este trabalho à sua memória, com gratidão e saudade.

Eu sou imensamente grato ao meu orientador, Gabriel, à minha coorientadora, Roberta, e à professora Ceça, que são do DEINFO da UFRPE, e não falo apenas de suas contribuições acadêmicas. Vocês participaram de meu amadurecimento como ser humano. Eu não poderia ter sido mais feliz ao ser aceito para trabalhar com vocês.

Heloise está em minha vida há 10 anos, e já a considero minha irmã. Ela tantas vezes me ouviu falar sobre dificuldades e conquistas que já é família mesmo. Eu conheci Taciana durante a graduação, no BSI da UFRPE, e fico feliz que tenha se tornado uma amizade para a vida. Também agradeço à Clarissa, que compartilhou vários momentos desta jornada. Agradeço a vocês três por todo o apoio e pela ajuda na revisão do trabalho.

Também escrevo agradecimentos ao meu chefe na Emprel, Cloves, e aos demais amigos e colegas de firma, principalmente Diego, Letícia, Giordano, Tennyson e Joelson, que tantas vezes me ouviram falar sobre este trabalho. Satisfação!

Ao longo do percurso, conheci e interagi com pessoas que vieram a participar deste trabalho direta ou indiretamente, muitas que nem foram citadas, mas que fazem parte. Eu sei que estas palavras contam histórias diferentes para cada um de vocês, mas a essência de gratidão é comum. Mais uma vez, a todos vocês, os meus sinceros agradecimentos.

Se a educação sozinha não transforma a sociedade, sem ela tampouco a sociedade muda.

(Paulo Freire)

Resumo

A evasão estudantil no ensino superior ocorre com o abandono do curso de graduação pelo estudante. Trata-se de um fenômeno multifacetado que afeta não somente o indivíduo e a instituição de ensino, mas toda a sociedade. A previsão da evasão pode auxiliar o desenvolvimento de estratégias de combate ao fenômeno e de estímulo à retenção estudantil e ao sucesso acadêmico nas instituições de ensino. Diante disso, este trabalho possui o objetivo de prever a evasão em diferentes momentos do curso de graduação com o uso de técnicas do aprendizado de máquina não supervisionado e supervisionado. O método proposto trata do aumento do volume de amostras de estudantes a partir do agrupamento de cursos com perfis semelhantes. Em seguida, há a construção de classificadores binários e multiclasse que visam prever a evasão precoce, a evasão tardia ou a formação estudantil por meio de seis cenários de análise. Um conjunto mínimo de variáveis é observado para a construção dos modelos, abordando, principalmente, o rendimento acadêmico do estudante, considerando a ética no uso de dados pessoais e facilitando a reprodução do trabalho em instituições que não dispõem de uma grande variedade de informações. Modelos são construídos e avaliados em cada cenário com o método de validação cruzada a partir dos algoritmos XGBoost, SVC e Logistic Regression. Além disso, em cada cenário, um *ensemble* dos modelos é proposto para a classificação com base na técnica de votação *soft*, que observa as probabilidades atribuídas a cada classe por cada classificador. Por fim, é executada a interpretação dos modelos com vistas a identificar variáveis importantes para a previsão da evasão a partir do uso de valores Shapley, uma abordagem baseada na teoria dos jogos que atribui de forma justa a contribuição de cada variável na previsão. Os resultados obtidos indicam que é viável considerar em conjunto dados de estudantes de cursos de graduação com características semelhantes para a previsão da evasão, uma vez que os modelos binários de previsão de evasão precoce e de evasão ou formação têm mais de 80% de acurácia, precisão e *recall*, chegando a 90% em alguns cenários, enquanto os modelos multiclasse que buscam prever a evasão precoce, a evasão tardia ou a formação apresentam cerca de 75% de acurácia, precisão e *recall*.

Palavras-chave: Previsão da Evasão no Ensino Superior. Aprendizado de Máquina. Agrupamento de Cursos de Graduação. Classificação da Evasão.

Abstract

Student dropout in higher education occurs when the student abandons the undergraduate course. It is a multifaceted phenomenon that affects not only the individual and the educational institution, but society as a whole. Predicting dropout rates can help develop strategies to combat the phenomenon and encourage student retention and academic success in educational institutions. Therefore, this work aims to predict dropout rates at different times during the undergraduate course using unsupervised and supervised machine learning techniques. The proposed method deals with increasing the volume of student samples by clustering courses with similar profiles. Next, there is the construction of binary and multiclass classifiers that aim to predict early dropout, late dropout or completion through six analysis scenarios. A minimum set of variables is observed to build the classifiers, mainly addressing the student's academic performance, considering ethics in the use of personal data and facilitating the reproduction of work in institutions that do not have a wide variety of information. Classifiers are built and evaluated in each scenario with the cross-validation method using the XGBoost, SVC, and Logistic Regression algorithms. Furthermore, in each scenario, a ensemble of the classifiers is proposed for classification based on the soft voting technique, which observes the probabilities assigned to each class by each classifier. Finally, the classifiers are interpreted to identify important variables for dropout prediction using Shapley values, an approach based on game theory that fairly attributes the contribution of each variable to the prediction. The results obtained indicate that it is feasible to jointly consider data from undergraduate students with similar characteristics to predict dropout rates, since the binary models for predicting early dropout and dropout or training have more than 80% accuracy, precision, and recall, reaching 90% in some scenarios, while multiclass models that seek to predict early dropout, late dropout, or completion present around 75% accuracy, precision, and recall.

Keywords: Higher Education Dropout Prediction. Machine Learning. Undergraduate Courses Clustering. Dropout Classification.

Lista de Figuras

Lista de tabelas

Sumário

1 Introdução

A Constituição da República Federativa do Brasil de 1988 estabelece, no caput de seu Artigo 205, que a educação é direito de todos e dever do Estado (BRASIL, 1988). Em seu Artigo 208, Inciso V, a Constituição determina que o dever do Estado com a educação será efetivado mediante a garantia de acesso aos níveis mais elevados do ensino (BRASIL, 1988). Trata-se da competência do Estado de fornecer educação superior pública de qualidade, buscando garantir o acesso equitativo e efetivo à formação acadêmica, o que evidencia um compromisso inalienável com a promoção da igualdade de oportunidades e o avanço socioeconômico do país. Contudo, não basta garantir o acesso ao ensino superior. É preciso dedicar esforços à formação integral dos estudantes, desenvolvendo e adotando políticas de permanência com o objetivo de mitigar a evasão estudantil (MEC, 2014).

A evasão estudantil no ensino superior é um fenômeno multifacetado que acarreta o abandono do curso de graduação pelo estudante. Trata-se de um problema complexo global que afeta não somente o estudante, mas também a instituição de ensino, o sistema educacional e toda a sociedade. Ao interromper a trajetória educacional, a evasão gera repercussões econômicas e sociais significativas (LOBO, 2012). Os motivos que levam à evasão são diversos e podem interagir de forma complexa, sendo necessária uma análise aprofundada e multidimensional com vistas a compreender os determinantes da evasão, buscando permitir o desenvolvimento de estratégias eficazes para mitigar as suas causas e apoiar a retenção e o sucesso dos estudantes (MEC, 2014).

Existem dois tipos de evasão no ensino superior: a evasão que ocorre em até dois anos após o ingresso no curso de graduação, denominada evasão precoce, e a evasão que acontece após dois anos do ingresso, chamada de evasão tardia (OECD, 2023; VELÁSQUEZ et al., 2003). A evasão precoce costuma estar relacionada a fatores como a falta de adaptação ao ambiente universitário, escolha inadequada do curso, dificuldades financeiras e insuficiência ou inexistência de suporte acadêmico e emocional. Estudantes que não conseguem se integrar à vida acadêmica ou que percebem um desalinhamento entre suas expectativas e a realidade do curso tendem a abandonar os estudos nos primeiros períodos (OECD, 2023; VELÁSQUEZ et al., 2003). Por outro lado, a evasão tardia frequentemente está associada a outros fatores, como o esgotamento acadêmico, desafios pessoais contínuos, mudanças de interesses ou prioridades e dificuldades em conciliar o trabalho com os estudos

(OECD, 2023; VELÁSQUEZ et al., 2003). Ambos os tipos de evasão têm consequências importantes tanto para os indivíduos quanto para as instituições de ensino, o que ressalta a necessidade de políticas eficazes de apoio e retenção estudantil.

Além de aumentar a oferta de vagas no ensino superior, o Estado brasileiro tem o dever de implementar abordagens que garantam a permanência e o sucesso acadêmico dos estudantes (TCU, 2014). É fundamental definir estratégias que ajudem a identificar padrões de evasão e que permitam a previsão do fenômeno, visando desenvolver ferramentas que possam basear intervenções proativas nas instituições de ensino (TCU, 2014). Técnicas de mineração de dados e aprendizado de máquina podem ser utilizadas com o objetivo de construir modelos preditivos que baseiem o desenvolvimento de intervenções personalizadas (COLPO et al., 2024). A depender dos modelos considerados, é possível usar a técnica *SHapley Additive exPlanations* (SHAP) para interpretá-los visando à obtenção de explicações claras e detalhadas sobre os fatores que contribuem para a evasão (LUNDBERG; LEE, 2017). Acredita-se que, combinando políticas públicas eficazes, apoio institucional e tecnologia, pode-se reduzir a evasão e promover um ambiente educacional inclusivo, contribuindo para o desenvolvimento socioeconômico do país.

Diversos trabalhos empregam técnicas de mineração de dados e aprendizado de máquina para a previsão da evasão estudantil em instituições de ensino superior no Brasil e ao redor do mundo, tratando da construção de modelos de classificação no âmbito de um ou mais cursos de graduação (COLPO et al., 2024). Os trabalhos comumente consideram dados de estudantes de apenas um curso, de um conjunto de cursos selecionados arbitrariamente ou de todos os cursos de uma instituição. Além disso, os estudos normalmente propõem a construção de classificadores binários com o objetivo de prever unicamente o enquadramento do estudante nas classes de *"Evasão"* ou *"Formação"*. É válido ressaltar que, a depender dos modelos utilizados, não se pode usar SHAP para interpretá-los, devido à sua complexidade ou natureza. Acredita-se que as estratégias que orientam o desenvolvimento de trabalhos de previsão de evasão podem ser aprimoradas.

Limitar a análise aos dados de estudantes de somente um curso pode acarretar uma série de desafios que possivelmente afetam os resultados obtidos. Primeiramente, a baixa quantidade de amostras de estudantes evadidos em determinados cursos pode inviabilizar a construção de modelos de classificação individuais, o que representa uma dificuldade de reprodução do estudo. Quando possível desenvolver modelos para determinado curso, a

seleção de dados de seus estudantes pode resultar em uma amostra extremamente específica. O viés de amostragem é uma preocupação importante, uma vez que os dados de um único curso certamente não capturam adequadamente a complexidade e as nuances presentes em toda a população estudantil. A falta de diversidade nos dados pode limitar a capacidade do modelo de generalizar para outros contextos, o que inviabiliza o uso de um modelo específico para prever a evasão em outros cursos. Dessa forma, ainda que haja dados em volume significativo para alguns cursos, pode haver cursos que não dispõem de quantidades razoáveis de dados, o que pode ocasionar a não realização do estudo em alguns contextos.

Por outro lado, selecionar arbitrariamente um conjunto de cursos para análise e construção de modelos de previsão de evasão apresenta desafios significativos que podem comprometer a integridade e a eficácia dos resultados obtidos. Essa abordagem está sujeita a potenciais vieses na seleção dos cursos, pois a escolha pode ser influenciada por critérios subjetivos ou preconceitos implícitos. O processo de seleção arbitrária pode ser complexo, requerendo uma avaliação cuidadosa da relevância dos fatores que levam à evasão e seu impacto para cada curso, o que pode ser difícil de quantificar de maneira objetiva. Além disso, a seleção arbitrária pode resultar em uma amostra não representativa da população acadêmica, limitando a capacidade dos modelos de generalizar para outros cursos. Essa abordagem pode introduzir inconsistências e dificuldade de comparar os resultados obtidos para diferentes conjuntos de cursos, prejudicando a interpretação dos modelos e a identificação de padrões consistentes de evasão.

Por fim, utilizar dados de estudantes de todos os cursos de uma instituição para construir os modelos de classificação oferece uma visão abrangente e pode solucionar o problema da baixa quantidade de amostras de estudantes evadidos, mas não é uma estratégia isenta de desvantagens. A diversidade de contextos acadêmicos, socioeconômicos e culturais entre os diferentes cursos pode aumentar a complexidade da análise. Pode ser necessário desenvolver abordagens sofisticadas para lidar com essa diversidade. Os padrões de evasão também podem variar entre os cursos, o que pode causar a necessidade de considerar variáveis adicionais para capturar as diferenças existentes ou mesmo exigir a criação de modelos específicos em vez um modelo genérico, o que pode levar ao problema da baixa disponibilidade de dados tratado anteriormente.

Quanto ao uso de modelos de classificação binária para previsão da evasão, trata-se de uma abordagem objetiva para lidar com o problema da disponibilidade de amostras

de estudantes em cada classe. Ocorre que observar somente duas classes, *"Evasão"* ou *"Formação"*, não permite a identificação da evasão em momentos específicos do curso, o que pode ser viabilizado com o uso de classificadores multiclasse. Porém, o problema da baixa disponibilidade de amostras é agravado com a modelagem multiclasse, uma vez que a análise se torna mais detalhada com o enfoque da evasão em momentos específicos, o que leva à necessidade de aumentar a quantidade de amostras de estudantes que se enquadram em determinadas classes.

Além da dificuldade relacionada ao volume de amostras de estudantes evadidos em momentos específicos do curso, há o problema da disponibilidade variável de informações qualitativas de estudantes, especialmente socioeconômicas, a depender da instituição de ensino e do curso de graduação considerados. Em algumas instituições, há uma menor quantidade de informações disponíveis para análise, seja por dificuldades na obtenção e na gestão dos dados, seja pelo tempo de existência do curso, entre outros fatores. Trata-se de um desafio considerável que impacta diretamente a possibilidade de reprodução de determinados estudos. Além disso, é crucial considerar a ética no uso de dados pessoais ao construir modelos preditivos, sendo fundamental evitar vieses na análise de dados sensíveis para garantir que os modelos não perpetuem discriminações ou injustiças, havendo a necessidade de tratar os dados com equidade e respeito aos indivíduos. O uso de dados socieconômicos pode representar desafios significativos para a interpretação dos resultados.

Acredita-se que (1) o desenvolvimento de uma abordagem que trate de agrupar de forma não supervisionada cursos de graduação com características semelhantes e (2) o uso de dados de estudantes de cursos que integrem um dos grupos gerados, especialmente dados que tratem de rendimento acadêmico, podem contribuir para o aprimoramento do processo de construção de modelos de previsão de evasão, sejam binários, trazendo a informação geral da evasão ou formação ao longo do curso, ou multiclasse, permitindo a identificação da evasão em momentos específicos da trajetória acadêmica. A previsão da evasão de forma detalhada e o uso de modelos passíveis de interpretação podem basear o desenvolvimento de medidas de retenção personalizadas, o que, por sua vez, pode levar ao aumento dos índices de formação estudantil. Além disso, considerar dados sobre o rendimento acadêmico e evitar informações socioeconômicas ao construir classificadores pode ajudar a evitar vieses de interpretação. Ao tratar essencialmente do desempenho acadêmico, os modelos podem avaliar os estudantes de forma mais objetiva e justa, sem

influências externas relacionadas às condições socioeconômicas, evitando discriminações ou preconceitos inadvertidos. A abordagem em questão pode promover uma análise mais equitativa, centrada nas habilidades e no progresso educacional dos estudantes. Diante disso, duas questões de pesquisa são propostas para o desenvolvimento deste trabalho:

- Questão de Pesquisa 1 (QP1): É possível gerar, de forma não arbitrária e não supervisionada, grupos coesos de cursos de graduação (que possuem características possivelmente semelhantes) a serem considerados para a previsão da evasão?
- Questão de Pesquisa 2 (QP2): É possível prever a evasão em momentos específicos do curso de graduação, considerando em conjunto amostras de estudantes de um grupo de cursos semelhantes e observando essencialmente o rendimento acadêmico, de forma a evitar o uso de informações socioeconômicas e a afastar vieses éticos?

1.1 Motivação e Justificativa

A abordagem proposta neste trabalho caracteriza uma estratégia que incorpora elementos das estratégias geralmente observadas em trabalhos relacionados, em que se consideram, para construção de classificadores, dados de estudantes de um único curso de graduação, de um conjunto de cursos ou de todos os cursos ao mesmo tempo. Com a observação de grupos de cursos de graduação, há uma maior quantidade de amostras de estudantes nas situações de interesse deste estudo, o que permite a análise da evasão precoce e da evasão tardia. Além disso, ao considerar em conjunto somente dados de estudantes de cursos semelhantes (com grupos de cursos sendo gerados de forma não supervisionada) em vez de dados de estudantes de conjuntos arbitrários de cursos ou de todos os cursos da instituição, há a redução do risco de o classificador construído ignorar possíveis padrões identificáveis em contextos específicos ou semelhantes.

A estratégia de considerar um conjunto de cursos para a construção de modelos visa permitir a previsão da evasão de estudantes nos cursos observados. Em outras palavras, um único modelo construído com dados de estudantes de um grupo de cursos semelhantes pode permitir a previsão da evasão em todos os cursos daquele grupo. Outra vantagem significativa está no fato de que novos cursos podem ser alocados em grupos existentes à medida que os seus dados estejam disponíveis e que o processo de agrupamento não supervisionado seja reexecutado. Isso permitiria, no âmbito de um novo curso, o uso de

modelos já construídos para um grupo previamente existente. Além disso, a estratégia do aumento do volume de dados de estudantes pode permitir a reprodução do estudo no âmbito de instituições de ensino que não dispõem de quantidades razoáveis de informações de estudantes que se enquadram em situações de evasão específicas.

O agrupamento não supervisionado possui vantagens em relação à seleção arbitrária de cursos. Os algoritmos de aprendizado de máquina são capazes de processar conjuntos de dados de forma eficiente (HAN et al., 2012). Além disso, o agrupamento permite a identificação de padrões complexos e não lineares nos dados (HAN et al., 2012), que podem não ser facilmente percebidos por meio de uma abordagem manual. Isso possibilita a descoberta de *insights* sobre os diferentes perfis de estudantes e suas características associadas à evasão. Outra vantagem está no fato de que o processo não supervisionado é menos suscetível a vieses pessoais ou subjetividade na seleção dos grupos (HAN et al., 2012), o que ajuda a garantir uma distribuição imparcial e representativa dos estudantes em cada grupo de cursos. Por fim, a abordagem tratada é escalável e replicável, permitindo sua aplicação em diferentes bases de dados ou contextos sem a necessidade de intervenção humana no agrupamento propriamente dito (HAN et al., 2012).

Os resultados obtidos neste trabalho indicam a viabilidade de considerar em conjunto dados de estudantes de cursos de graduação com características semelhantes para prever a evasão no ensino superior, pois os modelos binários de previsão de evasão precoce e de evasão ou formação geral têm mais de 80% de acurácia, precisão e *recall*, chegando a 90% em alguns contextos, enquanto os modelos multiclasse que buscam prever a evasão precoce, a evasão tardia ou a formação chegam a cerca de 75% de acurácia, precisão e *recall*.

Ao integrar técnicas do aprendizado de máquina não supervisionado para agrupar cursos de graduação e do aprendizado supervisionado para prever a evasão estudantil, este trabalho propõe uma abordagem robusta com o objetivo de permitir a implementação de intervenções direcionadas em instituições de ensino em geral, mas principalmente naquelas que não dispõem de grandes volumes de informações acerca da evasão e de características socioeconômicas de seus estudantes. Acredita-se que os resultados têm o potencial de basear políticas institucionais e contribuir para a criação de estratégias eficazes que fortaleçam a retenção e a formação dos estudantes em instituições de ensino superior.

1.2 Objetivos

Este trabalho tem como objetivo prever a evasão estudantil no ensino superior. O método proposto trata da geração de grupos de cursos de graduação com características semelhantes a serem considerados visando ao aumento do volume de amostras de estudantes evadidos de forma precoce ou tardia. Para a execução de agrupamentos, são utilizadas técnicas do aprendizado de máquina não supervisionado — os algoritmos K-Means e Aglomerativo. A partir disso, é proposta a construção de modelos de classificação binária, que buscam prever tanto a evasão ou a formação em geral, e multiclasse, que tratam de prever a evasão precoce e a evasão tardia. Os classificadores são construídos a partir de técnicas do aprendizado de máquina supervisionado — os algoritmos XGBoost, Support Vector Classifier e Logistic Regression e o método de *ensemble* de votação Voting Classifier. Após a construção dos modelos, este trabalho propõe interpretá-los utilizando a técnica SHAP, baseada na teoria dos jogos, que distribui de forma justa a contribuição de cada variável na previsão (LUNDBERG; LEE, 2017). Os valores Shapley são essenciais, pois oferecem explicações consistentes sobre a influência de cada característica observada, para facilitar a compreensão do comportamento do modelo (LUNDBERG; LEE, 2017).

Diante do exposto, este trabalho tem os seguintes objetivos específicos:

- Buscar trabalhos relacionados ao agrupamento de cursos de graduação e à previsão da evasão precoce e tardia em instituições de ensino superior;
- Gerar grupos coesos de cursos com características semelhantes;
- Selecionar um grupo de cursos semelhantes para aumentar o volume de amostras de estudantes evadidos em momentos específicos do curso;
- Prever a evasão ou formação geral em um grupo de cursos semelhantes;
- Prever a evasão precoce e a evasão tardia em um grupo de cursos semelhantes;
- Identificar variáveis importantes para a previsão da evasão precoce e tardia.

1.3 Estrutura do Trabalho

Este trabalho é constituído por sete capítulos, sendo este, Introdução, o primeiro, em que foram apresentados os objetos de estudo. A seguir, vêm os capítulos Referencial Teórico, Trabalhos Relacionados, Método e Ferramentas, Agrupamento de Cursos de Graduação, Classificação de Estudantes em Risco de Evasão e Considerações Finais.

2 Referencial Teórico

Este capítulo trata de conceitos relacionados ao método proposto e às ferramentas empregadas para o desenvolvimento deste trabalho. Abordam-se conceitos associados à mineração de dados, tratando dos processos *Knowledge Discovery in Databases* (KDD) e *Cross Industry Standard Process for Data Mining* (CRISP-DM). Em seguida, detalham-se técnicas de análise e processamento de dados, tratando de tarefas de limpeza, identificação e tratamento de *outliers* e de dados ausentes, padronização de dados e transformação de variáveis, além de métodos de codificação de variáveis, como *one hot encoding* e *label encoding*, e de balanceamento de classes, como o *undersampling*, entre outros. Por fim, são apresentados métodos do aprendizado de máquina, que são fundamentais para a análise preditiva e a modelagem de fenômenos complexos, como a evasão estudantil.

No âmbito do aprendizado de máquina, este capítulo trata tanto de técnicas de agrupamento quanto de classificação. Para o agrupamento, são apresentados os algoritmos K-Means e Aglomerativo, que são empregados para identificar padrões em dados. Diversas métricas de avaliação de agrupamentos são discutidas para assegurar a qualidade dos grupos formados. Em seguida, o capítulo explora métodos de classificação, incluindo XGBoost, Support Vector Classifier (SVC), Logistic Regression e Voting Classifier. Para a construção dos modelos, são utilizados os métodos de amostragem de dados validação cruzada (*cross-validation*) e *hold-out*. Por fim, o capítulo aborda as métricas de avaliação de classificação, que incluem acurácia, precisão, *recall*, *f1-score*, e AUC/ROC, além de tratar da interpretação dos modelos utilizando a técnica de SHAP.

2.1 Mineração de Dados

De acordo com Han et al. (2012), a mineração de dados compreende um processo de exploração e análise de grandes conjuntos de dados com o objetivo de descobrir padrões, tendências e conhecimento útil, que podem basear a tomada de decisões, a previsão de comportamentos futuros ou a melhor compreensão de fenômenos específicos. Cada vez mais percebe-se que, com o avanço da tecnologia e o aumento na disponibilidade de dados, a mineração de dados desempenha um papel fundamental na extração de valor a partir de grandes volumes de informações (HAN et al., 2012). O estudo de Fayyad et al. (1996)

demonstra que, há décadas, a mineração de dados é empregada em diversas áreas, tanto nas ciências quanto nos negócios.

Para a execução de trabalhos de mineração de dados, comumente são observados métodos amplamente difundidos, como o *Knowledge Discovery in Databases* e o *Cross Industry Standard Process for Data Mining*, que propõem etapas fundamentais para o uso de dados em processos de descoberta de conhecimento (FAYYAD et al., 1996; CHAPMAN et al., 2000). Os processos abrangem desde a compreensão dos dados até a utilização do conhecimento e dos artefatos obtidos, que geralmente são gerados com o emprego de técnicas de aprendizado de máquina.

O *Knowledge Discovery in Databases* é um processo sistemático de mineração de dados desenvolvido em ambiente acadêmico que visa à descoberta de conhecimento útil a partir de grandes volumes de dados (FAYYAD et al., 1996). Por meio de cinco etapas que compreendem tarefas diversas, o processo propõe a seleção e o processamento de dados para uso na mineração propriamente dita, etapa seguida pela interpretação e avaliação dos resultados percebidos. Ao introduzir o KDD, Fayyad et al. (1996) consideraram a expressão "*Knowledge Discovery in Databases*" um sinônimo de mineração de dados; entretanto, é válido ressaltar que "mineração de dados" também é o nome de uma das etapas do processo, aquela em que geralmente ocorre o emprego de técnicas de aprendizado de máquina para a extração de conhecimento.

O KDD é um processo iterativo, ou seja, cada uma de suas etapas deve ser refinada e reexecutada à medida que são obtidos novos *insights* ou que são ajustados os objetivos do trabalho. Trata-se de uma abordagem fundamental para pesquisas de mineração de dados, amplamente utilizada em áreas diversas, que permite a transformação de dados brutos em conhecimento útil e acionável (FAYYAD et al., 1996).

O *Cross Industry Standard Process for Data Mining* é um processo sistemático de mineração de dados desenvolvido em ambiente industrial que objetiva a descoberta de conhecimento com foco na compreensão do negócio (CHAPMAN et al., 2000). O CRISP-DM é constituído por seis fases, sendo as duas primeiras as fases de entendimento do negócio e entendimento dos dados, que surgem explícitas no processo em questão, diferentemente do que ocorre no KDD. As outras fases, que envolvem a preparação dos dados, modelagem, avaliação e implantação, completam o ciclo de vida do projeto.

Assim como o KDD, o CRISP-DM é iterativo, ou seja, deve-se refinar e reexecutar

fases do processo à medida que surge a necessidade. A flexibilidade do método permite o seu ajuste de acordo com as especificidades do projeto (CHAPMAN et al., 2000). Apesar de ser proveniente da indústria, o CRISP-DM é, há décadas, amplamente observado em trabalhos acadêmicos, uma vez que proporciona uma estrutura sólida para orientar o desenvolvimento de soluções baseadas em dados (MARTíNEZ-PLUMED et al., 2019).

Em resumo, o KDD e o CRISP-DM são metodologias amplamente utilizadas em trabalhos de mineração de dados, compartilhando o objetivo de extrair conhecimento útil a partir de grandes conjuntos de dados. Ambos os métodos seguem uma abordagem baseada em etapas, envolvendo desde a compreensão do problema até a implementação dos resultados. No entanto, suas diferenças estão na ênfase e na estrutura detalhada de cada etapa. O CRISP-DM é mais prescritivo e detalhado em termos de etapas específicas, enquanto o KDD oferece uma visão mais abrangente, englobando desde a seleção dos dados até a interpretação dos resultados. Combinar elementos de ambos os métodos para a execução de um trabalho de mineração de dados pode proporcionar uma abordagem completa e flexível, permitindo uma análise eficaz.

2.2 Análise e Processamento de Dados

A análise de dados constitui uma tarefa essencial em todas as etapas do processo de mineração de dados, baseando desde a compreensão dos dados para a definição das perguntas de pesquisa até a interpretação dos resultados obtidos na mineração propriamente dita (HAN et al., 2012). Ao possibilitar o entendimento dos dados considerados, a análise compreende a identificação de problemas existentes, bem como permite a definição de transformações necessárias na estrutura dos conjuntos visando ao uso dos dados para a mineração (CHAPMAN et al., 2000).

Diversos gráficos e visualizações podem ser empregados durante a análise de dados, designando ferramentas para a exploração e a compreensão dos conjuntos de dados. Com a análise, pode-se, por exemplo, identificar as quantidades de variáveis (atributos ou características) e de registros disponíveis nos conjuntos, bem como os tipos de dados de cada variável. Além disso, é possível detectar dados ausentes e observar a distribuição de frequências de dados por variável. Ainda tratando de análises comuns, em variáveis quantitativas pode-se identificar *outliers*, verificar se há variáveis colineares, observar a

escala das variáveis, além de perceber tendências, distribuições e padrões nos dados. As análises citadas são importantes para a avaliação da qualidade e integridade das informações disponíveis e para o levantamento de processamentos necessários (HAN et al., 2012).

Em um conjunto de dados numérico, *outliers* designam valores muito discrepantes ou anômalos, ou seja, valores que se diferenciam significativamente de outros, podendo ser observações muito distantes da média ou da mediana do conjunto (BRUCE; BRUCE, 2019). A presença de *outliers* em um conjunto de dados pode ter um impacto significativo nas análises estatísticas realizadas, influenciando medidas como a média e o desvio padrão, bem como pode influenciar o emprego de técnicas de aprendizado de máquina, em especial, de agrupamento de dados (HAN et al., 2012).

Existem várias razões pelas quais *outliers* podem ocorrer em conjuntos de dados, incluindo erros de medição ou digitação e variações genuínas. Identificar e lidar com *outliers* de forma apropriada é uma parte crucial da análise de dados que visa garantir a precisão e a confiabilidade do uso dos dados (BRUCE; BRUCE, 2019). O método do intervalo interquartil (*Interquartile Range* — IQR) e os diagramas de caixa (*boxplots*) são particularmente úteis para a identificação da existência de *outliers* em um conjunto numérico (BRUCE; BRUCE, 2019).

Há várias abordagens para lidar com *outliers*, a depender da causa de sua ocorrência e dos objetivos da análise. Entre os métodos de tratamento comuns, estão a remoção dos dados que designam *outliers* e a transformação dos dados considerados *outliers* (HAN et al., 2012). A simples remoção de *outliers* normalmente não é a melhor solução, pois os *outliers* podem conter informações valiosas sobre o fenômeno estudado, e sua exclusão pode distorcer a interpretação dos resultados. A substituição dos *outliers* pelo limite inferior ou superior do intervalo interquartil designa um tratamento que evita o descarte dos dados (HAN et al., 2012). Porém, é preciso identificar as situações em que os *outliers* realmente representam variações genuínas nos dados, o que dispensa a necessidade de tratamento.

O diagrama de caixa é um instrumento gráfico que fornece uma representação visual da distribuição estatística de um conjunto de dados numérico. O diagrama é útil para identificar a presença de *outliers* e avaliar a simetria e a dispersão dos dados (BRUCE; BRUCE, 2019). A caixa representa o intervalo interquartil, que é a diferença entre o primeiro quartil ($Q1$) e o terceiro quartil ($Q3$). A parte inferior da caixa é $Q1$, enquanto a linha no meio da caixa é a mediana ($Q2$), e a parte superior é $Q3$. A caixa,

portanto, abrange o IQR, que compreende a maioria dos dados. Os "bigodes" (*whiskers*) do diagrama se estendem a partir da caixa até os valores extremos que estão dentro de uma certa distância dos quartis. Geralmente, a distância padrão é definida como $1,5 * IQR$. Os pontos fora da distância são considerados *outliers* e são representados como pontos individuais (BRUCE; BRUCE, 2019).

Colinearidade é um termo utilizado em estatística para designar a situação em que duas ou mais variáveis independentes estão altamente correlacionadas entre si (BRUCE; BRUCE, 2019). Tratando de aprendizado de máquina não supervisionado, algoritmos de agrupamento afetados pela colinearidade podem ser sensíveis a pequenas alterações nos dados de entrada, o que pode levar a resultados menos confiáveis (HAN et al., 2012). Quanto ao aprendizado de máquina supervisionado, quando duas variáveis independentes estão altamente correlacionadas, torna-se difícil determinar a contribuição individual de cada variável para a previsão da variável dependente, pois as duas variáveis fornecem informações semelhantes ao modelo (HAN et al., 2012).

Algumas abordagens para lidar com colinearidade incluem a remoção de variáveis correlacionadas, a combinação de variáveis em uma única variável, quando apropriado, ou o uso de técnicas mais avançadas, como regularização. Em resumo, a colinearidade é uma preocupação importante ao utilizar algoritmos de aprendizado de máquina, e sua presença deve ser avaliada e tratada adequadamente para garantir a validade e interpretabilidade dos resultados (HAN et al., 2012).

Após a análise dos conjuntos de dados coletados, possíveis problemas que afetam a qualidade dos dados devem ser percebidos e as devidas soluções devem ser definidas para a realização de tratamentos, o que designa tarefas de processamento de dados (FAYYAD et al., 1996). Além disso, devem ser levantadas as alterações a serem efetuadas para a adequação dos dados observando o problema de pesquisa e visando à busca pelo atingimento dos objetivos do trabalho, o que caracteriza tarefas de transformação de dados (FAYYAD et al., 1996). O KDD apresenta a distinção entre as etapas de pré-processamento e transformação de dados, enquanto o CRISP-DM engloba todas as atividades citadas em sua fase de preparação de dados (CHAPMAN et al., 2000).

As tarefas de processamento de dados visam à adequação dos conjuntos de dados para uso na etapa de mineração, por exemplo, como entrada de algoritmos de aprendizado de máquina (FAYYAD et al., 1996; CHAPMAN et al., 2000). Entre operações comuns de

processamento dos dados, citam-se tarefas de tratamento de dados ausentes, inconsistentes ou incorretos, por meio de limpeza de dados, imputação de dados e tratamento de *outliers*. Também são comuns as operações que tratam da criação e alteração de variáveis, *one-hot encoding* (codificação ou binarização), *label encoding* (codificação de classes), padronização e *undersampling* (subamostragem) de dados.

Undersampling é uma técnica de seleção de dados comumente utilizada para lidar com conjuntos de dados desbalanceados, nos quais uma ou mais classes de uma variável têm uma quantidade muito maior de observações em comparação com outra(s) classe(s). O desbalanceamento pode ser problemático, especialmente em problemas de classificação, do aprendizado de máquina supervisionado, pois alguns algoritmos podem ter dificuldade em aprender sobre a(s) classe(s) minoritária(s), resultando em modelos tendenciosos em direção à(s) classe(s) majoritária(s) (HAN et al., 2012).

Se a variável tratada possui duas classes, o *undersampling* visa reduzir a quantidade de observações da classe majoritária para equilibrar a distribuição de frequências entre as classes. Se a variável tiver mais de duas classes, a lógica será análoga, buscando o equilíbrio da quantidade de observações entre as todas as classes, tendo a menor quantidade de observações das classes como objetivo. O *undersampling* pode ser realizado de diversas formas, sendo o *random undersampling* uma variante bastante empregada. O *random undersampling* trata de remover aleatoriamente observações da(s) classe(s) majoritária(s) até que o número de observações em nas classes seja mais equilibrado (HAN et al., 2012).

A imputação de dados é uma técnica usada com o objetivo de lidar com valores ausentes ou faltantes nos conjuntos de dados tratados. Valores ausentes podem ocorrer por diversos motivos, como erros de coleta, falhas nos instrumentos de medição, ou mesmo pela falta de registro de determinadas informações (HAN et al., 2012). A imputação de dados envolve preencher ou estimar os valores ausentes com base em informações disponíveis nos dados. Existem várias técnicas de imputação, e a escolha da abordagem depende do contexto e da natureza dos dados. Considerando variáveis quantitativas, algumas das técnicas comuns tratam da substituição dos dados ausentes por medidas de tendência central, com o uso da média, mediana ou moda dos dados existentes (HAN et al., 2012).

A padronização, também conhecida como normalização *z-score*, é um método de processamento de dados que visa colocar todas as variáveis em uma escala comum. O objetivo é garantir que as variáveis tenham a mesma ordem de grandeza, o que pode ser

importante para algoritmos de aprendizado de máquina que são sensíveis às diferenças nas escalas das variáveis (HAN et al., 2012). A padronização é particularmente importante em algoritmos que usam medidas de distância, como o K-Means, em que as diferenças absolutas nas escalas podem distorcer os resultados. Além disso, em métodos que envolvem otimização numérica, como diversos algoritmos de aprendizado de máquina supervisionado, a padronização pode acelerar a convergência (HAN et al., 2012).

One-hot encoding é uma técnica de codificação que visa à representação de variáveis qualitativas ou categóricas em formato quantitativo ou numérico. Em outras palavras, a técnica envolve a transformação de variáveis qualitativas em variáveis quantitativas (HAN et al., 2012). O *one-hot encoding* é comumente aplicado quando se lida com variáveis qualitativas nominais, ou seja, quando não há uma ordem intrínseca entre as categorias. A representação numérica de dados categóricos é fundamental para o uso de diversos algoritmos de aprendizado de máquina (HAN et al., 2012). O *one-hot encoding* compreende a criação de variáveis binárias para representar cada categoria única presente na variável qualitativa original. Cada nova variável binária indica a presença ou ausência da categoria correspondente para uma observação específica. A depender da quantidade de categorias das variáveis qualitativas empregadas no processo, pode haver um aumento considerável na quantidade de variáveis presentes no conjunto de dados final (HAN et al., 2012).

Assim como o *one-hot encoding*, o *label encoding* é uma técnica de representação de dados qualitativos de forma quantitativa, ou seja, trata da transformação de variáveis qualitativas em variáveis quantitativas. Porém, diferentemente do *one-hot encoding*, o uso do *label encoding* é apropriado quando se lida com dados categóricos ordinais, que possuem uma ordem intrínseca (HAN et al., 2012). O *label encoding* atribui a cada categoria da variável um número inteiro único, preservando a ordem original das categorias no conjunto de dados transformado. Em outras palavras, a técnica em questão é particularmente relevante quando as categorias apresentam relações de precedência ou hierarquia, uma vez que permite que algoritmos de aprendizado de máquina levem em consideração a ordem subjacente dos dados, o que pode proporcionar resultados mais significativos e contextualmente precisos (HAN et al., 2012).

2.3 Aprendizado de Máquina

O aprendizado de máquina (*machine learning)* é um subcampo da inteligência artificial que trata do desenvolvimento de algoritmos e modelos computacionais capazes de aprender a partir de dados, sem haver a necessidade de programação explícita. Os modelos construídos são capazes de identificar padrões nos dados e usá-los para fazer previsões ou tomar decisões com base em novas informações. Por meio da exposição a grandes volumes de dados, os modelos tornam-se eficazes em tarefas específicas, como agrupamento, classificação e regressão (HAN et al., 2012). O aprendizado de máquina é muito empregado em trabalhos de mineração de dados em áreas diversas, como processamento de linguagem natural, reconhecimento de imagem, diagnóstico médico e análises educacionais (FAYYAD et al., 1996; CHAPMAN et al., 2000; COLPO et al., 2024).

Entre os tipos de aprendizado de máquina, existem o aprendizado de máquina não supervisionado e o aprendizado de máquina supervisionado. No aprendizado não supervisionado, o algoritmo recebe como entrada conjuntos de dados não rotulados, com o objetivo de identificar padrões ou estruturas intrínsecas nos dados, como agrupamentos ou representações latentes (HAN et al., 2012). O agrupamento é uma técnica de aprendizado de máquina não supervisionado. No aprendizado supervisionado, o algoritmo é treinado com o uso de um conjunto de dados rotulado, no qual as entradas são associadas a rótulos (*labels*) ou saídas desejadas. O objetivo é capacitar o modelo a generalizar padrões e fazer previsões precisas para dados não vistos durante o treinamento (HAN et al., 2012). A classificação e a regressão são técnicas de aprendizado de máquina supervisionado.

2.3.1 Agrupamento

O agrupamento de dados, também chamado de clusterização, é uma técnica de aprendizado de máquina não supervisionado que envolve a divisão de um conjunto de dados em grupos ou *clusters*, de forma que os elementos que compõem cada grupo sejam mais semelhantes entre si do que semelhantes aos elementos de outros grupos. O objetivo do agrupamento é identificar estruturas intrínsecas nos dados e agrupar instâncias que compartilham características comuns (HAN et al., 2012).

O agrupamento é aplicado em diversas áreas, como segmentação de mercado, organização de grandes conjuntos de dados, compressão de imagem, reconhecimento de

padrões, entre outras. (HAN et al., 2012). Existem diversos algoritmos de agrupamento e cada um deles propõe abordagens distintas. A escolha do algoritmo a ser usado depende da natureza dos dados e dos objetivos da análise. Entre os algoritmos populares, estão o K-Means e o agrupamento hierárquico Aglomerativo (HAN et al., 2012). Entre tarefas comuns de mineração de dados com agrupamento, geralmente destacam-se a definição da quantidade de grupos a serem gerados, o emprego de métricas de avaliação dos grupos e a geração de gráficos para interpretação dos grupos gerados (HAN et al., 2012).

O K-Means é um popular algoritmo de agrupamento que pertence à categoria dos métodos de particionamento não hierárquicos. Em linhas gerais, o K-Means tem como objetivo dividir um conjunto de dados em k grupos distintos, sendo o valor k recebido como entrada, buscando atribuir cada observação do conjunto de dados a um grupo de forma a minimizar a variância entre os elementos do grupo (variância intra-grupo), ou seja, a soma das distâncias quadráticas entre cada ponto e o centroide do grupo ao qual o elemento pertence, ao mesmo tempo em que se deseja maximizar a variância entre elementos de grupos distintos (variância inter-grupo) (HAN et al., 2012).

O K-Means opera iterativamente, atribuindo inicialmente k centroides aleatórios, designando pontos ao grupo mais próximo, recalculando os centroides como a média dos pontos no grupo e repetindo esse processo até que a convergência seja alcançada. O K-Means é eficaz em espaços de alta dimensionalidade e é comumente aplicado em áreas como análise de dados, segmentação de clientes e processamento de imagem, sendo uma técnica rápida e escalável para a formação de grupos (HAN et al., 2012). Apesar de sua ampla aplicação, o K-Means tem limitações, como a sensibilidade à inicialização dos centroides. Estratégias de inicialização cuidadosas e o uso de variantes, como o *K-Means++*, podem ajudar a mitigar as limitações. Em geral, o K-Means oferece uma abordagem eficaz e computacionalmente eficiente para a tarefa de agrupamento em grandes conjuntos de dados (HAN et al., 2012).

Por sua vez, o agrupamento hierárquico Aglomerativo é um método de agrupamento em que cada instância de exemplo começa como um grupo individual e, iterativamente, grupos próximos são mesclados para formar grupos maiores até que todas as instâncias estejam em um único grupo (HAN et al., 2012). A escolha de quais grupos unir em cada etapa é determinada pelo método de ligação (*linkage*), que define a proximidade entre grupos. A medida de proximidade entre grupos é usualmente definida pela distância

euclidiana, mas outras métricas podem ser utilizadas. O agrupamento Aglomerativo resulta em uma árvore hierárquica, chamada de dendrograma, que exibe a relação de proximidade entre os grupos. A principal vantagem do agrupamento hierárquico Aglomerativo é sua capacidade de revelar estruturas hierárquicas nos dados, permitindo a análise em diferentes níveis de granularidade (HAN et al., 2012). A Figura 1 exibe um dendograma.

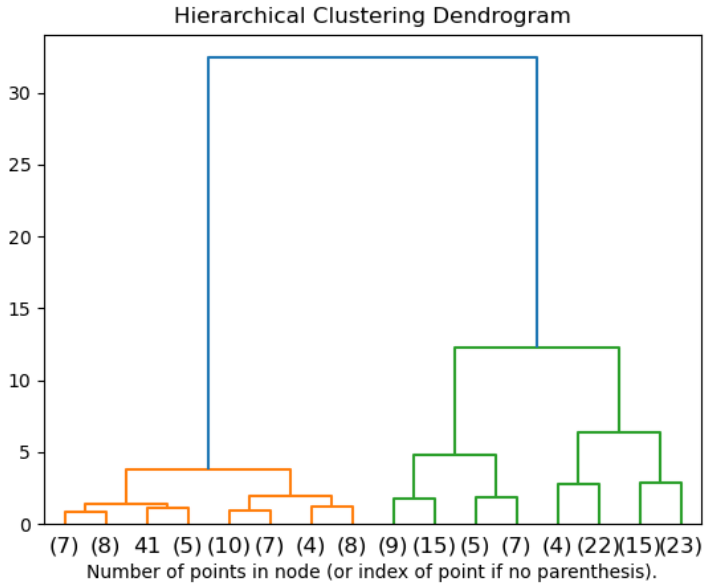

Figura 1 – Dendograma. Fonte: Scikit-Learn (2024).

Existem diversos métodos de ligação[1] (*linkage*), que têm características específicas. A ligação simples (*simple linkage*) calcula a menor distância entre pontos de diferentes grupos, sendo suscetível a formar grupos alongados e podendo ser afetados por ruído, resultando no efeito de "cadeia". Por outro lado, a ligação média (*average linkage*) calcula a média das distâncias entre todos os pares de pontos em diferentes grupos, representando um compromisso entre os métodos de ligação simples e completa, promovendo a formação de grupos de tamanhos semelhantes. A ligação completa (*complete linkage*) calcula a maior distância entre pontos de diferentes grupos, tendendo a criar grupos mais compactos e esféricos, sendo menos sensível ao ruído e *outliers*. Por fim, a ligação de Ward (*Ward linkage*) minimiza o aumento da soma dos quadrados das distâncias dentro do grupo ao combinar dois grupos. Esse método geralmente resulta em grupos de tamanhos mais próximos, sendo eficaz para minimizar a variabilidade interna dos grupos. A Figura 2 apresenta os distintos métodos de ligação do agrupamento Aglomerativo.

[1] https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html

Apesar de sua interpretabilidade, o agrupamento hierárquico Aglomerativo pode ser computacionalmente intensivo, especialmente em grandes conjuntos de dados, e sua sensibilidade à escolha da métrica de proximidade e ao método de ligação entre grupos pode afetar os resultados. A escolha adequada desses parâmetros é crucial para obter grupos significativos. Em resumo, o agrupamento hierárquico Aglomerativo é uma abordagem útil para explorar estruturas hierárquicas em dados, proporcionando uma visão detalhada das relações de proximidade entre grupos (HAN et al., 2012).

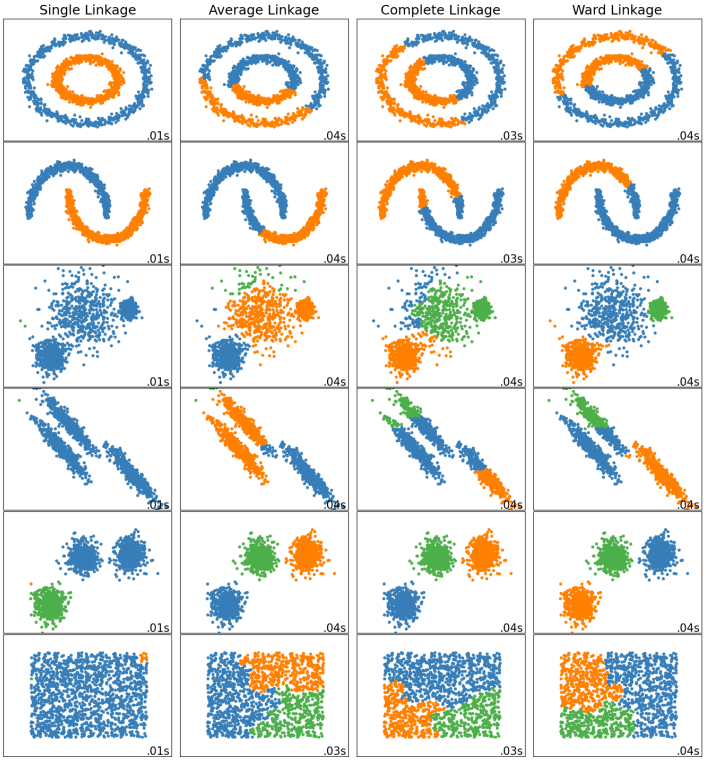

Figura 2 – Métodos de ligação do agrupamento Aglomerativo. Fonte: Scikit-Learn (2024).

Os algoritmos K-Means e Aglomerativo representam abordagens distintas para o agrupamento de dados. Enquanto o K-Means busca definir diretamente os grupos com base em centroides e minimização da inércia, o Aglomerativo adota uma abordagem hierárquica, unindo gradualmente os elementos mais semelhantes. Essa diferença fundamental reflete-se na forma como cada algoritmo lida com a similaridade entre os elementos e na estruturação dos grupos. Enquanto o K-Means usa a inércia como métrica para determinar a separação dos grupos, o Aglomerativo emprega medidas de distância e métodos de ligação para determinar a proximidade entre grupos. Além disso, o K-Means requer a definição prévia do número de grupos, enquanto o Aglomerativo não requer, o que permite uma abordagem

mais exploratória (HAN et al., 2012).

Métricas de avaliação são ferramentas importantes para entender a qualidade dos grupos gerados por algoritmos de agrupamento (HAN et al., 2012). Diferentes métricas podem ser relevantes dependendo da natureza dos dados e dos objetivos específicos da análise. Entre as métricas usadas, estão a inércia, que baseia o método do cotovelo, o coeficiente de silhueta e os índices de Calinski-Harabasz, Davies-Bouldin e Dunn.

A inércia é uma métrica que quantifica a compactação dos grupos formados, representando a soma das distâncias quadradas das amostras aos centros de seus respectivos grupos. Formalmente, a inércia é calculada como a soma das diferenças quadráticas entre cada ponto de dado e o centroide do grupo ao qual pertence, servindo como uma medida de quão bem os pontos estão agrupados. Menores valores de inércia indicam grupos mais coesos e bem definidos, enquanto valores mais altos sugerem maior dispersão dentro dos grupos. Essa métrica é frequentemente usada para avaliar a qualidade do agrupamento e para determinar o número ótimo de grupos em conjunto com outras técnicas, como o método do cotovelo (HAN et al., 2012).

O método do cotovelo[2] é uma técnica visual usada para determinar o número ideal de grupos k em um agrupamento. Ele envolve a execução do algoritmo de agrupamento para diferentes valores de k e a exibição da variância intra-grupo em relação a k. O ponto em que a curva forma um "cotovelo" indica um equilíbrio entre a redução da variância intra-grupo e a complexidade do modelo. Esse valor de k é frequentemente escolhido como o número ótimo de grupos do agrupamento. A Figura 3 exibe um gráfico que representa o uso do método do cotovelo.

O coeficiente de silhueta[3] mede a coesão interna e a separação entre grupos. Para cada observação i, o Coeficiente de Silhueta é calculado como $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, em que $a(i)$ é a média da distância intra-grupo (distância média para os pontos no mesmo grupo) do grupo i e b é a média da distância mais próxima inter-grupo (distância média para os pontos no grupo mais próximo). O coeficiente de silhueta varia de -1 a 1, em que valores mais altos indicam uma melhor separação entre grupos. Por sua vez, o índice de Calinski-Harabasz[4] é uma métrica que avalia a separação entre grupos em relação à dispersão intra-grupo. Quanto maior o índice, melhor a qualidade dos grupos. É calculado

[2] https://www.scikit-yb.org/en/latest/api/cluster/elbow.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html

como $ch(i) = \frac{(n-k)\cdot a}{(k-1)\cdot b(i)}$, em que n é o número total de amostras no conjunto de dados, k é o número de grupos, a é a dispersão inter-grupo e $b(i)$ é a dispersão intra-grupo do grupo i. A Figura 4 apresenta um gráfico construído a partir do coeficiente de silhueta.

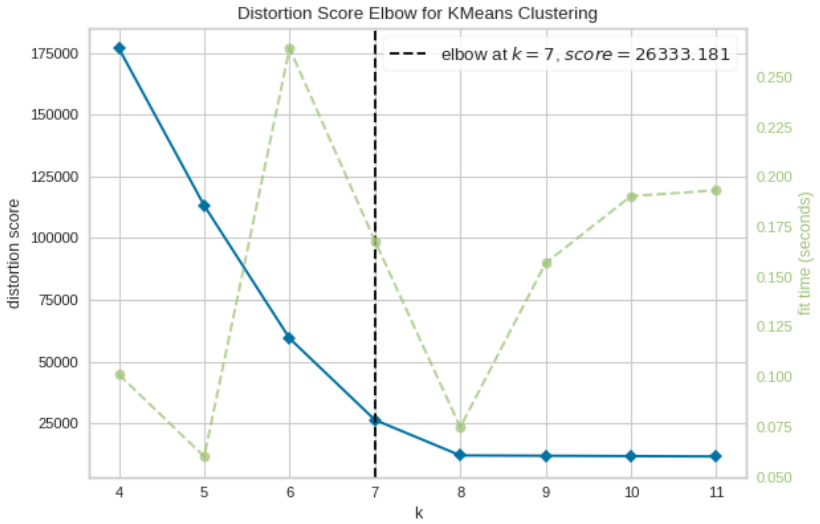

Figura 3 – Método do cotovelo. Fonte: Yellowbrick (2024).

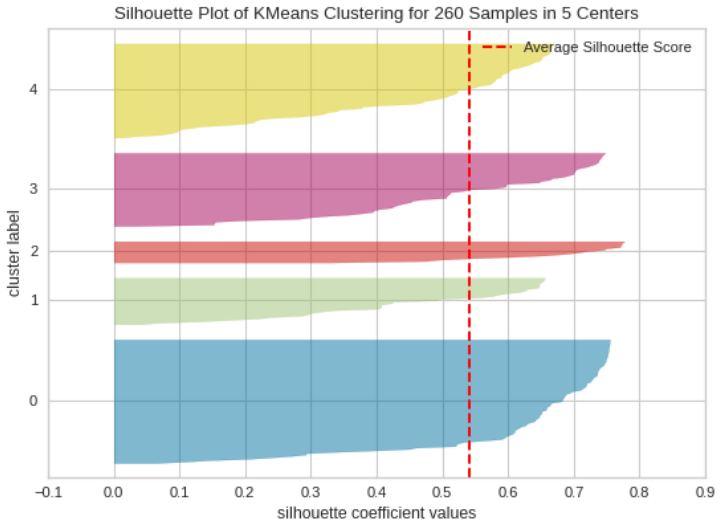

Figura 4 – Gráfico da silhueta. Fonte: Yellowbrick (2024).

O índice de Davies-Bouldin[5] mede a compactação e a separação dos grupos. Ele é calculado como $db = \frac{1}{k}\sum_{i=1}^{k}\left(\max_{j\neq i}\left(\frac{S_i+S_j}{d_{ij}}\right)\right)$, em que k é o número de grupos, $\sum_{i=1}^{k}\ldots$ é a soma entre todos os grupos, $\max_{j\neq i}\left(\frac{S_i+S_j}{d_{ij}}\right)$ é o máximo da razão entre a soma das

[5] https://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index

dispersões intra-grupo dos grupos i e j e a distância entre os centros dos grupos i e j (a maximização ocorre sobre todos os grupos $j \neq i$), $S(i)$ e $S(j)$ são as dispersões intra-grupo para os grupos i e j, respectivamente, e d_{ij} é a distância entre os centros dos grupos i e j. Valores mais baixos indicam uma melhor qualidade do agrupamento, considerando que grupos mais compactos e bem separados têm índices mais baixos.

O índice de Dunn é uma métrica que considera tanto a separação entre os grupos quanto a sua coesão. A fórmula do índice, $D = \frac{\min_{i \neq j}\{d(C_i, C_j)\}}{\max_k\{\Delta(C_k)\}}$, expressa essa relação, em que $d(C_i, C_j)$ representa a menor distância entre os pontos dos grupos C_i e C_j, e $\Delta(C_k)$ é a maior distância entre os pontos dentro do grupo C_k. Um valor mais alto de D indica uma melhor separação entre os grupos e uma menor dispersão dentro deles, sugerindo uma formação de grupos mais eficaz e distinta. No entanto, é importante considerar que o índice de Dunn é sensível a *outliers* e deve ser interpretado com cautela, levando em conta o contexto dos dados e a natureza dos grupos formados (DUNN, 1973).

2.3.2 Classificação

A classificação de dados é uma técnica de aprendizado de máquina comumente empregada para a mineração de dados. Ela envolve a atribuição de rótulos ou categorias a instâncias de dados com base em suas características. O objetivo é construir um modelo de classificação ou classificador capaz de generalizar padrões a partir de dados rotulados conhecidos, a fim de fazer previsões precisas sobre novos dados não rotulados. Dessa forma, a classificação é uma tarefa em que os modelos são treinados para categorizar instâncias em classes ou rótulos predefinidos (HAN et al., 2012).

A classificação de dados pode ser dividida em duas categorias principais: classificação binária e classificação multiclasse. A distinção entre essas duas categorias está relacionada ao número de classes ou categorias que o modelo busca prever (HAN et al., 2012). Na classificação binária, o modelo é treinado para classificar as instâncias de dados em duas classes distintas (HAN et al., 2012). Na classificação multiclasse, o modelo é treinado para classificar as instâncias de dados em mais de duas classes. Cada instância pode pertencer a uma de várias categorias exclusivas (HAN et al., 2012).

O treinamento e a avaliação dos modelos de classificação são etapas cruciais na construção dos modelos (HAN et al., 2012). Métodos de reamostragem de dados como *hold-*

out e validação cruzada (*cross-validation*) são amplamente empregados para a construção de modelos de classificação, compreendendo as tarefas de treinamento e avaliação. Para a construção de classificadores, costuma-se utilizar algoritmos ou modelos de classificação amplamente difundidos pela literatura e pela comunidade acadêmica e científica. Além disso, além de modelos de classificação individuais, é comum construir *ensembles* (comitês).

No método *hold-out*, o conjunto de dados é dividido em dois subconjuntos: um conjunto de treinamento e um conjunto de teste. O conjunto de treinamento é utilizado para treinar o modelo, enquanto o conjunto de teste é usado para avaliar a performance do modelo. Esta abordagem é simples e rápida, mas pode apresentar alta variância na estimativa do desempenho do modelo, especialmente para conjuntos de dados menores.

Na validação cruzada, ocorre a divisão do conjunto de dados em grupos (*folds*) para realizar o treinamento e avaliação do modelo em diferentes combinações de grupos, visando obter uma avaliação justa do desempenho do modelo (HAN et al., 2012). A validação cruzada visa garantir que o modelo construído generalize de forma adequada para dados não utilizados durante o treinamento. O objetivo principal é estimar a capacidade de generalização do modelo para novos dados, evitando que a avaliação seja enviesada pelo conjunto de dados específico usado para treinar e avaliar o modelo (HAN et al., 2012). Em resumo, vários modelos são construídos e avaliados durante o processo de validação cruzada, um em cada iteração do processo, sendo comum resumir os resultados gerais do modelo por meio da média das métricas (HAN et al., 2012).

Stratified k-fold[6] é uma técnica de validação cruzada que garante a preservação da distribuição das classes nos conjuntos de treino e teste. Nesse método, o conjunto de dados é dividido em k subconjuntos estratificados, sendo que cada subconjunto contém uma proporção equivalente das classes originais. Em cada iteração do *k-fold*, um subconjunto é retido como conjunto de teste enquanto os restantes são usados para treinamento do modelo, o que ajuda a reduzir o viés introduzido pela aleatoriedade na divisão dos dados, proporcionando uma estimativa mais confiável do desempenho do modelo. A Figura 5 exibe a representação de iterações do método de validação cruzada *stratified k-fold*.

Entre algoritmos de classificação comumente empregados na atualidade, citam-se XGBoost, Support Vector Machine (SVM) e Logistic Regression, entre outros. É válido notar que algoritmos de classificação binária podem ser utilizados em problemas multiclasse

[6] https://scikit-learn.org/stable/modules/cross_validation.html#stratified-k-fold

por meio das estratégias *um contra todos*[7] e *um contra um*[8]. A estratégia *um contra todos* envolve treinar um classificador binário para cada classe, em que cada classificador é treinado para distinguir uma classe específica das demais. Por outro lado, a estratégia *um contra um* treina um classificador binário para cada par possível de classes.

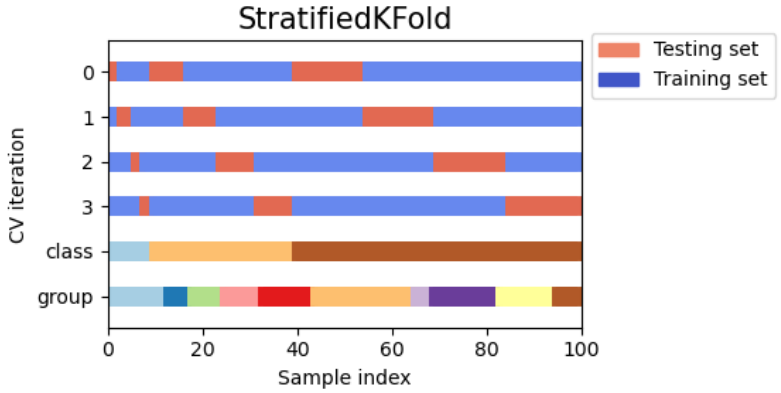

Figura 5 – Método de validação cruzada StratifiedKFold. Fonte: Scikit-Learn (2024).

O XGBoost, acrônimo de *eXtreme Gradient Boosting*, é um poderoso algoritmo de aprendizado de máquina que se destaca em problemas de classificação e regressão. Ele pertence à família de algoritmos de *boosting*, uma técnica de aprendizado de máquina que constrói um modelo preditivo por meio da combinação de múltiplos modelos de menor desempenho, geralmente árvores de decisão, para formar um modelo mais robusto e preciso. O XGBoost é conhecido por sua velocidade, escalabilidade e habilidade para lidar com grandes conjuntos de dados. Além disso, oferece recursos integrados para seleção de características, tratamento de dados ausentes e suporte a múltiplas métricas de avaliação do modelo. A combinação de sua flexibilidade, desempenho e facilidade de uso contribuiu significativamente para sua adoção generalizada em trabalhos de aprendizado de máquina, o que destaca o XGBoost como uma ferramenta versátil e poderosa (XGBOOST, 2021).

Support Vector Machines (SVMs) são algoritmos comumente utilizados para tarefas de classificação e regressão. O principal objetivo de uma Support Vector Machine (SVM) é encontrar um hiperplano de decisão que maximize a margem entre classes em um espaço de características (HAN et al., 2012). A margem é a distância entre o hiperplano e os pontos mais próximos de cada classe, que são chamados de vetores de suporte. SVMs são eficazes em espaços de alta dimensionalidade e podem lidar com dados não lineares

[7] https://scikit-learn.org/stable/modules/multiclass.html#ovr-classification
[8] https://scikit-learn.org/stable/modules/multiclass.html#ovo-classification

com o uso de truques de *kernel*, que mapeiam os dados para espaços de características de maior dimensionalidade. A interpretação dos resultados é facilitada pela capacidade das SVMs de fornecer uma fronteira de decisão clara, tornando-as amplamente aplicáveis em problemas de classificação (HAN et al., 2012). A flexibilidade das SVMs em lidar com diferentes tipos de dados e a capacidade de generalização para conjuntos de dados complexos as tornam uma escolha popular em várias áreas. No entanto, a sensibilidade à escolha de hiperparâmetros e a exigência de escalonamento adequado dos dados são considerações importantes ao implementar SVMs (HAN et al., 2012).

Support Vector Classifier[9] (SVC) é uma variação de Support Vector Machine usada para classificação. Utilizando uma abordagem de otimização, o SVC busca encontrar um hiperplano no espaço de características que melhor separe as classes de dados. Os "vetores de suporte", ou seja, os pontos de dados mais próximos da superfície de decisão, são fundamentais para determinar o hiperplano de separação ótimo. O SVC pode usar diferentes funções de *kernel*, como o linear, polinomial ou RBF, para mapear os dados em um espaço de características de alta dimensão, sendo capaz de lidar com problemas de classificação em que as classes não são linearmente separáveis, o que o torna uma ferramenta poderosa para classificação. A Figura 6 mostra os tipos de *kernel* para SVC.

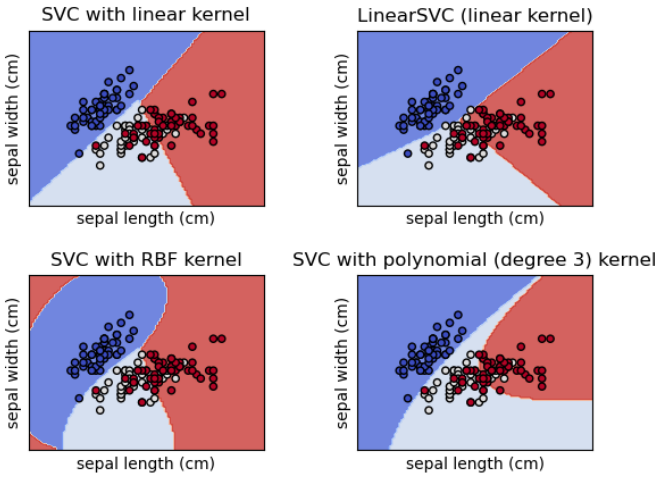

Figura 6 – *Kernels* do método Support Vector Machine. Fonte: Scikit-Learn (2024).

Logistic Regression[10] é uma técnica fundamental em estatística e aprendizado de máquina amplamente utilizada para modelar a probabilidade de ocorrência de um

[9] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[10] https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

evento binário. A função logística transforma uma combinação linear de características explicativas ponderadas por coeficientes em uma probabilidade, garantindo que a saída esteja no intervalo de 0 a 1. Os parâmetros do modelo são estimados por meio de técnicas como a máxima verossimilhança. A interpretação dos coeficientes para a técnica de Logistic Regression é crucial; um aumento em uma unidade nas características está associado a um aumento multiplicativo na *odds* de o evento ocorrer.

Uma das principais vantagens do Logistic Regression é sua capacidade de fornecer probabilidades e ser interpretável, o que permite a tomada de decisões informadas. Além disso, é menos suscetível à multicolinearidade e mais eficiente em termos computacionais em comparação com alguns outros métodos. No entanto, é importante destacar que o Logistic Regression pressupõe uma relação log-linear entre as variáveis independentes e a *log-odds* da variável dependente, o que pode limitar sua capacidade de capturar complexidades em determinados conjuntos de dados. A escolha criteriosa de características e a validação adequada do modelo são essenciais para o sucesso do algoritmo em diversas aplicações.

Ensemble learning[11] é uma estratégia que visa combinar múltiplos modelos com o objetivo de obter previsões mais precisas e robustas do que as que ocorrem com modelos individuais. Nesse contexto, há o interesse em explorar a diversidade dos modelos e agregar seus resultados para mitigar possíveis fraquezas individuais e capturar diferentes nuances dos dados (HAN et al., 2012). O Voting Classifier[12] é uma técnica de *ensemble* para problemas de classificação em que várias previsões de modelos base são combinadas por meio de votação. O Voting Classifier é útil para aproveitar as vantagens de diferentes modelos e para melhorar o desempenho geral do classificador. Essa abordagem não apenas aumenta a precisão das previsões, mas contribui para a estabilidade do modelo, tornando-o mais confiável em uma variedade de cenários de aplicação.

Existem duas estratégias comuns de votação: a votação majoritária (*hard*), na qual a classe com o maior número de votos é escolhida como previsão final, e a votação ponderada (*soft*), em que os modelos têm pesos associados e suas previsões são combinadas de acordo com esses pesos. Optar pela votação *soft* pode ser preferível em várias situações. Primeiramente, a votação *soft* considera não apenas as previsões de classe, mas a confiança associada a essas previsões. Isso é particularmente útil quando os classificadores base são capazes de estimar probabilidades de classe com precisão.

[11] https://scikit-learn.org/stable/modules/ensemble.html
[12] https://scikit-learn.org/stable/modules/ensemble.html#voting-classifier

Além disso, a votação *soft* tende a ser mais eficaz quando os classificadores base são diversos e produzem probabilidades de classe calibradas. Ao considerar a probabilidade de cada classe em vez de apenas a classe mais votada, a votação *soft* pode capturar nuances adicionais nos dados e fornecer uma decisão final mais informada. Isso é especialmente relevante em conjuntos de dados desbalanceados ou quando as classes têm sobreposição substancial em seus espaços de características. Além disso, em muitos casos, a votação *soft* permite interpretar a importância relativa das variáveis na decisão final. Ao ponderar as previsões de cada classificador base pela sua confiança, é possível observar como cada variável contribui para a previsão final, o que pode fornecer *insights* adicionais sobre o comportamento do modelo e os padrões nos dados.

A avaliação de classificadores é uma etapa crucial no desenvolvimento de modelos de aprendizado de máquina. Existem várias métricas disponíveis para avaliar a qualidade do desempenho de um classificador em problemas de classificação binária e multiclasse. Entre as métricas comuns, estão acurácia, precisão, *recall*, *f1-score*, além de representações visuais, como matriz de confusão e gráfico AUC/ROC (HAN et al., 2012). A acurácia é uma medida global que avalia a proporção de instâncias classificadas corretamente em relação ao total (HAN et al., 2012). Por sua vez, a precisão se concentra na capacidade do modelo de identificar corretamente instâncias positivas, calculando a proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos positivos (HAN et al., 2012). O *recall*, também conhecido como sensibilidade, avalia a habilidade do modelo de capturar todas as instâncias positivas, calculando a proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos negativos (HAN et al., 2012). A *f1-score* é uma métrica que combina precisão e *recall*, oferecendo um equilíbrio ponderado entre as duas (HAN et al., 2012).

Do ponto de vista visual, a matriz de confusão fornece uma visão detalhada do desempenho do modelo, podendo exibir as contagens de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos (HAN et al., 2012). Enquanto isso, a curva Receiver Operating Characteristic (ROC), que representa a taxa de verdadeiros positivos em função da taxa de falsos positivos, oferece *insights* sobre a sensibilidade e especificidade do modelo em diferentes configurações de limiar (HAN et al., 2012). A área sob a curva (AUC) em um gráfico ROC é uma métrica que destaca a capacidade do modelo de distinguir entre as classes. Quanto maior a AUC, melhor é a habilidade do modelo de classificação

em diferentes pontos de corte (HAN et al., 2012).

A interpretação de classificadores é crucial para compreender como os modelos tomam decisões e quais variáveis são mais influentes em suas previsões. A interpretação de modelos envolve a análise dos pesos atribuídos a cada variável pelo modelo (HAN et al., 2012). Métodos como a importância de variáveis, coeficientes de modelos lineares e visualizações de decisões de árvores são comumente utilizados. A técnica *SHapley Additive exPlanations* (SHAP) é amplamente empregada nesse contexto.

O SHAP é uma técnica avançada de interpretabilidade que se baseia nos valores Shapley da teoria dos jogos para atribuir níveis de importância a cada variável em uma previsão (LUNDBERG; LEE, 2017). Em essência, o SHAP procura distribuir de forma equitativa a contribuição de cada variável para a diferença entre a previsão do modelo e a média das previsões. Isso permite entender como cada variável contribui para uma previsão específica. Uma vantagem significativa do SHAP é sua capacidade de fornecer explicações individuais e globais. Explicações individuais destacam a contribuição de cada variável para uma única previsão, enquanto explicações globais oferecem uma visão geral do impacto médio de cada variável em todas as previsões (LUNDBERG; LEE, 2017). A Figura 7 mostra um gráfico de importância de variáveis gerado com valores SHAP para uma abordagem de classificação multiclasse[13].

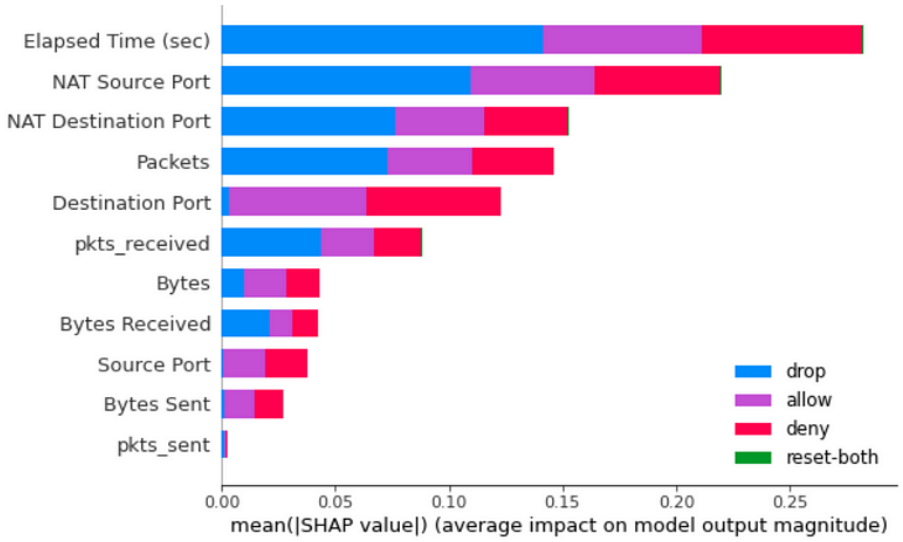

Figura 7 – Gráfico de importância de variáveis com valores Shapley. Fonte: Towards Data Science (2021).

[13] https://towardsdatascience.com/explainable-ai-xai-with-shap-multi-class-classification-problem-64dd30f97cea

3 Trabalhos Relacionados

A evasão estudantil é um desafio significativo enfrentado por instituições de ensino de todo o mundo, com implicações profundas para os estudantes, as próprias instituições e a sociedade em geral. Ao longo dos últimos anos, pesquisadores têm buscado entender e desenvolver estratégias para mitigar esse fenômeno complexo, empregando uma variedade de abordagens e técnicas. A mineração de dados tem desempenhado um papel importante nesse esforço, uma vez que permite a identificação de padrões e fatores de risco visando à criação de modelos com o objetivo de prever a evasão antes de sua ocorrência. Técnicas de mineração de dados e aprendizado de máquina como agrupamento e classificação têm sido amplamente utilizadas, recebendo como entrada conjuntos de dados educacionais, buscando a obtenção de *insights* que baseiem a gestão acadêmica e a implementação de medidas de intervenção. Neste capítulo, são apresentados estudos relacionados ao presente trabalho, tratando de evasão estudantil no ensino superior em instituições do Brasil e do mundo e destacando as diversas abordagens e contribuições para o campo. A Tabela 1 exibe um resumo dos trabalhos relacionados citados.

O presente trabalho é extremamente relacionado ao estudo de Costa et al. (2023), que aborda o desafio da evasão estudantil em instituições de ensino superior utilizando técnicas de mineração de dados e visando à previsão do fenômeno a partir da aplicação do processo KDD para análise de dados acadêmicos. O método do trabalho de Costa et al. (2023) se divide em duas etapas: a primeira, não supervisionada, emprega o algoritmo K-Means para agrupar cursos semelhantes, aumentando a quantidade de dados disponíveis para análise; e a segunda etapa, supervisionada, utiliza modelos de classificação, a partir do algoritmo Random Forest, para prever a evasão de estudantes.

Os dois trabalhos têm motivação compartilhada, tratando do interesse em aumentar o volume de amostras de estudantes de cursos de graduação para previsão da evasão. A baixa quantidade de amostras de estudantes evadidos pode impossibilitar a execução de um estudo que vise à previsão da evasão em determinados cursos e instituições de ensino. A instituição considerada no estudo de caso de Costa et al. (2023) foi desmembrada de outra recentemente e possui uma baixa quantidade de amostras de estudantes, sendo considerados somente 11 cursos. No estudo de caso executado no presente trabalho, foram observados 46 cursos presenciais de uma instituição de ensino brasileira.

Os atributos utilizados no agrupamento de Costa et al. (2023) foram criados pelos autores, observando, principalmente, as taxas de evasão de estudantes na instituição. O agrupamento efetuado a partir do algoritmo K-Means resultou em 3 grupos, que possuem 6, 3 e 2 cursos. O método do cotovelo foi empregado para definição da quantidade de grupos, bem como o coeficiente de silhueta foi usado para a avaliação do agrupamento. No presente trabalho, há a execução de dois agrupamentos para comparação de resultados, a partir dos algoritmos K-Means e Aglomerativo, com a observação do método do cotovelo, do coeficiente de silhueta, dos índices de Calinski-Harabasz, de Davies-Bouldin e de Dunn e do dendograma para definição da quantidade de grupos e avaliação dos resultados. Os atributos utilizados para a execução dos agrupamentos neste trabalho são oriundos do Censo da Educação Superior.

A etapa de classificação do trabalho de Costa et al. (2023) tratou da construção de modelos com uso de apenas um algoritmo, Random Forest. A modelagem emprega apenas duas classes, *Evasão* e *Não Evasão*, e o conjunto de atributos utilizado é composto por informações socioeconômicas e acadêmicas dos estudantes e de seus cursos. Ocorre que a disponibilidade de informações de estudantes em instituições de ensino superior é bastante variável, a depender da instituição e do curso de graduação tratados, o que pode dificultar a reprodução de um estudo relacionado. Foram construídos modelos para cada período do curso de graduação, do primeiro ao oitavo. Ocorre que o modelo do primeiro período tem dados de todos os evadidos e não evadidos, enquanto os modelos dos períodos seguintes têm uma diminuição significativa na quantidade de evadidos. O número de estudantes evadidos cai à medida que os períodos avançam, de forma que, no oitavo período, há um grande desbalanceamento de dados entre as classes consideradas. Os autores empregaram o coeficiente de correlação de Matthews (Matthews Correlation Coefficient — MCC) como tentativa para aprimorar a avaliação dos classificadores construídos com amostras desbalanceadas por classe. A métrica de *recall* também foi utilizada.

O presente trabalho emprega três algoritmos para a construção dos modelos de classificação, a saber: XGBoost, SVC e Logistic Regression. A modelagem busca considerar momentos distintos do curso de graduação para a previsão da evasão, visando detalhar a análise do fenômeno, o que levou à construção de seis modelos distintos. Além disso, neste trabalho, há a consideração da importância do balanceamento de amostras de estudantes em situações acadêmicas específicas, especialmente para a classe de evasão

tardia — que costuma ter uma baixa quantidade de amostras —, com interesse em reduzir o risco de *overfitting*. O conjunto de atributos usados para classificação possui somente dados acadêmicos relacionados ao rendimento do estudante e a informações gerais do curso e da instituição de ensino, pois a baixa disponibilidade de dados de uma instituição pode ser tanto de volume de amostras quanto de variedade de informações disponíveis e buscando evitar possíveis vieses introduzidos pelo uso de informações socioeconômicas. Para avaliação dos classificadores, utilizam-se as métricas de acurácia, precisão, *recall*, *f1-score* e AUC. Além disso, a técnica de SHAP é usada para interpretação dos modelos.

O trabalho de Costa et al. (2023) fez testes de hipótese para verificar a viabilidade da estratégia de agrupamento para aumento da quantidade de amostras de estudantes, considerando os resultados obtidos com a construção de modelos de cursos de graduação individuais. Ocorre que, se as amostras são limitadas e desbalanceadas em conjunto, a situação se torna ainda mais extrema ao observar os cursos individualmente. A baixa quantidade de amostras de estudantes disponíveis para a construção de modelos individuais no estudo de Costa et al. (2023) pode enviesar a análise, considerando, inclusive, que foi utilizado o método de reamostragem de validação cruzada de 10 grupos, o que torna a quantidade de dados de cada uma das classes em cada grupo ainda menor.

Quanto aos resultados de ambos os trabalhos considerando a abordagem de aumentar a quantidade de amostras de estudantes com agrupamento, nota-se que a proposta pode representar uma solução viável para resolver o problema da baixa quantidade de dados disponíveis para uso na construção de classificadores de evasão. Apesar da intencional não consideração de informações socioeconômicas de estudantes no presente trabalho — uma vez que há a preferência pelo uso de informações acadêmicas —, os resultados obtidos na previsão da evasão indicam a viabilidade do método proposto no presente trabalho. A consideração de atributos que tratam principalmente da trajetória acadêmica do estudante, como sugere este trabalho, pode facilitar a reprodução do estudo.

Alguns dos trabalhos apresentados a seguir tratam de agrupar cursos, instituições de ensino ou outros conceitos relacionados, enquanto os outros estudos citados tratam da previsão da evasão estudantil no ensino superior ou da interpretação de modelos de aprendizado de máquina. A pesquisa de Junior et al. (2013) adotou uma estratégia que visa formular indicadores de desempenho para os cursos de graduação na Universidade Federal do Espírito Santo (UFES) e avaliar sua eficácia para agrupar cursos por meio de

análise estatística de agrupamentos. O método envolveu a concepção de indicadores que incluem a demanda do curso no vestibular, quantidade de estudantes, índice de evasão e envolvimento em pesquisa, entre outras informações. Por meio da análise estatística de agrupamentos, os cursos foram dispostos em quatro grupos distintos, evidenciando suas características heterogêneas. Adicionalmente, a análise discriminante foi empregada para validar os resultados, demonstrando uma alta capacidade de classificação dos cursos. De acordo com os autores, essa abordagem metodológica oferece uma base sólida para o desenvolvimento de estratégias de melhoria e gestão acadêmica na UFES, ressaltando a importância da avaliação contínua do desempenho dos cursos de graduação.

O trabalho de Saraiva et al. (2021) propõe uma abordagem que utiliza técnicas de agrupamento para identificar perfis de estudantes em risco de evasão, com base em informações acadêmicas e socioeconômicas. Um estudo de caso é realizado em um curso técnico de Informática. O método adotado envolve a aplicação do algoritmo K-Means para agrupamento de estudantes, seguido por uma análise descritiva dos grupos formados. Os resultados apresentados pelo estudo indicam que variáveis como renda familiar, turno de estudo, sexo do estudante, faixa etária e grau de instrução da mãe influenciam o desempenho acadêmico do estudante. Houve a identificação de três perfis de estudantes, cada um associado a diferentes indicadores de desempenho acadêmico. De acordo com os autores, a análise descritiva revela que estudantes com melhor desempenho tendem a ter renda familiar mais alta, estudar no período noturno, ser do sexo masculino, ter faixa etária acima de 20 anos e ter mães com ensino médio completo. Por outro lado, estudantes com pior desempenho tendem a ter renda familiar mais baixa, estudar no período vespertino, ser do sexo feminino, ter faixa etária até 19 anos e ter mães com ensino médio incompleto. Os autores acreditam que esses resultados podem orientar intervenções pedagógicas para reduzir os índices de evasão e aumentar o sucesso estudantil. O uso de informações socioeconômicas em um estudo como esse pode levar a discussões sobre a consideração da ética no uso de dados pessoais, principalmente ao tratar de possíveis vieses ocorridos durante a análise.

O estudo realizado por Kirch et al. (2019) analisa os indicadores de desempenho de instituições federais de ensino superior no Brasil, utilizando dados do Tribunal de Contas da União (TCU). O trabalho emprega técnicas como análise de componentes principais (Principal Component Analysis — PCA) buscando reduzir a dimensionalidade

dos dados e análise de agrupamentos para identificar padrões entre as instituições. Os resultados revelam *insights* sobre o desempenho das instituições, destacando diferenças significativas nos indicadores, como custo corrente por aluno, taxa de aluno em tempo integral por professor e funcionário equivalente, grau de participação estudantil e índice de qualificação do corpo docente. Além disso, o estudo identifica grupos de instituições com características semelhantes. Embora reconheça que os indicadores utilizados têm limitações isoladas, o estudo enfatiza sua utilidade para análises comparativas, sugerindo que essas podem representar uma ferramenta valiosa para a avaliação e aprimoramento da gestão universitária no país, proporcionando *insights* para políticas educacionais mais eficazes e promovendo a autoavaliação institucional.

O trabalho de Digiampietri et al. (2016) aborda a evasão estudantil no ensino superior brasileiro, propondo um método baseado em mineração de dados para identificar precocemente estudantes com potencial de desistência. A abordagem se concentra no histórico de desempenho das disciplinas do primeiro ano do curso, dispensando fontes externas de dados. O trabalho usa o algoritmo Rotation Forest com o objetivo de identificar padrões relevantes nos dados e realizar a classificação dos alunos quanto ao risco de evasão. O estudo alcançou índices de acerto superiores a 90% ao prever o desempenho dos alunos do Bacharelado em Sistemas de Informação da Universidade de São Paulo (USP). Os autores afirmam que os resultados destacam a eficácia do método na identificação de padrões relevantes para prever a evasão universitária, fornecendo *insights* para ações direcionadas de orientação aos estudantes em risco e para o planejamento de intervenções futuras. Além disso, os autores sugerem que essa abordagem pode ser utilizada para melhorar a eficiência das instituições de ensino superior na redução da evasão, contribuindo para a formação de profissionais e para a otimização do uso dos recursos educacionais. Bem como o trabalho de Digiampietri et al. (2016), no presente estudo há a preferência pelo uso de informações acadêmicas para a construção de modelos de previsão de evasão.

O trabalho de Bonifro et al. (2020) propõe uma ferramenta para prever a evasão de estudantes no primeiro ano de curso de graduação com o uso de técnicas de aprendizado de máquina. A ferramenta visa estimar o risco de abandono de um curso acadêmico, considerando dados pessoais, histórico acadêmico do ensino médio e créditos obtidos no primeiro ano. Foram realizados experimentos utilizando dados de estudantes de 11 cursos de uma universidade. Dentre os algoritmos utilizados, estão Linear Discriminant Analysis

(LDA), Support Vector Machine (SVM) e Random Forest. Os autores afirmam que a inclusão de características como exigências de aprendizado adicionais e créditos do curso melhora o desempenho do modelo na previsão de evasão. Além disso, os autores dizem que, diante dos resultados obtidos, a ferramenta pode ser útil para intervenções precoces visando à redução da evasão universitária.

O estudo de Niyogisubizo et al. (2022) propõe um *ensemble stacking* baseado em uma combinação de modelos construídos por meio dos algoritmos Random Forest, XGBoost, Gradient Boosting e Feed-forward Neural Networks com o objetivo de prever a evasão de estudantes em cursos de graduação. Utilizando um conjunto de dados coletados de 2016 a 2020 na Universidade Constantino, o Filósofo, em Nitra, Eslováquia, o método proposto demonstrou melhor desempenho do *ensemble* em comparação com os modelos individuais, utilizando métricas de avaliação como precisão e AUC sob as mesmas condições. Os resultados indicam que é possível identificar estudantes em risco de evasão com base em fatores influentes, o que pode permitir intervenções precoces para evitar o abandono e tomar medidas preventivas proativas.

O estudo de Jha et al. (2019) investiga o desempenho de diferentes algoritmos de aprendizado de máquina na previsão da evasão e do resultado final de estudantes em Cursos Online Abertos e Massivos (Massive Open Online Courses — MOOCs). Utilizando o conjunto de dados OULAD da Open University do Reino Unido, o trabalho analisa atributos demográficos, notas de avaliação e interações com o ambiente virtual de aprendizagem (Virtual Learning Environment — VLE). De acordo com os autores, modelos baseados na interação do estudante com o VLE alcançaram alta performance, com AUC de até 0,91 para previsão de abandono e 0,93 para previsão de resultados acadêmicos, especialmente no caso do Gradient Boosting Machine. Os autores dizem que considerar informações demográficas e notas de avaliação, juntamente com as interações no VLE, resultou em um pequeno aumento de 0,01 na AUC. O estudo destaca a importância das interações dos estudantes com o VLE na predição de desempenho em MOOCs e sugere trabalhos futuros para melhorar a seleção e engenharia de atributos, incluindo métricas baseadas em tempo.

O estudo de Chaplot et al. (2015) propõe o uso de um algoritmo baseado em redes neurais artificiais para prever a evasão de estudantes em MOOCs a partir da análise de sentimentos dos estudantes. Os pesquisadores utilizam *logs* de cliques e postagens em fóruns do MOOC "Introduction to Psychology" da Coursera, juntamente com atributos como

semana do curso, semana do usuário e número de interações, para extrair características relevantes. A análise de sentimentos é realizada usando SentiWordNet 3.0 e os dados são modelados com uma rede neural artificial. Os autores afirmam que o método proposto supera o estado-da-arte em termos de kappa de Cohen na tarefa de previsão de evasão de estudantes. A importância dos sentimentos dos alunos e a abordagem para lidar com o desbalanceamento de dados são destacadas como contribuições significativas do estudo, com potenciais aplicações em MOOCs e ambientes de aprendizado digital.

O trabalho de Martins et al. (2023) propõe a construção de modelos de previsão de evasão estudantil, com foco na evasão tardia, a partir do uso de técnicas do aprendizado de máquina supervisionado. Utilizando um banco de dados com informações de mais de 30.000 alunos de graduação presencial da Universidade Federal de Juiz de Fora (UFJF) entre 2003 e 2020, foram treinados modelos com o uso de quatro algoritmos: Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression e Random Forest. Após o pré-processamento dos dados, incluindo tratamento de valores ausentes e transformação de variáveis categóricas em numéricas, os modelos foram avaliados em termos de acurácia e *f1-score*. Os resultados mostraram que os algoritmos apresentaram desempenho satisfatório, com acurácia em torno de 90% na base de dados geral de estudantes e resultados ligeiramente melhores para a categoria *Concluído* do que para a categoria *Evadido*. No entanto, em uma base de estudantes específica, o desempenho na categoria *Evadido* foi mais baixo devido à escassez de dados dessa classe, sugerindo a necessidade de considerar cuidadosamente o contexto de aplicação das previsões. Como trabalho futuro, pretende-se analisar a importância das variáveis de entrada para cada modelo proposto, visando entender melhor os fatores que influenciam o desempenho dos algoritmos. O trabalho em questão destaca dificuldades do uso de uma baixa quantidade de amostras de estudantes evadidos de forma tardia para a construção de classificadores, o que está relacionado à motivação do presente trabalho.

A pesquisa de Santos et al. (2021) propõe uma abordagem que utiliza técnicas de classificação visando prever a evasão de estudantes do ensino superior no Brasil. O objetivo é identificar estudantes em risco de evasão para oferecer suporte adequado. O sistema desenvolvido gera diferentes modelos preditivos baseados no desempenho dos alunos em disciplinas cursadas até o momento da previsão. São criadas bases de treinamento e teste para cada período do curso, utilizando o algoritmo CART para gerar árvores de decisão. Os resultados dos experimentos, utilizando uma base de dados de uma universidade

brasileira, demonstram que os modelos alcançam acurácia entre 79,31% e 98,25%. A análise comparativa com o algoritmo Random Forest também é realizada, mostrando que, embora os modelos construídos com esse algoritmo tenham um desempenho superior em algumas métricas, a simplicidade e interpretabilidade das árvores de decisão favorecem sua escolha para compor o sistema de monitoramento proposto.

O trabalho de Kappel (2020) trata da evasão estudantil nas instituições públicas de ensino superior no Brasil, buscando identificar padrões e características dos estudantes com maior propensão a abandonar os estudos. Utilizando dados publicados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), cinco técnicas de aprendizado de máquina foram aplicadas: Naive Bayes, K-Nearest Neighbors, Decision Tree, Random Forest e redes neurais. Os resultados indicam que os modelos construídos com o algoritmo Random Forest obtiveram o melhor desempenho, alcançando uma taxa de acerto de cerca de 80% na previsão da evasão, destacando a idade, participação em atividades extracurriculares e carga horária do curso como atributos determinantes. Os autores dizem que a identificação desses atributos é crucial para o desenvolvimento de estratégias de redução da evasão, com potencial impacto na política educacional. O estudo também ressalta a importância da técnica de validação cruzada para a avaliação estatística dos resultados obtidos pelos modelos de classificação. Por fim, os autores sugerem que trabalhos futuros considerem também o desempenho acadêmico dos estudantes.

O estudo de Carvalho et al. (2019) teve como objetivo identificar atributos que se correlacionam fortemente com o risco de evasão, detectáveis nas duas primeiras semanas do curso, analisando dados de estudantes de cursos de graduação na área de Computação na Universidade Federal do Amazonas (UFAM). Realizada como uma pesquisa exploratória quantitativa que trata da avaliação de três cursos presenciais, a pesquisa utilizou um questionário socioeconômico-demográfico para coletar dados sobre os ingressantes. Foram utilizadas árvores de regressão para identificar os indicadores do questionário associados ao desempenho acadêmico inicial. De acordo com os autores, os resultados revelaram que a participação nas atividades de acolhimento dos calouros, políticas afirmativas, modalidade de ensino médio, conhecimento prévio de linguagem de programação e escolha do curso como primeira opção estão significativamente associados a um bom coeficiente de rendimento no primeiro período letivo. A participação nas atividades de acolhimento foi identificada como o principal indicador associado ao coeficiente de rendimento, seguido

pela experiência prévia em linguagens de programação. O estudo destaca a importância de tais indicadores para a previsão do desempenho acadêmico e para a identificação precoce de estudantes em risco de evasão, fornecendo *insights* para a gestão e aprimoramento dos programas de acolhimento e suporte aos estudantes.

O estudo de Melo e Medeiros (2021) aborda a evasão escolar precoce (EEP) em cursos de graduação na modalidade de educação a distância (EaD). O objetivo é analisar a evasão precoce no curso de Licenciatura em Letras na modalidade EaD do Instituto Federal da Paraíba (IFPB), com foco nos ingressantes no primeiro semestre de 2019. A pesquisa adota uma abordagem qualitativa e quantitativa, abordando pesquisa bibliográfica, documental e estudo de caso. Os resultados mostram que a EEP é significativa, com cerca de dois terços dos alunos evadindo no início do curso. O trabalho destaca variáveis como idade, estado civil, etnia e local de residência como indicadores de vulnerabilidade para a evasão. A correlação entre ano de nascimento e ano de conclusão do ensino médio é forte, indicando a importância dessas variáveis. Os autores dizem que as estratégias para combater a evasão incluem definir políticas de retenção, analisar as probabilidades individuais de evasão de cada estudante desde o seu ingresso e definir práticas pedagógicas mais inclusivas e colaborativas. Essas informações podem orientar os gestores e professores na implementação de medidas eficazes para promover a retenção e o sucesso dos estudantes.

O trabalho de Lundberg et al. (2020) aborda a explicabilidade de modelos de aprendizado de máquina com foco em valores Shapley, propondo uma forma de atribuição de importância para cada variável ou característica em uma previsão. O estudo trata de modelos baseados em árvores, como *random forests*, *decision trees* e *gradient trees*. Os autores dizem que, embora esses modelos sejam amplamente utilizados devido à sua precisão, sua interpretabilidade é um desafio. Os autores introduzem três principais contribuições para melhorias: (1) um algoritmo de tempo polinomial para calcular explicações ótimas baseadas na teoria dos jogos; (2) um novo tipo de explicação que mede diretamente os efeitos de interação local das características; e (3) um conjunto de ferramentas para entender a estrutura global do modelo, combinando muitas explicações locais. Essas ferramentas foram aplicadas a três problemas médicos, demonstrando como explicações locais de alta qualidade podem ser combinadas para representar a estrutura global do modelo, mantendo a fidelidade local. As ferramentas permitiram identificar fatores de risco de mortalidade raros e significativos, destacar subgrupos populacionais com características

de risco compartilhadas, identificar efeitos de interação não lineares entre fatores de risco para doença renal crônica e monitorar o desempenho de um modelo em um hospital, percebendo características que diminuem seu desempenho ao longo do tempo.

O estudo de Messalas et al. (2019) discute a importância da interpretabilidade dos modelos de aprendizado de máquina, especialmente os complexos, como *ensembles* ou redes neurais profundas, para garantir confiança e transparência nas decisões automatizadas. Os valores Shapley, que são baseados na teoria dos jogos, oferecem explicações precisas atribuindo uma importância a cada variável ou característica, mas são computacionalmente caros, exceto em modelos baseados em árvores de decisão. Modelos substitutos, que buscam repetir o comportamento de modelos complexos, são mais fáceis de interpretar, porém produzem explicações aproximadas e nem sempre confiáveis. O método proposto pelos autores combina essas abordagens, usando modelos substitutos treinados com XGBoost e extraindo valores Shapley via Tree SHAP com o objetivo de alcançar alta fidelidade e interpretabilidade. No trabalho, foi introduzida a métrica TopjSimilarity para avaliar a similaridade das explicações entre o modelo original e o substituto. De acordo com os autores, resultados experimentais mostram que a eficácia da abordagem depende da estrutura dos dados e da complexidade do modelo. O trabalho sugere melhorias futuras na construção de modelos substitutos para garantir maior fidelidade e interpretabilidade.

Tabela 1 – Trabalhos relacionados. Fonte: o autor (2024).

Trabalho	**Ano**	**Objetivo**	**Método**	**Resultados**
Costa et al. (2023)	2023	Prever evasão estudantil em instituições de ensino superior	Agrupamento; Classificação; K-Means; Random Forest	Agrupamento aumentou dados disponíveis; solução viável para baixa quantidade de dados
Junior et al. (2013)	2013	Formular indicadores de desempenho e avaliar sua eficácia para categorizar cursos na UFES	Agrupamento	4 grupos de cursos; validação por análise discriminante
Saraiva et al. (2021)	2021	Identificar perfis de estudantes em risco de evasão	Agrupamento	3 perfis de estudantes; variáveis influenciam desempenho; orientações pedagógicas

Continua na próxima página...

Tabela 1 – Trabalhos relacionados. Fonte: o autor (2024).

Trabalho	**Ano**	**Objetivo**	**Método**	**Resultados**
Kirch et al. (2019)	2019	Analisar indicadores de desempenho de instituições federais de ensino superior no Brasil	Agrupamento; PCA	Identificação de grupos de instituições; diferenças significativas em indicadores de desempenho
Digiampietri et al. (2016)	2016	Identificar precocemente estudantes com potencial de evasão	Classificação; Rotation Forest	Acerto superior a 90%; eficácia na previsão da evasão; orientações para intervenções
Bonifro et al. (2020)	2020	Prever a evasão de estudantes no primeiro ano de graduação	Classificação; LDA, SVM, Random Forest	Inclusão de características melhora previsão; ferramenta útil para intervenções
Niyogisubizo et al. (2022)	2022	Prever a evasão de estudantes em cursos de graduação	Classificação; *Ensemble stacking*	Melhor desempenho comparado a modelos individuais; identificação de estudantes em risco
Jha et al. (2019)	2019	Prever a evasão e os resultados finais em MOOCs	Classificação; Gradient Boosting Machine	Alta performance; AUC até 0,91 para evasão; importância das interações no ambiente virtual
Chaplot et al. (2015)	2015	Prever a evasão em MOOCs	Classificação; Redes neurais; análise de sentimento	Importância do sentimento e abordagem de dados desbalanceados
Martins et al. (2023)	2023	Prever a evasão tardia em cursos de graduação	Classificação; Decision Tree, KNN, Logistic Regression, Random Forest	Acurácia em torno de 90%;
Santos et al. (2021)	2021	Prever a evasão em cursos de graduação	Classificação; CART	Acurácia entre 79,31% e 98,25%; Random Forest superior em algumas métricas
Kappel (2020)	2020	Identificar padrões de evasão em instituições públicas de ensino superior no Brasil	Classificação; Naive Bayes, KNN, Decision Tree, Random Forest, redes neurais	Melhor desempenho com Random Forest; idade e carga horária como atributos determinantes

Continua na próxima página...

Tabela 1 – Trabalhos relacionados. Fonte: o autor (2024).

Trabalho	**Ano**	**Objetivo**	**Método**	**Resultados**
Carvalho et al. (2019)	2019	Identificar atributos correlacionados ao risco de evasão em cursos de TI	Classificação; Árvores de regressão	Acolhimento e experiência em programação associados a bom desempenho
Melo e Medeiros (2021)	2021	Analisar evasão precoce em cursos EaD	Misto (Qualitativo e Quantitativo)	Evasão precoce significativa; idade, estado civil, etnia e residência como indicadores de risco
Lundberg et al. (2020)	2020	Melhorar a explicabilidade de modelos de aprendizado de máquina	Valores de Shapley; PCA; Agrupamento	Melhor explicabilidade em modelos de árvores; identificação de fatores de risco em problemas médicos
Messalas et al. (2019)	2019	Garantir confiança e transparência em modelos complexos	Valores de Shapley; Modelos substitutos	Alta fidelidade e interpretabilidade; métrica TopjSimilarity avaliada
Este trabalho	2024	Agrupar cursos semelhantes, prever a evasão precoce e tardia e identificar fatores relevantes para evasão	K-Means, Aglomerativo; XGBoost, SVC, Logistic Regression e Voting Classifier; SHAP	Grupos de cursos e classificadores de evasão precoce e tardia

3.1 Considerações Finais

Os trabalhos relacionados que tratam de prever a evasão estudantil por meio da construção de modelos de classificação, com o uso de técnicas do aprendizado de máquina supervisionado, costumam observar uma das seguintes estratégias de seleção de amostras de estudantes: ou consideram dados de estudantes de apenas um curso, ou utilizam dados de estudantes de um conjunto de cursos selecionados arbitrariamente, ou empregam dados de estudantes de todos os cursos no âmbito de uma instituição de ensino.

Ocorre que, quando tratam de um único curso, os trabalhos costumam promover dificuldade de reprodução em instituições que não dispõem de volumes razoáveis de amostras de estudantes evadidos. Limitar a análise a um único curso pode dificultar a

construção de modelos de previsão de evasão devido à falta de amostras representativas e ao viés de amostragem, comprometendo a generalização dos resultados. Em último caso, a baixa quantidade de amostras nas classes observadas pode inviabilizar a construção dos modelos. Por outro lado, selecionar um conjunto arbitrário de cursos pode introduzir vieses na escolha dos cursos e limitar a representatividade da amostra, dificultando a interpretação dos resultados. Por fim, quando os trabalhos consideram todos os cursos de uma instituição ao mesmo tempo, os classificadores podem não ser capazes de compreender determinados padrões e podem não considerar as especificidades de todos os cursos.

O presente trabalho usa técnicas de mineração de dados, análise de agrupamentos e modelagem preditiva visando lidar com o desafio da evasão estudantil no ensino superior. O método propõe a execução de agrupamentos não arbitrários e não supervisionados de cursos de graduação para o aumento do volume de amostras de estudantes evadidos no âmbito de uma instituição de ensino. Em seguida, há o uso de dados de estudantes para o treinamento de modelos de classificação com o objetivo de prever a evasão em um grupo de cursos. Com a abordagem de agrupar cursos semelhantes e construir modelos de previsão de evasão, espera-se contribuir para o desenvolvimento de políticas institucionais personalizadas que propiciem uma gestão proativa para mitigar o abandono dos cursos.

A interpretabilidade dos modelos de aprendizado de máquina é um aspecto crucial para garantir a confiança e compreensão nas decisões, especialmente em modelos complexos como *ensembles*. A técnica *SHapley Additive exPlanations* (SHAP) se destaca como uma ferramenta valiosa para atribuir importâncias às características das previsões, fornecendo *insights* sobre as decisões do modelo. A melhoria direta na interpretação dos valores Shapley e a combinação de modelos substitutos com SHAP são abordagens promissoras para equilibrar precisão e interpretabilidade. A aplicação de SHAP pode ter implicações significativas em diversas áreas, quando se busca a transparência das decisões dos modelos.

Percebe-se que a abordagem proposta pelo presente trabalho está alinhada ao que vem sendo estudado por pesquisadores de todo o mundo, tratando de um método alternativo para a seleção de amostras de estudantes com o objetivo de construir classificadores para prever a evasão precoce e a evasão tardia. Acredita-se que essa abordagem tem o potencial de fornecer *insights* para gestores educacionais, permitindo o desenvolvimento de estratégias personalizadas de retenção de estudantes para aprimorar os índices de sucesso acadêmico.

4 Método e Ferramentas

Este capítulo apresenta o método desenvolvido e as ferramentas empregadas visando responder às questões de pesquisa que orientam este trabalho. O método é constituído por duas partes, como exibe a Figura 8, sendo cada parte relacionada a uma questão de pesquisa. A primeira parte, abordada na Seção 4.1, trata da execução de agrupamentos de cursos de graduação com o objetivo de gerar, de forma não supervisionada, grupos coesos de cursos com perfis semelhantes para o aumento do volume de amostras de estudantes evadidos. Os dados utilizados na etapa de agrupamento são proveninentes do Censo da Educação Superior de 2021. Por sua vez, a segunda parte, que é discutida na Seção 4.2, trata de construir, para um grupo de cursos obtido após os agrupamentos, modelos de classificação binária e multiclasse para a previsão da evasão em momentos distintos do curso de graduação. Os dados usados na etapa de classificação são oriundos de bases de dados internas da instituição de ensino analisada, a Universidade Federal Rural de Pernambuco (UFRPE), universidade pública federal brasileira com sede em Recife, Pernambuco.

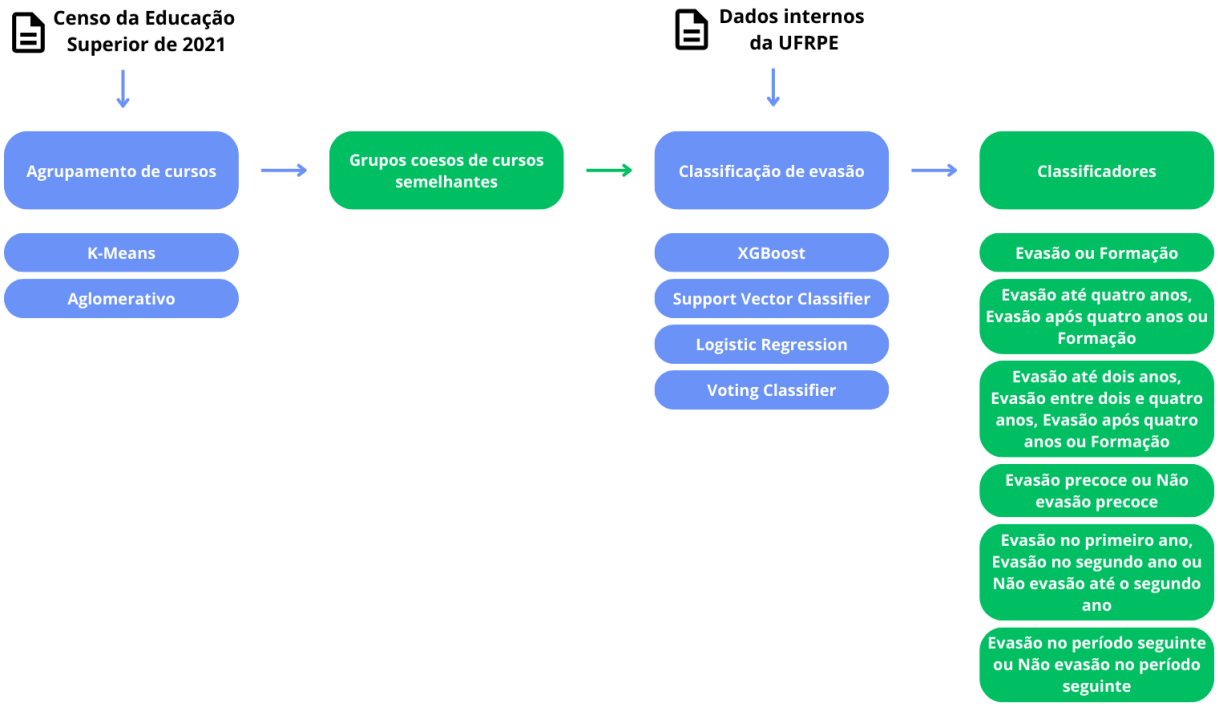

Figura 8 – Visão geral do método. Fonte: o autor (2024).

Cada uma das partes deste trabalho corresponde a um processo de mineração de dados distinto, compreendendo as etapas a serem seguidas e as técnicas a serem utilizadas

para o atingimento dos objetivos específicos do estudo. O método proposto para a execução de cada parte emprega técnicas previstas pelos processos *Knowledge Discovery in Databases* e *Cross Industry Standard Process for Data Mining*.

4.1 Agrupamento de Cursos de Graduação

Para responder à Questão de Pesquisa 1 (QP1), este trabalho propõe a realização de agrupamentos de cursos de graduação de uma instituição de ensino com o interesse em obter grupos coesos de cursos com características semelhantes, o que pode permitir, a partir da consideração de determinados cursos em conjunto, o aumento do volume de amostras de estudantes em situações acadêmicas específicas, como a evasão nos dois primeiros anos do curso (evasão precoce) e a evasão nos anos finais (evasão tardia).

A motivação da execução de agrupamentos está baseada no interesse em identificar e selecionar de forma não supervisionada cursos de graduação de características semelhantes cujos dados de seus estudantes serão empregados para a construção de modelos de previsão da evasão durante a segunda parte do trabalho. As etapas que compõem o processo de agrupamento estão expostas na Figura 9 e são apresentadas a seguir.

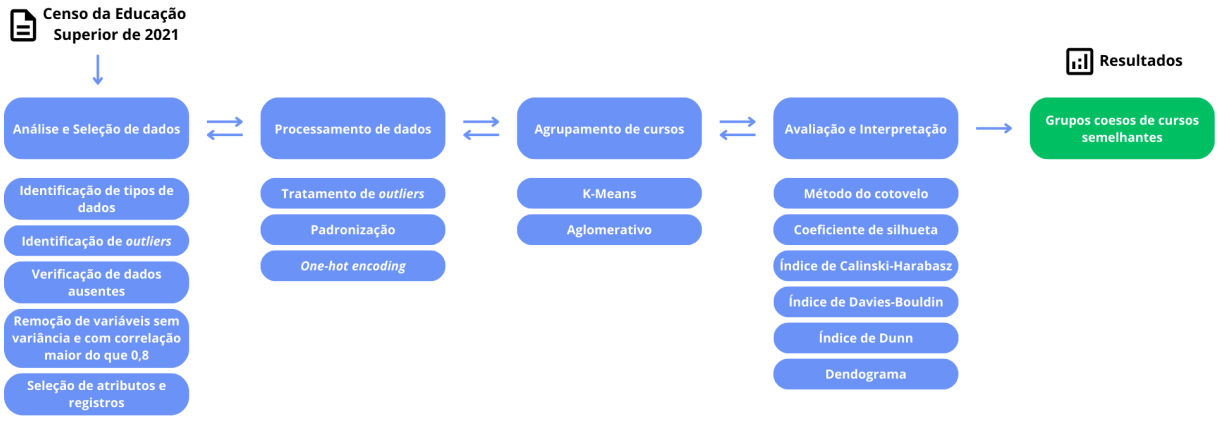

Figura 9 – Visão geral da etapa de agrupamento. Fonte: o autor (2024).

O presente trabalho visa à realização de agrupamentos de cursos de graduação de uma universidade pública brasileira utilizando dados do Censo da Educação Superior de 2021, que é uma base de dados aberta. Essa base contém informações sumarizadas sobre diversos fatores educacionais, incluindo quantidades de vagas, ingressantes, matrículas e concluintes. O uso de dados abertos é incentivado, pois permite a reprodução deste estudo para quaisquer instituições incluídas na base, garantindo transparência e possibilitando

comparações de resultados obtidos ao considerar instituições distintas. O objetivo da análise é gerar grupos de cursos de graduação com perfis semelhantes, facilitando a identificação de padrões e tendências. No entanto, é importante salientar que o Censo da Educação Superior não inclui informações detalhadas sobre o perfil ou a grade curricular dos cursos. Apesar dessa limitação, os dados quantitativos disponíveis são fundamentais, pois fornecem uma visão abrangente e objetiva sobre a distribuição e o comportamento das vagas, ingressos, matrículas e concluintes, elementos essenciais para a análise dos cursos.

4.1.1 Análise e Seleção de Dados

Após a coleta dos conjuntos de dados a serem considerados para a execução de testes de agrupamento, os dados devem ser analisados para que as suas características e peculiaridades sejam conhecidas. Por meio da análise, pode-se identificar problemas a serem resolvidos, bem como pode-se perceber a necessidade de processar ou transformar variáveis (também denominadas atributos ou características) e registros, visando ao uso dos dados como entrada para algoritmos de agrupamento. Além disso, os próprios objetivos do estudo podem ser alterados ou ajustados a partir do conhecimento obtido nas análises. Diversos métodos disponíveis nas bibliotecas de Python[1] Pandas[2] e YData Profiling[3] podem ser empregados durante as análises de dados.

Entre as tarefas a serem realizadas durante as análises, destacam-se a observação de quantidades de variáveis e registros e a identificação de tipos de dados, o que possibilita o levantamento de variáveis qualitativas e quantitativas, além da identificação de dados ausentes, de existência de *outliers*, de diferenças de escalas em variáveis quantitativas e de outras inconsistências nos dados. Após a identificação de possíveis problemas, deve haver a concepção de processamentos que os resolvam. Além disso, com a identificação de variáveis qualitativas e quantitativas, surge a necessidade de processar os dados para a adequação da estrutura de determinadas variáveis, visando ao seu uso para a mineração. As tarefas em questão devem ser executadas durante a etapa de processamento de dados, que é tratada em seção posterior deste capítulo.

Uma vez que as análises sejam feitas, as bases de dados sejam compreendidas e

[1] https://www.python.org/
[2] https://pandas.pydata.org/
[3] https://docs.profiling.ydata.ai/

os possíveis problemas sejam identificados, devem ser selecionados os registros dos cursos de graduação considerados no estudo. Além disso, devem ser selecionadas variáveis de interesse para o agrupamento, com a observação de determinados critérios semânticos e técnicos, considerando métodos de mineração de dados propostos na literatura, que são amplamente utilizados em trabalhos relacionados.

Tratando de critérios semânticos de seleção de variáveis e registros, deve-se observar as particularidades do estudo realizado. Considerando o objetivo de agrupar cursos de graduação presenciais de uma instituição, devem ser selecionados os registros de cursos de interesse, o que implica a remoção de registros de cursos da modalidade EaD. Além disso, devem ser selecionadas variáveis de interesse semântico para o estudo. No estudo de caso a ser proposto neste trabalho, dados do Censo da Educação Superior[4], disponibilizados pelo INEP, são utilizados para a obtenção de informações sobre a universidade pública cujos cursos são objeto deste estudo. Isso acarreta a remoção de variáveis que tratam especificamente de cursos de graduação de instituições de ensino privadas, que também estão presentes na base de dados do Censo.

Quanto aos critérios técnicos de seleção de variáveis, variáveis sem variância de dados devem ser removidas. Para isso, pode ser executada uma seleção automática de variáveis por meio do método Variance Threshold[5], da biblioteca de Python Scikit-Learn[6], que, a partir de um valor de limiar (*threshold*), seleciona variáveis considerando a sua variância. O valor padrão de limiar, zero, deve ser utilizado, de forma a eliminar apenas variáveis sem variância de dados. Também é proposta a remoção de variáveis altamente correlacionadas, com correlação maior do que 0,8, uma vez que variáveis colineares podem atrapalhar a interpretação dos modelos criados. A colinearidade das variáveis pode ser verificada com o uso da biblioteca Pandas.

4.1.2 Processamento de Dados

Os problemas percebidos no conjunto de dados por meio das análises realizadas devem ser solucionados durante a etapa de processamento de dados. Além disso, quaisquer transformações necessárias para a adequação dos dados, considerando tanto os objetivos

[4] https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior
[5] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html
[6] https://scikit-learn.org/

do estudo quanto os requisitos técnicos para o uso de algoritmos de agrupamento, devem ser efetuadas. São diversas as tarefas de processamento que podem ser executadas para possibilitar o uso dos dados como entrada para algoritmos de agrupamento.

O estudo de caso proposto para o agrupamento neste trabalho trata do uso de dados de duas fontes distintas, o Censo da Educação Superior e os Indicadores de Qualidade da Educação Superior[7]. Após a seleção das variáveis e registros de interesse para o agrupamento de cursos, os dados oriundos das distintas bases devem ser unidos em um único conjunto, que será utilizado como entrada dos algoritmos de agrupamento. A junção em questão pode ser feita por meio de métodos disponíveis na biblioteca Pandas.

Entre as principais tarefas de processamento a serem realizadas visando à resolução de problemas comumente percebidos em conjuntos de dados, cita-se a limpeza de dados, incluindo o ato de renomear variáveis e categorias de variáveis qualitativas para seguir as convenções estabelecidas pela literatura, além da imputação de dados para o preenchimento de dados ausentes e da substituição de *outliers* pelo limite inferior ou superior do intervalo interquartil. Uma forma de imputação de dados que costuma ser amplamente utilizada em variáveis quantitativas trata do uso da média dos valores existentes, o que é proposto neste trabalho. Além disso, a substituição de *outliers* é essencial para o uso de algoritmos de agrupamento como o K-Means[8], que é sensível a valores discrepantes. Todas essas tarefas podem ser executadas a partir de métodos implementados pela biblioteca Pandas.

Por sua vez, as tarefas de processamento para a transformação de dados alteram variáveis e registros de acordo com necessidades específicas, observando os objetivos do estudo e os requisitos de uso dos algoritmos de mineração de dados. As tarefas de transformação são realizadas de acordo com o tipo de dados considerado. Como exemplos de tarefas de transformação de dados, há o *one-hot encoding* de variáveis qualitativas e a padronização de variáveis quantitativas.

Variáveis qualitativas normalmente não podem ser empregadas diretamente como entrada de determinados algoritmos de agrupamento como o K-Means. Diante disso, pode ser executado o método *one-hot encoding* com o objetivo de representar variáveis qualitativas numericamente, criando variáveis quantitativas. Uma implementação do método *one-hot encoding* está disponível na biblioteca Scikit-Learn por meio da classe

[7] https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior

[8] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

OneHotEncoder[9].

Por sua vez, a padronização pode ser executada sobre variáveis quantitativas para alterar e aproximar as suas escalas, o que é importante antes do uso de dados numéricos para agrupamento, uma vez que algoritmos costumam ser sensíveis a grandes variações de escala. Um método de padronização que pode ser usado está disponível na biblioteca Scikit-Learn por meio da classe StandardScaler[10].

4.1.3 Agrupamentos de Cursos de Graduação

Considerando os objetivos deste trabalho, após a realização dos processamentos necessários sobre os conjuntos de dados, os dados devem ser empregados para a execução de agrupamentos de cursos de graduação, por meio de métodos do aprendizado de máquina não supervisionado. Pertencendo a categorias de algoritmos de particionamento distintas, o K-Means e o Aglomerativo[11], algoritmos amplamente difundidos e empregados, são usados neste estudo. A inicialização *K-Means++* é utilizada para o K-Means. Quanto ao Aglomerativo, é empregado o método de ligação (*linkage*) de Ward. Ambos os algoritmos estão disponíveis na biblioteca Scikit-Learn. As variáveis utilizadas para a realização dos agrupamentos estão expostas na Tabela 2.

O K-Means é um método particional que categoriza os dados em grupos com o objetivo de minimizar a variância intra-grupo e maximizar a distância inter-grupo. Por sua vez, o Aglomerativo é um algoritmo hierárquico que constrói dendrogramas para representar a proximidade entre os pontos e mescla grupos iterativamente, sendo capaz de identificar estruturas mais complexas nos dados. O K-Means pode ser mais eficiente para conjuntos de dados com grupos bem definidos e formatos mais simples, enquanto o Aglomerativo pode capturar relações mais sutis e hierarquias.

A motivação do uso dos algoritmos em questão para a realização de agrupamentos está baseada no interesse em desenvolver uma abordagem robusta que proporcione uma compreensão profunda da estrutura subjacente dos dados considerados. O uso de diferentes algoritmos permite a validação dos resultados obtidos. Caso ambos os algoritmos convirjam para soluções semelhantes, a confiança nos grupos identificados será fortalecida. Dessa

[9] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
[10] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
[11] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Tabela 2 – Variáveis comuns aos testes de agrupamento, oriundas do Censo da Educação Superior de 2021. Fonte: o autor (2024).

Variável	**Tipo**	**Valores**
area_geral_cine_brasil	Texto	Diversos
grau_academico	Texto	Bacharelado Licenciatura Tecnólogo
enade_faixa	Inteiro	Diversos
cpc_faixa	Inteiro	Diversos
qt_vg_total	Inteiro	Diversos
qt_vg_remanesc	Inteiro	Diversos
qt_inscrito_total	Inteiro	Diversos
qt_insc_vg_remanesc	Inteiro	Diversos
qt_ing_fem	Inteiro	Diversos
qt_ing_masc	Inteiro	Diversos
qt_ing_vestibular	Inteiro	Diversos
qt_ing_vg_remanesc	Inteiro	Diversos
qt_ing_preta	Inteiro	Diversos
qt_ing_amarela	Inteiro	Diversos
qt_ing_indigena	Inteiro	Diversos
qt_ing_cornd	Inteiro	Diversos
qt_conc	Inteiro	Diversos
qt_aluno_deficiente	Inteiro	Diversos
qt_ing_deficiente	Inteiro	Diversos
qt_ing_reserva_vaga	Inteiro	Diversos
qt_conc_reserva_vaga	Inteiro	Diversos
qt_sit_trancada	Inteiro	Diversos
qt_sit_desvinculado	Inteiro	Diversos
qt_conc_procescpublica	Inteiro	Diversos
qt_apoio_social	Inteiro	Diversos
qt_ing_apoio_social	Inteiro	Diversos
qt_conc_apoio_social	Inteiro	Diversos
qt_ativ_extracurricular	Inteiro	Diversos
qt_mob_academica	Inteiro	Diversos

forma, analisar e comparar os resultados obtidos com a utilização desses algoritmos pode fundamentar análises abrangentes no contexto do agrupamento de cursos de graduação.

Para definir a quantidade ideal de grupos a serem gerados em um agrupamento definitivo, são realizados agrupamentos prévios, cujos resultados devem ser avaliados por meio de métricas de avaliação convencionais amplamente difundidas. Para o K-Means, podem ser usados o método do cotovelo, o coeficiente de Silhueta e os índices de Calinski-Harabasz, de Daivid-Bouldin e de Dunn. Para o Aglomerativo, pode ser gerado um dendograma. Todos os recursos de avaliação de agrupamentos citados estão disponíveis nas bibliotecas de Python Scikit-Learn e Yellowbrick[12].

Os grupos obtidos após os agrupamentos devem ser analisados e interpretados. Espera-se que sejam gerados grupos coesos de cursos de graduação de perfis semelhantes, o que pode viabilizar a execução da segunda parte deste trabalho, em que um grupo de cursos será escolhido para selecionar um conjunto de dados de estudantes dos cursos do grupo em questão, o que deve aumentar a quantidade de amostras de estudantes evadidos. Além disso, é esperado que sejam gerados grupos semelhantes nos agrupamentos com K-Means e Aglomerativo, o que pode validar os resultados obtidos em cada método.

4.2 Classificação de Estudantes em Risco de Evasão

Com o objetivo de responder à Questão de Pesquisa 2 (QP2), este trabalho tem o propósito de prever a evasão estudantil no ensino superior, considerando dados de estudantes de cursos de graduação de uma instituição de ensino, por meio da construção de modelos de classificação, com o uso de algoritmos de aprendizado de máquina supervisionado. Seis modelos são propostos neste estudo com o objetivo de prever a evasão ou formação de estudantes, com a observação de momentos distintos da trajetória do discente durante o curso, tratando tanto de evasão precoce quanto de evasão tardia.

O uso de modelos de classificação, com o emprego de técnicas de aprendizado de máquina supervisionado, pode desempenhar um papel fundamental na abordagem da evasão estudantil no ensino superior. A evasão é um desafio significativo enfrentado por instituições educacionais em todo o mundo, e as técnicas de aprendizado de máquina podem ajudar a compreender e lidar com esse problema. Nesse contexto, os classificadores têm a

[12] https://www.scikit-yb.org/

capacidade de identificar padrões complexos nos dados, transcender análises tradicionais e oferecer *insights* sobre as variáveis que contribuem para a evasão.

A motivação deste estudo está baseada no fato de que, por meio de modelos de classificação, é possível realizar previsões sobre a probabilidade de um estudante abandonar os seus estudos. Coordenações de curso e gestores acadêmicos podem usar os modelos para a categorização de estudantes com base em seu risco de evasão ao longo do curso, o que permite o acompanhamento individual dos discentes. Essa capacidade preditiva permite que as instituições de ensino identifiquem estudantes em risco e adotem medidas proativas para oferecer o suporte necessário com o objetivo de mitigar a evasão. Para isso, o processo de mineração de dados pode utilizar uma ampla gama de recursos, considerando o desempenho acadêmico e outros dados disponíveis, buscando proporcionar uma visão abrangente da situação. As fases que constituem a etapa de classificação deste trabalho estão exibidas na Figura 10 e são apresentadas a seguir.

Figura 10 – Visão geral da etapa de classificação. Fonte: o autor (2024).

4.2.1 Análise e Seleção de Dados

No método proposto neste trabalho, os dados empregados para a construção de modelos de classificação caracterizam semestres do histórico acadêmico de estudantes de graduação. No estudo de caso executado, os dados são provenientes de uma universidade pública brasileira e foram coletados a partir de sistemas informatizados da universidade em

questão, após as devidas autorizações. Bem como ocorreu durante os testes de agrupamento, as bases de dados coletadas para o desenvolvimento de classificadores de evasão devem ser analisadas com o objetivo de conhecer os dados e processá-los de acordo com questões semânticas, observando os objetivos do estudo, e critérios técnicos, levando em consideração os requisitos de uso de algoritmos de aprendizado de máquina supervisionado.

Entre as análises a serem realizadas, destacam-se aquelas que visam ao entendimento dos dados coletados, tanto para identificar problemas quanto para definir os processamentos necessários para o devido uso. Diante disso, há análises que são semelhantes às realizadas durante a execução dos testes de agrupamento, que buscam a identificação de variáveis qualitativas e quantitativas, de dados ausentes, da existência de *outliers* e de diferenças nas escalas de variáveis quantitativas. As análises citadas podem ser realizadas com o auxílio de métodos disponíveis nas bibliotecas Pandas e YData Profiling.

Quanto à seleção de variáveis a serem empregadas para a modelagem, este trabalho tem interesse em selecionar um conjunto mínimo de variáveis que tratem da realidade acadêmica dos estudantes, que será empregado na construção dos modelos, considerando, por exemplo, a carga horária cursada no semestre e as quantidades de sucessos e insucessos em disciplinas do semestre, além de informações sobre o curso de graduação, como a área do conhecimento e o turno. É válido ressaltar que, a depender da variável analisada, a presença de *outliers* é justificada pela existência de variações genuínas nos dados, o que pode ocorrer, por exemplo, em variáveis que tratam de carga horária, não devendo haver tratamento nesses casos.

Diante do exposto, este trabalho tem o objetivo de utilizar variáveis acadêmicas do estudante para a previsão da evasão, além de variáveis que trazem informações sobre o curso, intencionalmente desconsiderando informações socioeconômicas do estudante. A motivação de usar apenas informações acadêmicas está baseada no fato de que a reprodução do estudo pode ser facilitada em instituições que não dispõem de uma ampla variedade de informações socioeconômicas de seus estudantes, além do interesse em afastar possíveis vieses causados pelo uso de dados pessoais. A seleção de variáveis pode ser executada por meio da biblioteca de Python Pandas.

Um objetivo fundamental deste trabalho trata da necessidade de considerar, durante a coleta e seleção dos registros de semestres de estudantes, os resultados dos agrupamentos de cursos de graduação previamente executados. Para a construção dos classificadores,

devem ser utilizados dados de estudantes de cursos que constituem um grupo obtido após os agrupamentos. Buscando respeitar a temporalidade dos dados, devem ser selecionados dados de estudantes que ingressaram nos cursos em um período específico. O período em questão será tratado na apresentação do estudo de caso realizado, no Capítulo 6.

Como citado anteriormente, os registros do conjunto que contém dados de estudantes representam semestres do histórico acadêmico de cada estudante. Dessa forma, devem ser ordenados os registros dos estudantes por seu identificador único e duração de vínculo, de maneira a garantir que os registros de um mesmo estudante estejam juntos em blocos e ordenados de forma crescente, considerando o período letivo. As *queries* de seleção de registros, a ordenação citada e todas as outras tarefas compreendidas nesta etapa podem ser executadas com o uso da biblioteca Pandas.

Conforme será tratado na seção Modelagem deste capítulo, este trabalho propõe diversos modelos de classificação, cada um com um objetivo específico. A depender do modelo e da necessidade de balanceamento das classes da variável dependente, a seleção de registros de estudantes considerará regras de busca diferentes, de forma que conjuntos de registros distintos serão empregados para construir cada modelo tratado. As tarefas de seleção de registros para cada modelo serão tratadas em detalhes quando os modelos forem apresentados.

4.2.2 Processamento de Dados

Após a análise de dados, é crucial identificar quaisquer problemas a serem resolvidos. Para isso, são definidos processamentos destinados a limpar e preparar os dados para serem utilizados como entrada em algoritmos de aprendizado de máquina supervisionado. Esses processamentos variam de acordo com o tipo de dados presentes nas variáveis, exigindo abordagens distintas para variáveis qualitativas e quantitativas. Além dos procedimentos de limpeza e preparação, é necessário realizar determinadas transformações nos dados das variáveis existentes, seja para atender a requisitos do estudo, seja para garantir sua adequação considerando critérios técnicos para uso dos algoritmos. Adicionalmente, a partir dos dados coletados, é possível criar novas variáveis de interesse, que podem agregar para a consecução dos objetivos do trabalho.

Tratando das variáveis de interesse para a modelagem, especificamente de variáveis

quantitativas, deve ser feita a padronização dos dados, de forma a preservar a justiça em suas escalas distintas. Este trabalho usa o método de validação cruzada para a construção e avaliação de modelos, como será tratado na seção Modelagem deste capítulo. Como sugere a literatura, a padronização das variáveis quantitativas deve ser feita no contexto de cada iteração da validação cruzada. A padronização pode ser realizada com o emprego da classe StandardScaler da biblioteca Scikit-Learn.

Quanto às variáveis qualitativas nominais empregadas, este trabalho sugere o uso do *one-hot encoding* para convertê-las para o formato numérico, tornando-as adequadas para a construção de modelos de classificação. Como alguns algoritmos de aprendizado de máquina não aceitam dados textuais como entrada, é necessário representar esses dados de forma numérica. Além disso, o *one-hot encoding* preserva a natureza das variáveis categóricas, pois não impõe ordem entre as categorias. O OneHotEncoder da biblioteca Scikit-Learn é uma ferramenta conveniente para realizar essa transformação.

Observando o processamento da variável dependente (*target*) dos modelos, tendo em mente o possível desbalanceamento de dados por classe, deve-se considerar estratégias de balanceamento como o *undersampling* ou o *oversampling*. Enquanto o *undersampling* reduz a quantidade de registros da classe majoritária para equilibrar com a classe minoritária, o *oversampling*, exemplificado pelo método SMOTE[13], gera novos exemplos sintéticos da classe minoritária para igualar as proporções.

Este trabalho propõe o uso do *undersampling* em vez do *oversampling*, tendo em vista que o *undersampling* tende a preservar as características dos dados originais, uma vez que não introduz dados sintéticos, o que é vantajoso, considerando que a qualidade dos dados é crucial para a precisão do modelo. O método RandomUnderSampler[14] da biblioteca Imbalanced-Learn[15] pode ser usado para a execução do balanceamento. É válido ressaltar que, caso a estratégia de agrupamento para o aumento da quantidade de amostras de estudantes evadidos não fosse utilizada neste trabalho, o *undersampling* resultaria em quantidades de amostras muito baixas para alguns dos modelos propostos neste estudo, o que certamente inviabilizaria o seu uso.

Ainda tratando das variáveis de interesse, devem ser criadas as variáveis de período final do discente e a situação final do discente, que devem armazenar, respectivamente, o

[13] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html
[14] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html
[15] https://imbalanced-learn.org/stable/

último período de vínculo do discente no curso e a sua situação final no curso, que pode ser *"Evadido"* ou *"Formado"*. A criação de variáveis é uma tarefa comum a ser executada com o uso de recursos da biblioteca Pandas.

4.2.3 Modelagem

Este trabalho tem o objetivo de prever a evasão estudantil no ensino superior por meio da construção de modelos de classificação. Com a identificação precoce de indivíduos em risco de evasão, pode-se realizar intervenções direcionadas para melhorar a retenção e o sucesso estudantil em instituições de ensino. A utilidade dos modelos está baseada em sua capacidade preditiva, uma vez que podem fornecer informações acionáveis a instituições de ensino para implementação de estratégias preventivas e personalizadas.

Neste trabalho, são apresentados seis cenários de estudo distintos, que compreendem a modelagem de classificadores binários ou multiclasse, com propósitos específicos. Para cada cenário, devem ser construídos classificadores com a utilização do método de validação cruzada estratificada de cinco grupos, empregando os algoritmos XGBoost[16], Support Vector Classifier[17] e Logistic Regression[18], o que permite a avaliação e a comparação dos modelos propostos. Além disso, em cada cenário, após a avaliação com validação cruzada, deve ser construído um *ensemble* a partir dos três algoritmos propostos, usando Voting Classifier[19], com o método de votação *soft*, designando o modelo definitivo do cenário. Para a construção do *ensemble*, deve ser usado o método de reamostragem de dados *hold-out*, com 70% do conjunto de dados sendo utilizado para treino e 30% para teste.

Os algoritmos citados estão disponíveis em bibliotecas de Python. O XGBoost pode ser obtido por meio da biblioteca homônima, enquanto os outros algoritmos têm implementações incluídas na biblioteca Scikit-Learn. Para o uso dos algoritmos, foram utilizados os seus parâmetros-padrão. Para o XGBoost, foi utilizada a função *binary:logistic*. Para o SVC, foi utilizado o *kernel* RBF. Para o Logistic Regression, foi empregada a penalidade *l2*.

Em linhas gerais, um *ensemble* de votação fornece duas abordagens para combinar as previsões dos classificadores base: votação *hard* e votação *soft*. Na votação *hard*, a

[16] https://xgboost.readthedocs.io/
[17] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[18] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[19] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html

classe prevista é determinada pela maioria dos votos entre os classificadores individuais. Por outro lado, na votação *soft*, as previsões são ponderadas pela probabilidade estimada de cada classificador para cada classe e então combinadas para gerar uma previsão final. A votação *soft* é preferível quando se deseja aprimorar a confiança das previsões, lidar com a diversidade nos classificadores base, obter resultados mais robustos e interpretar a importância das variáveis na decisão final, o que é de interesse deste estudo.

Os cenários de classificação propostos são motivados pelo interesse em compreender e antecipar os fatores determinantes que influenciam a trajetória acadêmica dos estudantes. Os modelos de cada cenário têm o objetivo geral de prever a evasão com base em dados que representam semestres do histórico acadêmico de estudantes. Em cada cenário, os modelos têm variáveis dependentes (*targets*) com classes distintas, havendo a consideração de registros de estudantes distintos de acordo com o modelo. A depender do objetivo do cenário, os modelos podem ser binários ou multiclasse, o que designa uma abordagem abrangente e detalhada com interesse em entender diferentes aspectos da evolução acadêmica dos estudantes. Por meio dos cenários, busca-se desenvolver meios de identificar a evasão estudantil em distintos momentos do curso de graduação.

Quanto aos dados utilizados para a construção dos modelos, um conjunto de variáveis comum a todos os cenários é selecionado, contendo as variáveis expostas na Tabela 3. As variáveis resultam dos processamentos e trasformações anteriormente tratados. Sobre a seleção dos registros de semestres do histórico acadêmico de estudantes, há um processo compartilhado entre os cenários, que é realizado com a exclusão dos registros em que a situação do período letivo seja *"Suspenso"*, bem como dos registros de estudantes com ano de ingresso anterior a um ano determinado pelo estudo de caso realizado. Em seguida, o conjunto de dados é refinado para conter apenas aqueles estudantes que apresentam múltiplos registros de semestres, representando o histórico acadêmico, sendo os registros ordenados com base nos identificadores dos discentes e nos períodos letivos.

Além da estratégia de seleção de registros de estudantes compartilhada entre os cenários, cada cenário possui uma lógica própria de construção da variável dependente do modelo. Em cada cenário, o status final do discente, considerando a situação de seu último período letivo e a duração total do vínculo, é utilizado para a criação da variável dependente. A diferença na criação da variável está na situação final do estudante, como será tratado a seguir, com a apresentação dos cenários. Em cada cenário, a técnica de

random undersampling deve ser utilizada para balancear as classes da variável dependente, garantindo uma representação equitativa entre estudantes evadidos e não evadidos. Como tratado anteriormente, a estratégia do uso da validação cruzada com o *undersampling* da variável dependente é fortalecida por meio do aumento da quantidade de amostras de estudantes evadidos. Caso não houvesse a iniciativa de aumentar o conjunto de dados de estudantes evadidos para análise, o método proposto poderia ser inviabilizado, considerando a separação dos grupos na validação cruzada e a subamostragem da variável dependente. Os modelos propostos são apresentados a seguir.

Tabela 3 – Variáveis comuns aos testes de classificação. Fonte: o autor (2024).

Variável	**Tipo**	**Valores**
duracao_vinculo	Inteiro	Diversos
area_cnpq	Texto	Ciências Agrárias Ciências Biológicas Ciências da Saúde Ciências Exatas e da Terra
diferenca_conc_em_inicio_graduacao	Inteiro	Diversas
forma_ingresso	Texto	Diversos
periodo_ingresso	Inteiro	1 2
turno	Texto	Matutino Vespertino Noturno
media_final	Decimal	Diversos
ch_total	Inteiro	Diversos
ch_aprovacoes_acum	Inteiro	Diversos
ch_reprovacoes_acum	Inteiro	Diversos
ch_cancelamentos_acum	Inteiro	Diversos
situacao_final_discente	Texto	Depende do experimento

1. Evasão ou Formação. O primeiro cenário proposto compreende a construção de um modelo de classificação binária que têm o objetivo de prever, a partir da consideração de dados de semestres do histórico acadêmico, a evasão ou a formação de estudantes em qualquer momento do curso de graduação. As classes propostas são *Evasão* e *Formação*.

As vantagens deste modelo incluem sua simplicidade e facilidade de implementação, permitindo maior clareza na interpretação dos resultados. O modelo fornece uma visão geral da evasão e formação ao longo do curso, o que é útil para uma compreensão inicial do problema. Propõe-se o uso do modelo quando se deseja uma visão ampla sobre os padrões de evasão e formação dos estudantes, sendo ideal no começo da análise para identificar a magnitude do problema de evasão. Este modelo é semelhante aos modelos que comumente são apresentados em trabalhos relacionados de previsão de evasão a partir de classificadores.

2. Evasão até quatro anos, Evasão após quatro anos ou Formação. O segundo cenário trata de prever a evasão ou a formação de estudantes em momentos específicos do curso de graduação por meio da construção de classificadores multiclasse. As classes propostas são *Evasão até quatro anos*, *Evasão após quatro anos* e *Formação*. Este cenário é motivado pela necessidade de compreender o fenômeno da evasão de uma maneira mais granular, observando momentos específicos do curso. Esta abordagem oferece uma visão mais detalhada e específica, o que pode permitir a identificação precoce de padrões distintos associados a cada classe. A utilidade desse cenário está na capacidade do modelo de diferenciar as classes propostas, possibilitando intervenções personalizadas e estratégias específicas para mitigar a evasão em diferentes estágios do percurso acadêmico.

3. Evasão até dois anos, Evasão entre dois e quatro anos, Evasão após quatro anos ou Formação. Assim como o segundo cenário, o terceiro cenário busca a previsão da evasão ou formação de estudantes com o uso de modelos de classificação multiclasse, mas utilizando uma maior quantidade de classes. As classes sugeridas são *Evasão até dois anos*, *Evasão entre dois e quatro anos*, *Evasão após quatro anos* e *Formação*. Este cenário surge do interesse em aprofundar a compreensão dos diferentes tipos de evolução acadêmica dos estudantes no ensino superior. O emprego de quatro classes visa possibilitar análises ainda mais refinadas entre os momentos de evasão, o que pode permitir uma identificação mais precisa dos fatores contribuintes em cada classe. A utilidade deste modelo está em sua capacidade de classificar de forma mais específica, o que pode possibilitar intervenções altamente direcionadas, bem como estratégias personalizadas, para enfrentar os desafios específicos associados a cada tipo de evasão. Vale ressaltar que, à medida que se aumenta a quantidade de classes de um modelo, é esperado que se torne mais difícil classificar corretamente. Em outras palavras, há um desafio inerente ao uso de

um maior número de classes para a previsão da evasão ou formação estudantil.

4. Evasão precoce ou Não evasão precoce. O quarto cenário visa à previsão da evasão ou da não evasão de estudantes até o segundo ano do curso de graduação com o uso de modelos de classificação binários. As classes sugeridas são *Evasão precoce* ou *Não evasão precoce*. Sabe-se que a evasão é mais frequente nos dois primeiros anos de curso, designando a evasão precoce (OECD, 2023), o que leva à proposta de construção de classificadores para prever se o estudante evadirá ou não no período em questão. Este cenário é motivado pela necessidade premente de identificar estudantes em risco de abandonar a formação acadêmica de maneira precoce. Com o modelo sugerido, a instituição de ensino pode ter a capacidade de antecipar potenciais casos de evasão precoce, permitindo a implementação de intervenções proativas no início do curso. Diante disso, este modelo deve ser usado nos estágios iniciais do curso, para identificar e intervir em potenciais casos de evasão precoce, sendo ideal para ações imediatas que visam aumentar a retenção nos primeiros anos. Por fim, destaca-se o fato de que pode haver estudantes da classe *Não evasão* com situação acadêmica *"Formado"* ou *"Cursando"* no conjunto de dados do cenário.

5. Evasão no primeiro ano, Evasão no segundo ano ou Não evasão até o segundo ano. Como o quarto cenário, o quinto cenário trata da previsão da evasão ou da não evasão de estudantes até o segundo ano de curso de graduação, mas utilizando modelos multiclasse. As classes propostas são *Evasão no primeiro ano*, *Evasão no segundo ano* e *Não evasão até o segundo ano*. A motivação deste cenário é semelhante à do cenário anterior, estando associada à necessidade de desenvolver estratégias mais segmentadas e personalizadas, permitindo intervenções específicas para cada fase da jornada acadêmica do estudante. Porém, este cenário se destaca por seu objetivo de oferecer uma compreensão ainda mais detalhada dos padrões da evasão precoce. As vantagens deste cenário incluem a segmentação detalhada dos padrões de evasão e a precisão nas intervenções, facilitando ações específicas para cada ano crítico. Este modelo deve ser usado durante os primeiros anos do curso, quando é crucial entender e diferenciar os padrões de evasão entre o primeiro e o segundo ano, sendo ideal para estratégias de retenção focadas nos momentos iniciais do curso. Ao prever a evasão potencial em estágios distintos, como primeiro ou segundo ano, e ao identificar a não evasão, o modelo pode ajudar na implementação de medidas preventivas mais precisas e eficazes.

6. Evasão no período seguinte ou Não evasão no período seguinte. O último

cenário visa à construção de modelos de classificação binários capazes de prever a evasão ou a não evasão do estudante no período seguinte a um determinado período considerado. Neste trabalho, é sugerida a construção de modelos com o uso de dados do primeiro período do curso, para a previsão da evasão no segundo período. As classes propostas são *Evasão* ou *Não evasão*. A motivação deste cenário está baseada na necessidade de abordar a evasão de maneira imediata e proativa. Com um enfoque mais direcionado ao curto prazo, o modelo busca identificar estudantes em risco iminente de evasão, o que pode permitir o monitoramento contínuo e intervenções rápidas e personalizadas. Com um modelo como este, as instituições de ensino podem ter a capacidade de antecipar desafios imediatos enfrentados pelos estudantes no início do curso com vistas a oferecer suporte oportuno. A previsão de evasão no período seguinte capacita as instituições a agir prontamente, melhorando, assim, as chances de retenção e sucesso dos estudantes. Além disso, este modelo pode ser usado após cada período acadêmico, não se restringindo ao início do curso, o que permitiria identificar rapidamente os estudantes em risco de evasão no próximo período, sendo útil para um monitoramento dinâmico e contínuo. O modelo pode contribuir para a promoção de um ambiente de monitoramento ágil e responsivo.

Após o treinamento, a avaliação dos modelos é um aspecto crítico para garantir a sua eficácia e confiabilidade. Diversas métricas são comumente usadas para a avaliação de desempenho, incluindo acurácia, precisão, *recall*, *f1-score* e a área sob a curva ROC (AUC/ROC). Além disso, a interpretação dos modelos é fundamental para compreender como as características dos dados influenciam as previsões. A técnica *SHapley Additive exPlanations* (SHAP) proporciona uma explicação intuitiva e individualizada sobre o impacto de cada variável na decisão do modelo. Durante a construção dos classificadores, a importância das variáveis deve ser observada, de forma a compreender a sua relevância para a classificação. A análise pode ser efetuada por meio da biblioteca de Python SHAP[20].

A interpretação dos modelos com SHAP é valiosa não apenas para compreender as previsões individuais, mas para identificar padrões gerais de influência das variáveis, o que pode permitir aumentar a transparência do processo de classificação e auxiliar tanto a tomada de decisões informadas quanto a identificação de possíveis intervenções para mitigar a evasão. A interpretação dos modelos, quando combinada com as métricas de avaliação convencionais, proporciona uma compreensão abrangente de seu desempenho e

[20] https://shap.readthedocs.io/

uso, permitindo aprimoramentos contínuos e a confiança em suas previsões.

É válido ressaltar que a avaliação e interpretação dos modelos não são apenas etapas finais, mas componentes integrantes de um ciclo contínuo de refinamento. Essas práticas levam a ajustes específicos para otimizar o desempenho, aumentam a compreensão sobre as dinâmicas subjacentes dos dados e fortalecem a confiabilidade do modelo, contribuindo para uma abordagem mais eficaz na mitigação da evasão no ensino superior. A interpretação e a compreensão dos modelos capacitam os *stakeholders*, como educadores e gestores educacionais, a tomar decisões informadas com base nas previsões, o que é fundamental, especialmente no âmbito educacional, em que a confiança nas decisões do modelo é essencial para a implementação bem-sucedida de estratégias de prevenção da evasão.

4.3 Considerações Finais

O método proposto neste trabalho aborda o agrupamento de cursos de graduação e a classificação de semestres do histórico acadêmico de estudantes com o objetivo de prever a evasão estudantil no ensino superior. Um dos principais objetivos é aumentar o volume de dados de estudantes para a construção dos modelos preditivos de evasão. Esse aumento é fundamental, uma vez que possibilita a criação de modelos mais robustos e generalizáveis, especialmente em contextos em que a evasão é um evento relativamente raro e os dados podem ser escassos. Ao agrupar cursos de características semelhantes de maneira não arbitrária e não supervisionada, é possível integrar dados de múltiplos cursos, ampliando significativamente o conjunto de amostras de estudantes e permitindo a identificação de padrões de evasão compartilhados entre os cursos. Dessa forma, o método proposto visa à previsão da evasão para um grupo de cursos de graduação, o que permite o compartilhamento dos modelos entre os cursos considerados na análise.

A escolha dos algoritmos K-Means e Aglomerativo para os agrupamentos baseia-se em sua ampla aceitação e eficácia na identificação de padrões e coesão nos dados, o que potencializa a qualidade dos grupos obtidos. Acredita-se que a execução dos agrupamentos proporciona uma base sólida para a fase subsequente de construção dos modelos de previsão de evasão. Ao utilizar métodos de aprendizado de máquina não supervisionado para agrupar os cursos, é possível lidar de forma eficiente e não arbitrária com a heterogeneidade dos dados, evitando vieses que uma abordagem manual de seleção de cursos poderia introduzir.

Na etapa de classificação, a proposta de seis cenários distintos permite uma análise abrangente da evasão estudantil, abordando tanto a evasão precoce quanto a evasão tardia. O uso dos algoritmos XGBoost, Support Vector Classifier e Logistic Regression designa a aplicação de técnicas avançadas e variadas para a modelagem preditiva. A estratégia de validação cruzada estratificada de cinco grupos é fundamental para avaliar a robustez e a generalização dos modelos, proporcionando uma comparação rigorosa entre eles. O balanceamento das classes das variáveis dependentes visa assegurar que os modelos não sejam tendenciosos em relação às classes majoritárias. A criação de um *ensemble* por meio do Voting Classifier com votação *soft* destaca-se como uma abordagem que combina os pontos fortes de diferentes algoritmos. Esse método potencializa a precisão preditiva e a capacidade de generalização dos modelos, resultando em um modelo final mais robusto e confiável para cada cenário. A reamostragem de dados empregando o método *hold-out*, com a divisão de 70% do conjunto para treino e 30% para teste, visa assegurar uma avaliação imparcial e consistente do desempenho dos modelos.

Em síntese, o método proposto neste trabalho trata da combinação de técnicas de agrupamento e classificação para a construção de modelos de previsão de evasão, capazes de fornecer *insights* para as instituições de ensino. Como tratado ao longo deste capítulo, ao permitir a identificação de estudantes em risco de evasão com antecedência, os modelos preditivos oferecem uma base sólida para a implementação de estratégias proativas de intervenção. Dessa forma, acredita-se que o presente trabalho pode contribuir para a mitigação da evasão estudantil, considerando que os modelos podem ser utilizados para o desenvolvimento de medidas que visem à retenção e ao aumento do sucesso acadêmico.

5 Agrupamento de Cursos de Graduação

Este capítulo apresenta os resultados obtidos após a execução dos agrupamentos realizados no estudo de caso proposto neste trabalho. O estudo considera 46 cursos de graduação presenciais da Universidade Federal Rural de Pernambuco (UFRPE), que são divididos em quatro unidades acadêmicas, a saber: Belo Jardim (UABJ), Cabo de Santo Agostinho (UACSA), Recife (SEDE) e Serra Talhada (UAST). Os dados utilizados são oriundos das bases de dados do Censo da Educação Superior de 2021 e dos Indicadores de Qualidade da Educação Superior de 2020, que são disponibilizadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (INEP, 2021b; INEP, 2021a). As edições das bases de dados em questão eram as mais recentes disponíveis durante a execução deste estudo, o que justifica a sua coleta.

Como propõe o método apresentado, dois testes de agrupamento foram executados. O primeiro teste compreende o uso do algoritmo K-Means, enquanto o segundo teste trata do emprego do algoritmo Aglomerativo. A motivação para o uso de dois algoritmos de agrupamento é desenvolver uma abordagem robusta que proporcione uma compreensão profunda da estrutura dos dados. Usar diferentes algoritmos permite validar os resultados obtidos: se ambos convergirem para soluções semelhantes, a confiança nos grupos obtidos aumenta. Ambos os testes de agrupamento envolvem a realização de agrupamentos de 2 a 10 grupos para comparação dos resultados obtidos para diversas métricas, o que pode fundamentar análises abrangentes no contexto do agrupamento de cursos de graduação.

A Tabela 4 exibe informações estatísticas sobre as variáveis quantitativas usadas nos testes de agrupamento. As figuras 11 e 12 mostram *boxplots* das variáveis em questão. As figuras 13, 14, 15, 16 e 17 exibem as distribuições de frequências dos cursos para as variáveis de Conceito ENADE, Conceito Preliminar de Curso (CPC), área do conhecimento, grau acadêmico e vagas ofertadas por curso. Os gráficos das distribuições de frequências trazem as quantidades de cursos que se enquadram nos valores para as variáveis observadas.

5.1 K-Means

Com o uso do algoritmo K-Means, foram realizados agrupamentos de 2 a 10 grupos visando à comparação dos resultados obtidos para a métrica de inércia, o coeficiente de

silhueta e os índices de Calinski-Harabasz, Davies-Bouldin e Dunn. Os resultados são exibidos na Tabela 5. Ao interpretar os resultados dos agrupamentos, é crucial considerar uma variedade de métricas comumente empregadas para avaliar a coesão, a separação e a interpretabilidade dos grupos, com o objetivo de encontrar um equilíbrio entre a complexidade do método e a capacidade de distinguir padrões significativos nos dados.

Tabela 4 – Variáveis quantitativas utilizadas nos testes de agrupamento. Fonte: o autor (2024).

Variável	**Min.**	**Máx.**	**Média**	**Desv. Pad.**	**Mediana**
qt_vg_total	70	150	108,48	24,03	100
qt_vg_remanesc	0	40	21,74	14,50	20
qt_inscrito_total	0	1567	716,68	418,97	611
qt_insc_vg_remanesc	0	22	5,75	6,93	3
qt_ing_fem	0	84	32,15	19,32	30,50
qt_ing_masc	0	100	42,89	24,76	36
qt_ing_outra_forma	0	3	0,70	1,01	0
qt_ing_preta	0	24	9,13	5,32	7,5
qt_ing_amarela	0	3	0,76	0,81	1
qt_ing_cornd	0	8	1.95	2,19	1
qt_conc	0	33	9,42	8,63	9
qt_ing_deficiente	0	3	0,89	1,02	1
qt_ing_reserva_vaga	0	17	6,67	5,94	4
qt_sit_trancada	0	19	5,96	5,44	4,5
qt_sit_desvinculado	0	111	53,57	27,88	54
qt_ing_procescprivada	0	141	55,15	30,53	58
qt_conc_procescpublica	0	5	1,37	1,85	0
qt_apoio_social	0	60	17,52	15,62	13,5
qt_conc_apoio_social	0	7	1,64	2,41	0
qt_ativ_extracurricular	0	36	11,09	10,33	8,5

A Figura 18 apresenta um gráfico com os valores de inércia obtidos. Na figura, nota-se a existência de uma curva cujo ponto de inflexão indica uma possível quantidade ideal de grupos, designando o método do cotovelo. A análise dos agrupamentos baseou a definição da quantidade ideal de grupos para a realização de um agrupamento K-Means definitivo. É importante observar a métrica de inércia, que representa a soma das distâncias

quadráticas entre os pontos de dados e os centroides de seus grupos. Como é possível notar, à medida que o número de grupos aumenta, a inércia tende a diminuir, o que é esperado, pois há a tendência de que os grupos se ajustem melhor aos dados à medida que a quantidade de grupos aumente. No entanto, é preciso atentar à diminuição da taxa de diferença da inércia (Δ Inércia) à medida que se aumenta o número de grupos, pois isso pode indicar que há uma adição de complexidade desnecessária.

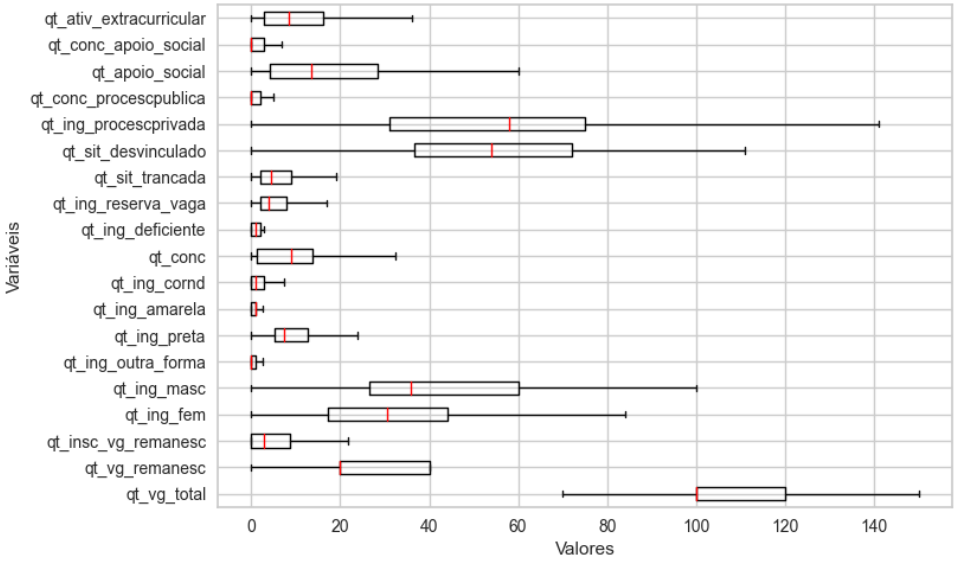

Figura 11 – *Boxplots* das variáveis quantitativas utilizadas na etapa de agrupamento. Fonte: o autor (2024).

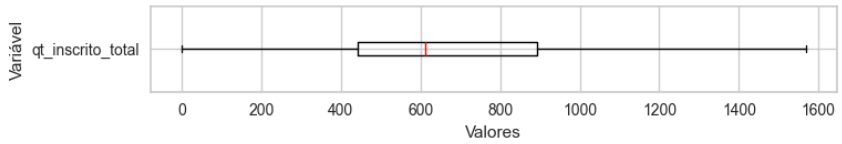

Figura 12 – *Boxplot* da variável *qt_ inscrito_ total* utilizada na etapa de agrupamento. Fonte: o autor (2024).

Outra métrica importante a ser considerada é o coeficiente de silhueta, que fornece uma medida da coesão dentro dos grupos e da separação entre os grupos. Idealmente, deseja-se que o coeficiente de silhueta seja o mais próximo possível de 1, o que indica uma boa separação entre os grupos. Diante dos resultados, observa-se que, à medida que o número de grupos aumenta, o coeficiente de silhueta tende a diminuir ligeiramente, sugerindo que a interpretação dos grupos pode se tornar mais ambígua à medida que são gerados mais grupos.

Conceito ENADE	Contagem	Frequência (%)
SC	15	32.6%
4	15	32.6%
3	9	19.6%
2	4	8.7%
5	3	6.5%

Figura 13 – Frequências do Conceito ENADE. Fonte: o autor (2024).

Conceito Preliminar de Curso	Contagem	Frequência (%)
4	25	54.3%
3	10	21.7%
SC	9	19.6%
5	2	4.3%

Figura 14 – Frequências do Conceito Preliminar de Curso. Fonte: o autor (2024).

Área	Contagem	Frequência (%)
Educação	12	26.1%
Agricultura, silvicultura, pesca e veterinária	11	23.9%
Engenharia, produção e construção	10	21.7%
Ciências sociais, comunicação e informação	4	8.7%
Computação e tecnologias da informação e comunicaç...	3	6.5%
Negócios, administração e direito	2	4.3%
Ciências naturais, matemática e estatística	2	4.3%
Serviços	2	4.3%

Figura 15 – Frequências da área do conhecimento. Fonte: o autor (2024).

Grau acadêmico	Contagem	Frequência (%)
Bacharelado	33	71.7%
Licenciatura	12	26.1%
Tecnológico	1	2.2%

Figura 16 – Frequências do grau acadêmico. Fonte: o autor (2024).

Vagas por ano	Contagem	Frequência (%)
70	5	10.9%
80	5	10.9%
100	15	32.6%
120	12	26.1%
140	5	10.9%
150	4	8.7%

Figura 17 – Frequências de vagas por ano. Fonte: o autor (2024).

Além disso, os índices de Calinski-Harabasz e Davies-Bouldin também fornecem informações sobre a coerência e a separação dos grupos. O índice de Calinski-Harabasz tende a aumentar à medida que são gerados mais grupos, indicando uma maior separação entre eles, enquanto o índice de Davies-Bouldin tende a diminuir, sugerindo uma melhor separação. No entanto, deve-se atentar ao fato de que essas métricas podem ser sensíveis à distribuição dos dados. Por fim, o índice de Dunn avalia a compactação dos grupos em relação à separação entre eles. Valores mais altos indicam uma melhor separação entre os grupos. Nos resultados obtidos, nota-se uma tendência geral de aumento no índice de Dunn à medida que o número de grupos aumentas.

Tabela 5 – Resultados dos agrupamentos de 2 a 10 grupos realizados com K-Means, considerando a inércia, a diferença da inércia, o coeficiente de silhueta e os índices de Calinski-Harabasz, Davies-Bouldin e Dunn. Fonte: o autor (2024).

Grupos	**Inércia**	**Δ Inércia**	**Silhueta**	**Calinski-Harabasz**	**Davies-Bouldin**	**Dunn**
2	817,951785	817,951785	0,185819	11,876792	1,439152	0,736784
3	678,487173	-139,464613	0,189036	11,415716	1,783741	0,682984
4	582,793968	-95,693205	0,217835	10,952811	1,527131	0,701285
5	491,494375	**-91,299593**	**0,244393**	**11,412661**	**1,393688**	**0,983090**
6	439,115038	**-52,379337**	**0,235714**	**10,924228**	**1,368609**	**0,843174**
7	411,560603	-27,554435	**0,230336**	9,905371	1,460253	**0,851281**
8	390,527058	-21,033545	0,181161	9,010555	1,524040	0,769276
9	364,192745	-26,334313	0,185195	8,566280	1,406329	0,782984
10	345,583370	-18,609375	0,178193	8,023022	1,399694	0,791747

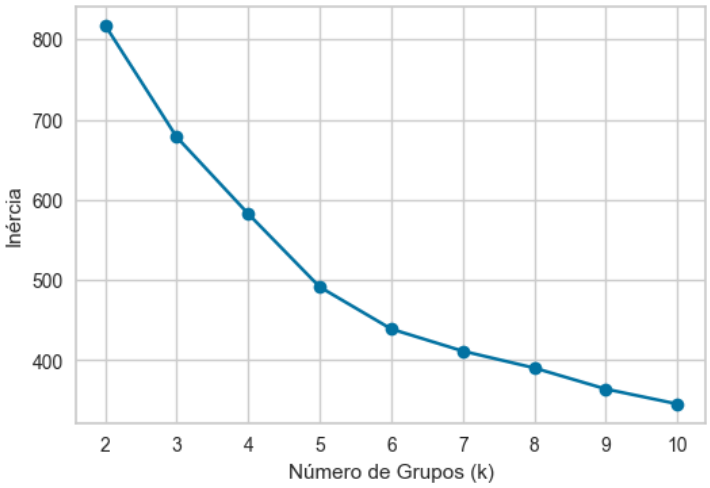

Figura 18 – Método do cotovelo. Fonte: o autor (2024).

Com base na análise das métricas de desempenho consideradas, justifica-se selecionar o valor de k igual a 5 para a definição de um agrupamento definitivo com K-Means, uma vez que esse valor apresenta uma redução significativa na inércia em comparação com $k = 2$, indicando uma melhor adequação dos dados aos grupos. Além disso, o coeficiente de silhueta para $k = 5$ é relativamente alto, sugerindo uma boa separação entre os grupos. O índice de Calinski-Harabasz também é favorável, indicando uma alta dispersão entre os grupos. Por outro lado, o índice de Davies-Bouldin é relativamente baixo, o que sugere uma boa separação e distinção entre os grupos. Dessa forma, o valor de $k = 5$ parece

equilibrar a redução da inércia com a maximização da coesão e separação dos grupos, resultando em grupos possivelmente mais significativos e úteis para análises posteriores. Outros valores possíveis para o k seriam 4 e 6, uma vez que os agrupamentos para $k = 4$ e para $k = 6$ têm resultados próximos aos do agrupamento de $k = 5$. A Figura 19 expõe um gráfico gerado a partir do coeficiente de silhueta para o agrupamento K-Means de $k = 5$.

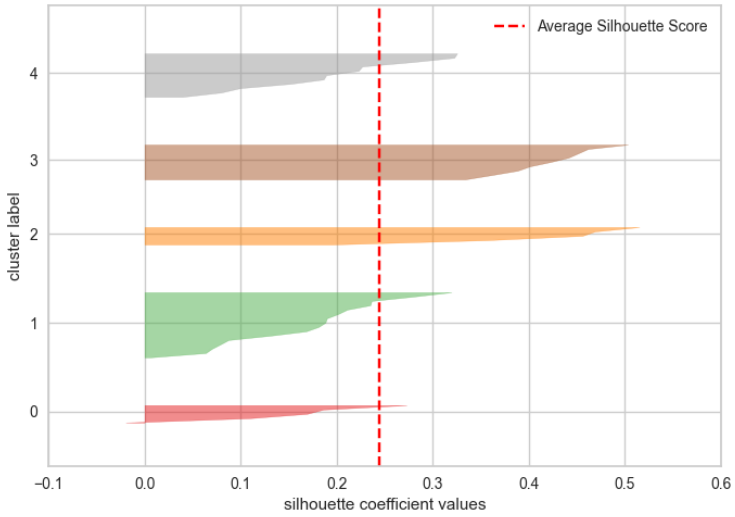

Figura 19 – Gráfico gerado com o coeficiente de silhueta para o agrupamento K-Means de 5 grupos. Fonte: o autor (2024).

Os resultados do agrupamento de 5 grupos com K-Means podem fornecer *insights* valiosos, considerando que os grupos obtidos podem destacar semelhanças entre os cursos com base em diferentes características percebidas nos dados oriundos do Censo da Educação Superior. As tabelas 6, 7, 8, 9 e 10 apresentam, respectivamente, os cursos que compõem os grupos 1, 2, 3, 4 e 5 gerados com o agrupamento K-Means com $k = 5$.

No Grupo 1, percebe-se uma ênfase em cursos relacionados à área de ciências agrárias, com cursos como agronomia, medicina veterinária e biologia. O Grupo 2 abrange uma variedade de cursos, com destaque para ciências biológicas e engenharias, sugerindo uma diversidade de disciplinas científicas e tecnológicas. O Grupo 3 é dominado por cursos de engenharia, com foco em diferentes especializações, como civil, de materiais, eletrônica e elétrica. O Grupo 4 agrupa cursos de engenharia com enfoque em computação, controle e automação, hidráulica e química, além de cursos relacionados à agroecologia e aquicultura. Por fim, o Grupo 5 é composto por cursos relacionados à administração, computação, ciências sociais, educação e gastronomia, indicando uma diversidade de disciplinas de humanas e tecnológicas.

Tabela 6 – Os 16 cursos que compõem o Grupo 1 do agrupamento definitivo com K-Means. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Ciências Biológicas
		Ciências do Consumo
		Engenharia Agrícola e Florestal
		Engenharia de Pesca
		Engenharia Florestal
		Zootecnia
	Licenciatura	Física
UAST	Bacharelado	Administração
		Agronomia
		Ciências Biológicas
		Ciências Econômicas
		Engenharia de Pesca
		Sistema de Informação
		Zootecnia
	Licenciatura	Letras - Português e Inglês
		Química

Tabela 7 – Os cinco cursos que compõem o Grupo 2 do agrupamento definitivo com K-Means. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Agronomia
		Medicina Veterinária
	Licenciatura	Ciências Biológicas
		Matemática
		Química

Os grupos formados no agrupamento K-Means de $k = 5$ parecem capturar certas similaridades entre os cursos em termos de habilidades desenvolvidas pelos estudantes, temas considerados ou abordagens metodológicas empregadas, o que pode ser útil para a continuidade do trabalho.

Tabela 8 – Os cinco cursos que compõem o Grupo 3 do agrupamento definitivo com K-Means. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
UACSA	Bacharelado	Engenharia Civil
		Engenharia de Materiais
		Engenharia Eletrônica
		Engenharia Elétrica
		Engenharia Mecânica

Tabela 9 – Os nove cursos que compõem o Grupo 4 do agrupamento definitivo com K-Means. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Agroecologia
		Economia Doméstica
		Engenharia Ambiental
	Licenciatura	Ciências Agrícolas
	Tecnólogo	Aquicultura
UABJ	Bacharelado	Engenharia de Computação
		Engenharia de Controle e Automação
		Engenharia Hídrica
		Engenharia Química

5.2 Aglomerativo

Tal como ocorreu no primeiro teste de agrupamento realizado, o segundo teste trata da realização de agrupamentos de 2 a 10 grupos, mas com o uso do método Aglomerativo, considerando o objetivo de comparar os resultados obtidos para o coeficiente de silhueta e os índices de Calinski-Harabasz, Davies-Bouldin e Dunn, que são apresentados na Tabela 11. A Figura 20 exibe um gráfico de dendrograma, que é uma representação visual de um agrupamento hierárquico, como o Aglomerativo.

Diferentemente do K-Means, o método de agrupamento Aglomerativo emprega uma abordagem hierárquica que começa tendo cada amostra como um grupo separado e, em seguida, une iterativamente os grupos mais similares até que todos os dados estejam em um único grupo. O dendograma permite uma compreensão intuitiva da estrutura do

agrupamento, revelando a proximidade entre os elementos e como eles são agrupados em diferentes níveis. A interpretação dos resultados do agrupamento hierárquico envolve uma análise cuidadosa das métricas de desempenho em conjunto para garantir a identificação de grupos significativos e úteis. A abordagem hierárquica oferece uma visão detalhada da estrutura dos dados, permitindo a identificação de padrões em diferentes níveis de granularidade. É essencial encontrar o equilíbrio entre a complexidade dos grupos e a necessidade de mantê-los distintos e interpretáveis.

Tabela 10 – Os 11 cursos que compõem o Grupo 5 do agrupamento definitivo com K-Means. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Administração
		Ciência da Computação
		Ciências Econômicas
		Ciências Sociais
		Gastronomia
		Sistema de Informação
	Licenciatura	Computação
		Educação Física
		História
		Letras - Português e Espanhol
		Pedagogia

Como dito anteriormente, o coeficiente de silhueta é uma métrica fundamental que avalia a coesão e a separação dos grupos. Ele fornece uma medida da distância média entre uma amostra ou ponto e os outros pontos no mesmo grupo, em comparação com a distância média entre o ponto e os pontos em outros grupos. Um valor de silhueta próximo de 1 indica que os pontos estão bem agrupados, enquanto valores negativos sugerem que os pontos podem ter sido atribuídos ao grupo errado. Ao examinar os resultados dos agrupamentos, observa-se uma tendência inicial de aumento do coeficiente de silhueta com o aumento do número de grupos, indicando uma melhor separação e coesão dos grupos.

Ao observar o dendrograma, é possível identificar seis ramificações distintas que sugerem a presença de seis grupos bem definidos, o que justifica a definição de $k = 6$ para um agrupamento definitivo com o método Aglomerativo. Cada divisão no dendrograma

representa uma bifurcação em que ocorre a separação dos grupos, indicando que há seis pontos de divisão claros ao longo do processo de agrupamento hierárquico. Além disso, o coeficiente de silhueta apresenta um valor razoável com 6 grupos, enquanto o índice de Calinski-Harabasz, o índice de Davies-Bouldin e o índice de Dunn também demonstram um desempenho consistente para $k = 6$.

Tabela 11 – Resultados dos agrupamentos de 2 a 10 grupos realizados com o método Aglomerativo, considerando o coeficiente de silhueta e os índices de Calinski-Harabasz, Davies-Bouldin e Dunn. Fonte: o autor (2024).

Grupos	**Silhueta**	**Calinski-Harabasz**	**Davies-Bouldin**	**Dunn**
2	0,197533	11,741929	1,271981	0,406658
3	0,181909	11,101357	1,838512	0,387871
4	0,209901	10,804716	1,560183	0,504084
5	**0,237440**	**11,055284**	**1,396174**	**0,552781**
6	**0,215787**	**10,505056**	**1,464867**	**0,542596**
7	**0,215642**	9,604535	**1,492474**	**0,542596**
8	**0,214593**	8,989599	**1,446882**	**0,592002**
9	0,203552	8,464518	1,343649	0,577640
10	0,183873	8,108693	1,339840	0,474444

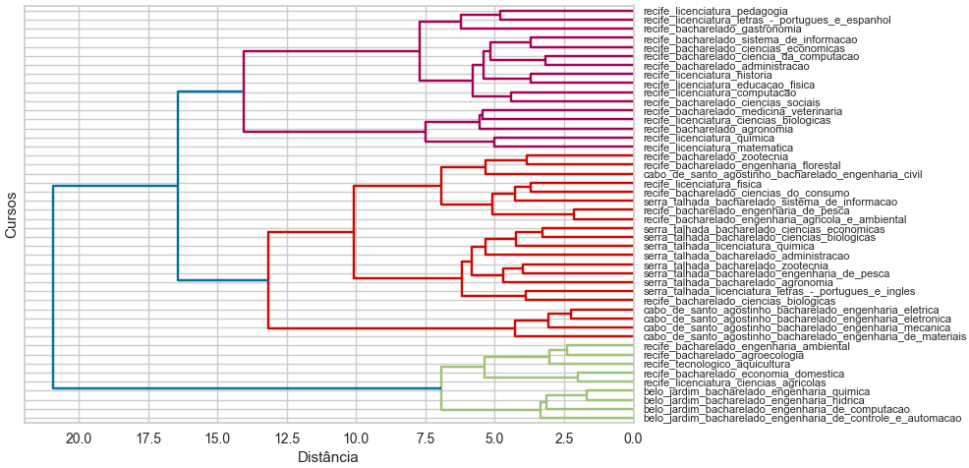

Figura 20 – Dendograma do agrupamento Aglomerativo. Fonte: o autor (2024).

Embora haja variações nos valores dessas métricas conforme o número de grupos aumenta, a escolha de 6 grupos parece ser uma opção que equilibra a coesão dentro dos grupos com a separação dos grupos, resultando em uma estrutura de agrupamento

robusta e bem definida. Portanto, a análise abrangente das métricas de avaliação de agrupamento, juntamente com a evidência visual fornecida pelo dendrograma, motiva a escolha de 6 grupos como uma opção adequada. As tabelas 12, 13, 14, 15, 16 e 17 apresentam, respectivamente, os cursos que compõem os grupos 1, 2, 3, 4, 5 e 6, gerados com o agrupamento Aglomerativo definitivo.

No Grupo 1, há uma concentração de cursos relacionados à área de ciências sociais e de tecnologia, como administração, ciência da computação, economia e sistemas de informação. O Grupo 2 abrange principalmente cursos das áreas de agronomia, biologia e química, indicando um foco em disciplinas das ciências naturais e da saúde. O Grupo 3 tem em sua maioria cursos de engenharia, com especializações em áreas como civil, agrícola, de pesca e florestal. No Grupo 4, encontram-se cursos de engenharia com foco em computação, controle e automação, química e ambiental, sugerindo uma abordagem integrada das ciências exatas e aplicadas. O Grupo 5 é composto exclusivamente por cursos de engenharia, incluindo materiais, eletrônica, elétrica e mecânica. Por fim, o Grupo 6 engloba uma variedade de cursos, desde biologia e agronomia até letras e química, indicando uma abordagem mais ampla e diversificada em relação às disciplinas oferecidas pela universidade.

Tabela 12 – Os 11 cursos que compõem o Grupo 1 do agrupamento definitivo com o método Aglomerativo. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Administração
		Ciência da Computação
		Ciências Econômicas
		Ciências Sociais
		Gastronomia
		Sistema de Informação
	Licenciatura	Computação
		Educação Física
		História
		Letras - Português e Espanhol
		Pedagogia

Tabela 13 – Os cinco cursos que compõem o Grupo 2 do agrupamento definitivo com o método Aglomerativo. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Agronomia
		Medicina Veterinária
	Licenciatura	Ciências Biológicas
		Matemática
		Química

Diante dos resultados obtidos, pode-se afirmar que, no contexto da universidade em questão, o agrupamento Aglomerativo de 6 grupos corresponde a uma abordagem útil para formar grupos de cursos de perfis semelhantes, pois houve a identificação de padrões e afinidades entre os cursos com base em suas características e disciplinas.

Tabela 14 – Os oito cursos que compõem o Grupo 3 do agrupamento definitivo com o método Aglomerativo. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Ciências do Consumo
		Engenharia Agrícola e Ambiental
		Engenharia de Pesca
		Engenharia Florestal
		Zootecnia
	Licenciatura	Física
UAST	Bacharelado	Sistema de Informação
UACSA	Bacharelado	Engenharia Civil

5.3 Discussão dos Resultados

Antes da discussão dos resultados de agrupamento, deve-se considerar a Questão de Pesquisa 1 (QP1): *"É possível gerar, de forma não arbitrária e não supervisionada, grupos coesos de cursos de graduação (que possuem características possivelmente semelhantes) a serem considerados para a previsão da evasão?"*.

Ao comparar os resultados dos agrupamentos executados a partir dos métodos K-Means e Aglomerativo, percebem-se semelhanças e diferenças entre os grupos obtidos.

Tabela 15 – Os nove cursos que compõem o Grupo 4 do agrupamento definitivo com o método Aglomerativo. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Agroecologia
		Economia Doméstica
		Engenharia Ambiental
	Licenciatura	Ciências Agrícolas
	Tecnólogo	Aquicultura
UABJ	Bacharelado	Engenharia de Computação
		Engenharia de Controle e Automação
		Engenharia Hídrica
		Engenharia Química

Tabela 16 – Os quatro cursos que compõem o Grupo 5 do agrupamento definitivo com o método Aglomerativo. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
UACSA	Bacharelado	Engenharia de Materiais
		Engenharia Eletrônica
		Engenharia Elétrica
		Engenharia Mecânica

Conceitualmente, há variações significativas na forma como os dois métodos formam os seus grupos. Uma diferença notável está na maneira como os métodos lidam com a similaridade entre os cursos e a estruturação dos grupos. Enquanto o K-Means usa a inércia como métrica para determinar a separação dos grupos, o Aglomerativo utiliza uma abordagem hierárquica, mesclando gradualmente os cursos com base em sua proximidade.

Considerando os resultados obtidos com o K-Means, observa-se uma tendência de redução gradual na inércia à medida que o número de grupos aumenta, o que é esperado, já que o algoritmo busca minimizar a soma dos quadrados das distâncias entre os pontos e os centroides dos grupos. No entanto, há uma variação na taxa de diminuição da inércia, com um decréscimo mais acentuado inicialmente e, posteriormente, uma diminuição mais suave à medida que o número de grupos aumenta. Isso sugere que a formação de grupos mais distintos e separados é mais evidente com um número de grupos não tão elevado.

Observando os resultados do agrupamento com o método Aglomerativo, há uma tendência semelhante em relação ao coeficiente de silhueta, em que ocorre um aumento

Tabela 17 – Os nove cursos que compõem o Grupo 6 do agrupamento definitivo com o método Aglomerativo. Fonte: o autor (2024).

Campus	**Grau Acadêmico**	**Curso**
SEDE	Bacharelado	Ciências Biológicas
UAST	Bacharelado	Administração
		Agronomia
		Ciências Biológicas
		Ciências Econômicas
		Engenharia de Pesca
		Zootecnia
	Licenciatura	Letras - Português e Inglês
		Química

inicial seguido por uma diminuição à medida que o número de grupos aumenta. Isso sugere que a qualidade dos grupos pode ser comprometida com um número excessivo de grupos, indicando a presença de grupos sobrepostos ou mal definidos. Também se nota que tanto o índice de Calinski-Harabasz quanto o índice de Davies-Bouldin apresentam uma tendência a diminuir à medida que o número de grupos aumenta, o que pode indicar uma redução na separação e distinção entre os grupos.

No presente estudo, alguns cursos de graduação frequentemente considerados semelhantes pelo senso comum, como determinadas licenciaturas e suas correspondentes versões bacharelado, foram alocados em grupos distintos devido às diferenças significativas em seus perfis quantitativos e comportamentais. Deve-se considerar o fato de que, neste estudo, os perfis curriculares dos cursos não foram utilizados, havendo preferência pelo uso de variáveis que trazem sumarizações quanto a quantidades de vagas, ingressantes, matrículas e concluintes, disponíveis no Censo da Educação Superior de 2021. Embora compartilhem a área do conhecimento, a análise revelou que alguns cursos teoricamente próximos possuem padrões diferenciados em termos de ingresso e formação de estudantes. As licenciaturas tendem a atrair alunos interessados em seguir carreiras na educação, enquanto os bacharelados frequentemente se destinam a estudantes com foco em pesquisa e outras áreas profissionais não diretamente ligadas à educação. Além disso, os cursos considerados neste estudo se dividem em quatro unidades acadêmicas localizadas em diferentes regiões do Estado de Pernambuco, e fatores regionais podem afetar esses padrões, influenciando as preferências e o comportamento dos estudantes. Essas diferenças

refletem-se nos dados quantitativos, justificando a separação em grupos distintos para permitir uma análise mais precisa e relevante, destacando as particularidades e tendências de cada curso.

Ambos os métodos mostram uma tendência de convergência para um número de grupos em torno de 5 ou 6, haja vista que os grupos apresentam uma boa coesão e separação, conforme indicado pelo coeficiente de silhueta. Além disso, ambos os métodos produziram grupos que refletem a diversidade de disciplinas acadêmicas, abrangendo áreas que vão desde as ciências naturais e engenharias até as ciências sociais e humanas. No agrupamento definitivo com K-Means, cinco grupos foram obtidos, sendo o Grupo 1 o maior em termos de número de cursos (16 cursos), abrangendo principalmente disciplinas das áreas de ciências biológicas, engenharias e letras. Esse grupo reflete uma ampla gama de disciplinas, sugerindo uma abordagem mais generalista ou multidisciplinar. Em contraste, o agrupamento definitivo com o método Aglomerativo gerou seis grupos, com o Grupo 1, o maior grupo obtido (11 cursos), compreendendo cursos das áreas de administração, ciência da computação, ciências econômicas e outras disciplinas relacionadas. Ambos os métodos agruparam cursos de engenharia em grupos distintos, sugerindo uma forte especialização nessa área dentro da universidade.

Apesar das diferenças nas quantidades de grupos dos agrupamentos definitivos de cada método, uma semelhança notável pode ser percebida ao comparar os grupos gerados. Trata-se da presença de três grupos idênticos nos resultados de ambos os métodos, o que merece destaque. Especificamente, o Grupo 5 do K-Means e o Grupo 1 do Aglomerativo são iguais (11 cursos), assim como o Grupo 4 de ambos (9 cursos) e o Grupo 2 de ambos (5 cursos), que são compostos pelos mesmos cursos. O padrão de consistência entre os métodos sugere que a natureza desses cursos e suas características intrínsecas levaram a grupos semelhantes, independentemente do método utilizado. Essa consistência ressalta a robustez e confiabilidade dos agrupamentos propostos, pois eles destacam de maneira substancial a afinidade entre cursos específicos no contexto da universidade em questão. Diante do exposto, considera-se viável selecionar um dos grupos idênticos com o objetivo de observar os dados de estudantes de seus cursos em conjunto para aumento da quantidade de amostras a serem usadas na construção de modelos de previsão da evasão, o que responde de forma afirmativa à Questão de Pesquisa 1 (QP1).

6 Classificação de Estudantes em Risco de Evasão

Este capítulo apresenta os resultados obtidos após a construção dos modelos de classificação propostos. O método trata de seis cenários que visam à previsão da evasão no ensino superior, observando momentos distintos da trajetória acadêmica. Em cada cenário, são construídos modelos a partir dos algoritmos XGBoost, SVC e Logistic Regression, com o uso do método de validação cruzada estratificada de cinco grupos. Também é construído, em cada cenário, um *ensemble* de classificadores por meio da estratégia de votação *soft*, com o método de reamostragem *hold-out*, com 70% do conjunto de dados sendo empregado para treino e 30% para teste. Os classificadores são avaliados com as métricas de acurácia, precisão, *recall*, *f1-score* e AUC.

Os dados do histórico acadêmico de estudantes usados para a construção dos modelos são selecionados com a observação dos agrupamentos realizados na primeira parte do trabalho, que tratou de agrupar cursos de graduação presenciais da Universidade Federal Rural de Pernambuco (UFRPE). Especificamente, foram selecionados dados de estudantes dos cursos que compõem um dos grupos gerados, o Grupo 2, com ingresso a partir de 2010. Nos agrupamentos K-Means e Aglomerativo, o Grupo 2 é composto pelos mesmos cinco cursos da mesma unidade acadêmica, a SEDE: Bacharelado em Agronomia, Licenciatura em Ciências Biológicas, Licenciatura em Matemática, Bacharelado em Medicina Veterinária e Licenciatura em Química. Os dados dos estudantes foram coletados a partir de sistemas de gestão acadêmica da UFRPE, após as devidas autorizações.

Quanto às análises dos dados dos semestres acadêmicos dos estudantes, as figuras 21 e 22 exibem, respectivamente, as distribuições de frequência das variáveis do curso e do período de ingresso, enquanto as figuras 23, 24, 25 e 26 exibem, respectivamente, os histogramas das variáveis de média do semestre, carga horária de aprovações acumulada, carga horária de reprovações acumulada e carga horária de cancelamentos acumulada.

Curso	Contagem	Frequência (%)
BACH. EM AGRONOMIA (SEDE)	5796	26.6%
LIC. EM CIÊNCIAS BIOLÓGICAS (SEDE)	5505	25.2%
BACH. EM MEDICINA VETERINÁRIA (SEDE)	4252	19.5%
LIC. EM QUÍMICA (SEDE)	3191	14.6%
LIC. EM MATEMÁTICA (SEDE)	3081	14.1%

Figura 21 – Frequências do curso. Fonte: o autor (2024).

Período de ingresso	Contagem	Frequência (%)
1	11796	54.0%
2	10029	46.0%

Figura 22 – Frequências do período de ingresso. Fonte: o autor (2024).

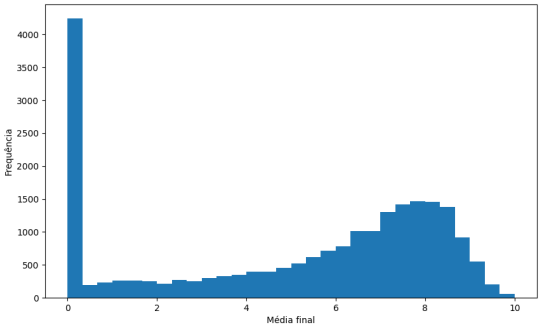

Figura 23 – Frequências da média final. Fonte: o autor (2024).

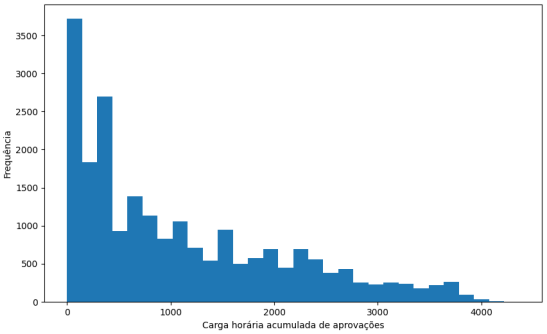

Figura 24 – Frequências da carga horária de aprovações. Fonte: o autor (2024).

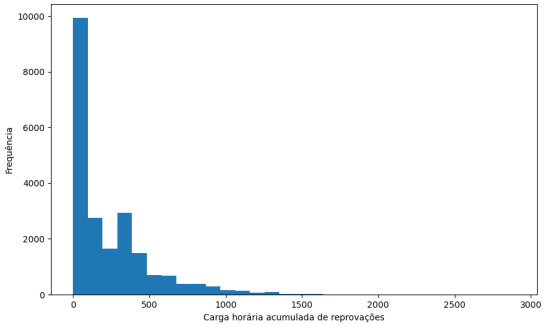

Figura 25 – Frequências da carga horária de reprovações. Fonte: o autor (2024).

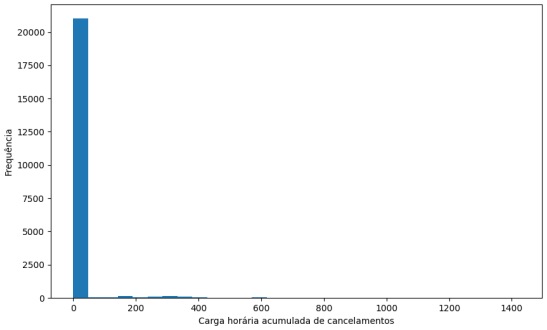

Figura 26 – Frequências da carga horária de cancelamentos. Fonte: o autor (2024).

Os histogramas apresentados consideram todas as amostras de semestres acadêmicos existentes na base de dados observada no estudo de caso executado, que possui dados de estudantes que ingressaram a partir de 2010 nos cursos do grupo de cursos selecionado.

6.1 Evasão ou Formação

A Tabela 18 apresenta a distribuição de frequências para cada classe empregada no primeiro cenário. A Tabela 19 exibe informações estatísticas sobre as variáveis quantitativas utilizadas. As figuras 27 e 28 apresentam *boxplots* das variáveis em questão.

Tabela 18 – Distribuição de frequências por classe do primeiro cenário. Fonte: o autor (2024).

Classe	**Quantidade de Ocorrências**
1. Evasão	9075
2. Formação	9075

Tabela 19 – Variáveis quantitativas utilizadas no primeiro cenário. Fonte: o autor (2024).

Variável	**Min.**	**Máx.**	**Média**	**Desv. Pad.**	**Mediana**
duracao_vinculo	1	18	4,01	3,01	3
diferenca_conc_em_inicio_graduacao	1	50	5,52	6,01	4
media_final	0	10	4,76	3,30	5,80
ch_total	15	4200	378,36	264,65	360
ch_aprovacoes_acum	0	4170	989,85	986,83	630
ch_reprovacoes_acum	0	2895	243,68	284,93	180
ch_cancelamentos_acum	0	1425	12,61	75,87	0

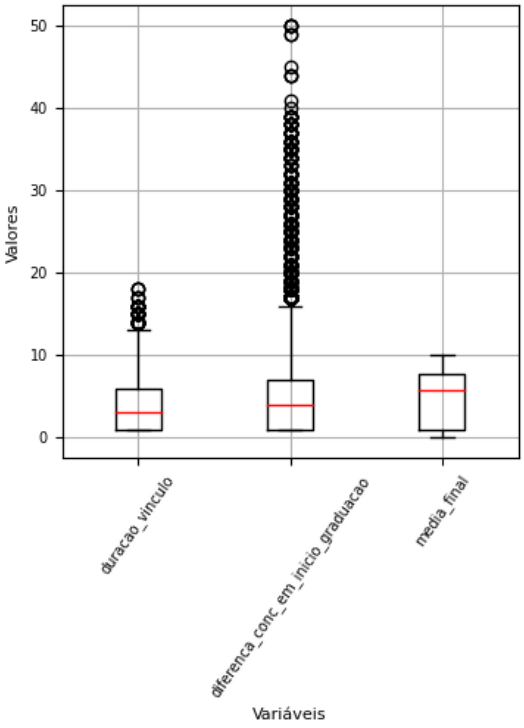

Figura 27 – *Boxplots* - Cen. 1 (parte 1).
Fonte: o autor (2024).

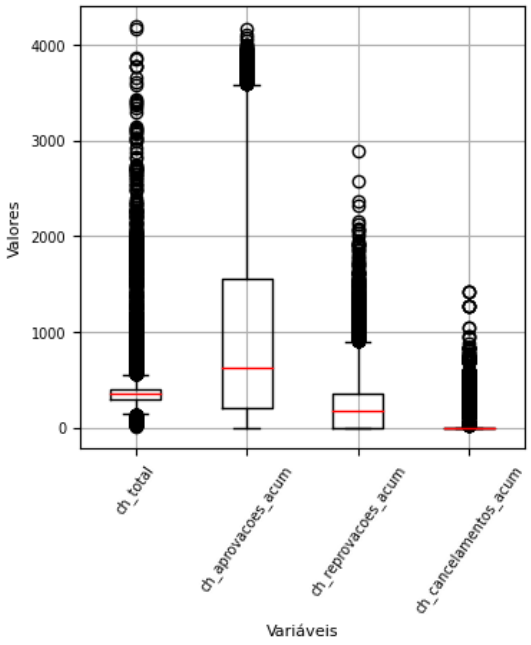

Figura 28 – *Boxplots* - Cen. 1 (parte 2).
Fonte: o autor (2024).

A Tabela 20 exibe o resumo dos resultados obtidos pelos modelos construídos com validação cruzada de cinco grupos (*5-fold CV*), além dos resultados do *ensemble* Voting Classifier construído com *hold-out* 70/30, observando as métricas de acurácia, precisão, *recall* e *f1-score*. No caso dos modelos construídos com validação cruzada, os valores mostrados por classe indicam a média dos resultados obtidos para cada classe, por modelo, em cada iteração do processo.

Tabela 20 – Resultados dos modelos construídos no primeiro cenário. Fonte: o autor (2024).

	5-fold CV									***Hold-out 70/30***		
	XGBoost			**Support Vector Machine**			**Logistic Regression**			**Voting Classifier**		
	Precisão	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***
1. Evasão	0,906222	0,874270	0,889946	0,906374	0,846501	0,875393	0,881378	0,854215	0,867564	0,903955	0,877193	0,890373
2. Formação	0,878597	0,909532	0,893785	0,856026	0,912507	0,883348	0,858563	0,884959	0,871543	0,879570	0,905869	0,892526
Média	0,892410	0,891901	0,891865	0,881200	0,879504	0,879370	0,869971	0,869587	0,869554	0,891762	0,891531	0,891449
Acurácia	0,891901			0,879504			0,869587			0,891460		

As figuras 29, 31 e 33 mostram o resumo das matrizes de confusão dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 30, 32 e 34 apresentam o resumo dos gráficos AUC/ROC dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 35 e 36 mostram, respectivamente, a matriz de confusão e o gráfico AUC/ROC do Voting Classifier.

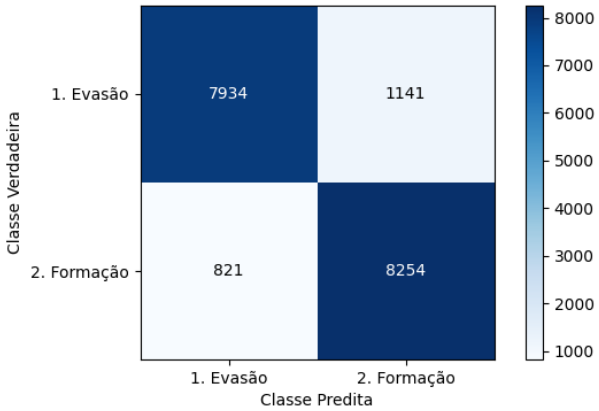

Figura 29 – Matriz de confusão - XGBoost - Cen. 1. Fonte: o autor (2024).

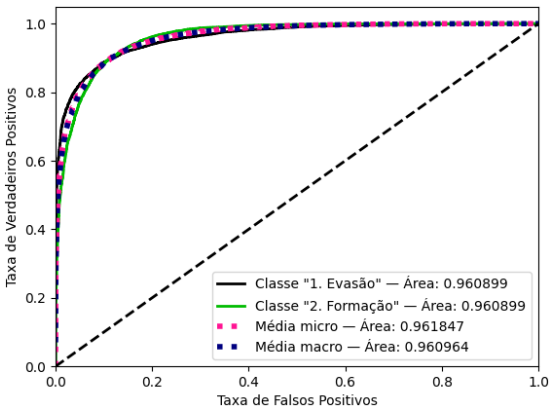

Figura 30 – AUC/ROC - XGBoost - Cen. 1. Fonte: o autor (2024).

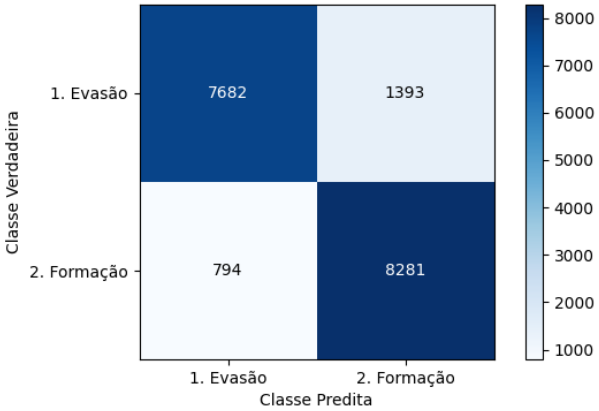

Figura 31 – Matriz de confusão - SVC - Cen. 1. Fonte: o autor (2024).

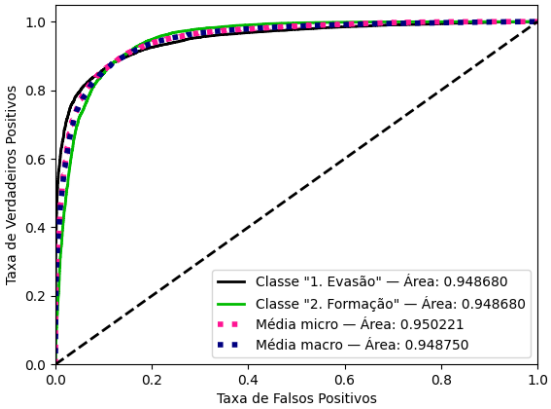

Figura 32 – AUC/ROC - SVC - Cen. 1. Fonte: o autor (2024).

Tratando dos modelos individuais construídos com validação cruzada, o XGBoost apresenta uma acurácia alta, indicando que a maioria das previsões está correta. A precisão, *recall* e *f1-score* estão próximos, sugerindo um bom equilíbrio entre a capacidade de classificar corretamente entre *"Evasão"* e *"Formação"*. O modelo SVC apresenta uma acurácia ligeiramente inferior à do XGBoost, mas ainda relativamente alta. As métricas de precisão, *recall* e *f1-score* são próximas, indicando um bom desempenho geral do modelo.

O modelo de Logistic Regression apresenta a menor acurácia entre os três, mas ainda está em um nível razoável. As métricas de precisão, *recall* e *f1-score* são consistentes entre si.

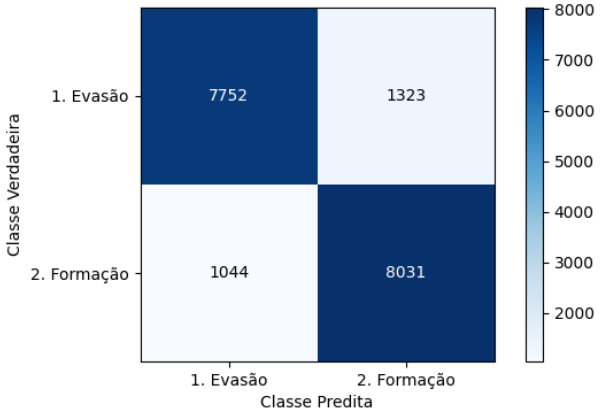

Figura 33 – Matriz de confusão - Log. Regress. - Cen. 1. Fonte: o autor (2024).

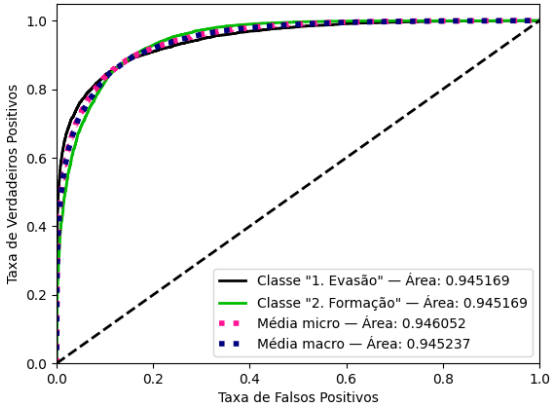

Figura 34 – AUC/ROC - Log. Regress. - Cen. 1. Fonte: o autor (2024).

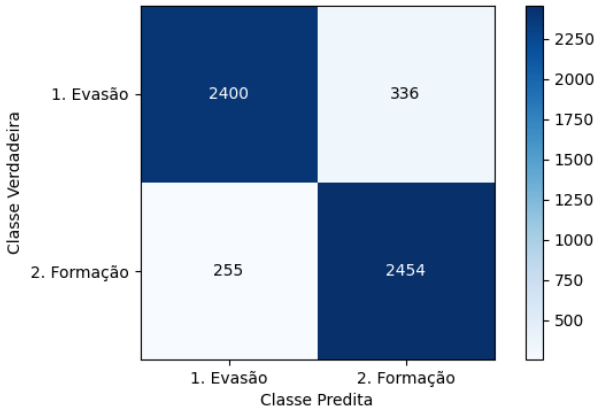

Figura 35 – Matriz de confusão - Voting Classifier - Cen. 1. Fonte: o autor (2024).

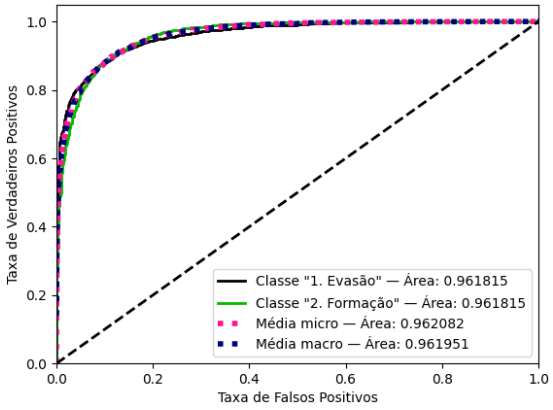

Figura 36 – AUC/ROC - Voting Classifier - Cen. 1. Fonte: o autor (2024).

Quanto aos resultados do *ensemble* Voting Classifier, para a classe *"Evasão"*, o modelo apresentou uma precisão de 90,40% e um *recall* de 87,72%. Esses números indicam que a grande maioria dos exemplos classificados como *"Evasão"* pelo modelo realmente pertencem a essa classe, considerando a alta precisão, e que o modelo conseguiu capturar uma proporção significativa dos exemplos de *"Evasão"* presentes no conjunto de dados, tendo em vista o *recall* razoavelmente alto. O valor de *f1-score* para essa classe foi de 89,04%, sugerindo um equilíbrio entre precisão e *recall*, o que é importante para garantir uma boa performance em tarefas de classificação.

No que diz respeito à classe *"Formação"*, o *ensemble* obteve uma precisão de 87,96%

e um *recall* de 90,59%. Isso indica que a grande maioria dos exemplos classificados como *"Formação"* pelo modelo realmente pertencem a essa classe, observando a alta precisão, e que o modelo conseguiu capturar a maioria dos exemplos de *"Formação"* presentes no conjunto de dados, considerando o *recall* elevado. O valor de *f1-score* para essa classe foi de 89,25%, sugerindo um bom equilíbrio entre precisão e *recall*.

Observa-se que, para ambas as classes, o *ensemble* apresenta resultados bastante sólidos, com precisão e *recall* relativamente altos e valores de *f1-score* equilibrados. Isso sugere que o modelo proposto é capaz de realizar uma classificação confiável tanto para *"Evasão"* quanto para *"Formação"*. A Figura 37 exibe as 15 variáveis mais importantes para o *ensemble*, observando as classes utilizadas. Como este cenário trata da construção de classificadores binários, a importância de cada variável é tipicamente a mesma para cada classe. Por meio da figura, nota-se que as variáveis *ch_ aprovacoes_ acum*, *media_ final*, *ch_ reprovacoes_ acum*, *ch_ total* e *duracao_ vinculo* foram as mais importantes para a previsão tanto da formação quanto da evasão no contexto abordado. Esses resultados ressaltam diversos aspectos do fenômeno da evasão.

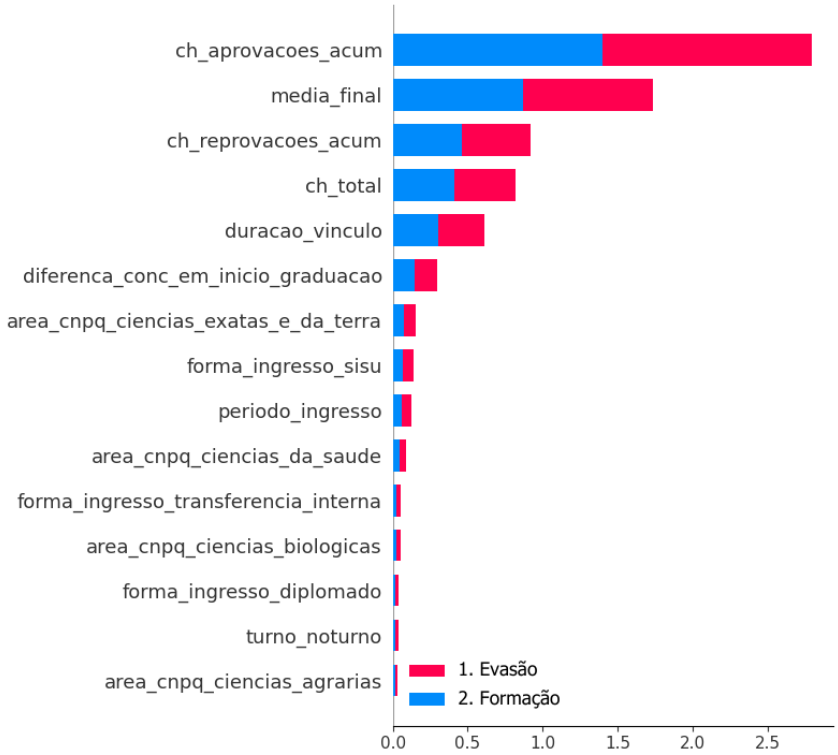

Figura 37 – Importância das variáveis - Voting Classifier - Cen. 1. Fonte: o autor (2024).

A variável *ch_ aprovacoes_ acum* está relacionada ao desempenho acadêmico do estudante, sugerindo que um histórico de aprovações pode ser um indicativo de persistência

no curso. Por outro lado, a *ch_reprovacoes_acum* pode sinalizar dificuldades enfrentadas pelo aluno, podendo influenciar negativamente na sua permanência. A *media_final* é um indicador direto do desempenho acadêmico global e pode refletir o comprometimento do estudante com seus estudos. Já a variável *ch_total* está relacionada à carga horária total do curso, possivelmente indicando a intensidade e o rigor do programa acadêmico, o que pode afetar a decisão de permanecer ou abandonar o curso. Por fim, a *duracao_vinculo* aponta para a duração do vínculo do estudante no curso, sendo algo crucial a ser analisado para o entendimento da evasão ou formação.

6.2 Evasão até quatro anos, Evasão após quatro anos ou Formação

A Tabela 21 apresenta a distribuição de frequências para cada classe empregada no segundo cenário. A Tabela 22 exibe informações estatísticas sobre as variáveis quantitativas utilizadas. As figuras 38 e 39 apresentam *boxplots* das variáveis em questão.

Tabela 21 – Distribuição de frequências por classe do segundo cenário. Fonte: o autor (2024).

Classe	**Quantidade de Ocorrências**
1. Evasão até 4 anos	1518
2. Evasão após 4 anos	1518
3. Formação	1518

Tabela 22 – Variáveis quantitativas utilizadas no segundo cenário. Fonte: o autor (2024).

Variável	**Min.**	**Máx.**	**Média**	**Desv. Pad.**	**Mediana**
duracao_vinculo	1	18	4,33	3,20	3
diferenca_conc_em_inicio_graduacao	1	50	5,77	6,23	4
media_final	0	10	4,34	3,22	4,96
ch_total	30	3780	363	248,59	330
ch_aprovacoes_acum	0	4005	923,57	904,20	630
ch_reprovacoes_acum	0	2895	317,92	347,80	240
ch_cancelamentos_acum	0	1425	19,76	93,18	0

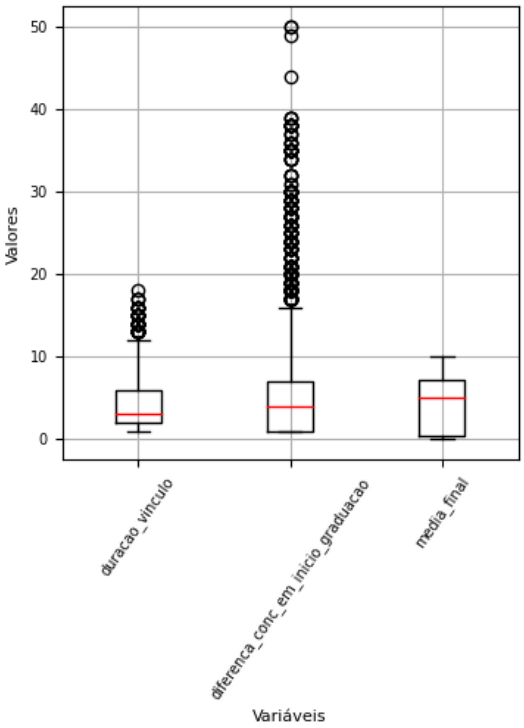

Figura 38 – *Boxplots* - Cen. 2 (parte 1).
Fonte: o autor (2024).

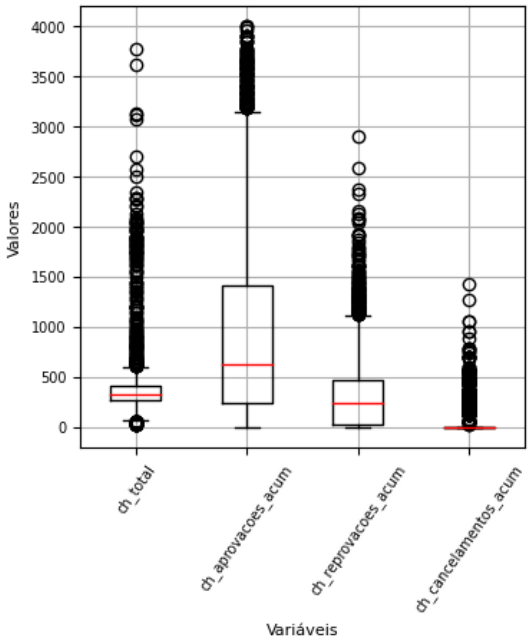

Figura 39 – *Boxplots* - Cen. 2 (parte 2).
Fonte: o autor (2024).

A Tabela 23 exibe o resumo dos resultados obtidos pelos modelos construídos com validação cruzada de cinco grupos (*5-fold CV*), além dos resultados do *ensemble* Voting Classifier construído com *hold-out* 70/30, observando as métricas de acurácia, precisão, *recall* e *f1-score*. No caso dos modelos construídos com validação cruzada, os valores mostrados por classe indicam a média dos resultados obtidos para cada classe, por modelo, em cada iteração do processo.

As figuras 40, 42 e 44 mostram as matrizes de confusão dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 41, 43 e 45 apresentam os gráficos AUC/ROC dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 46 e 47 mostram, respectivamente, a matriz de confusão e o gráfico AUC/ROC do Voting Classifier.

Tabela 23 – Resultados dos modelos construídos no segundo cenário. Fonte: o autor (2024).

	5-fold CV									***Hold-out 70/30***		
	XGBoost			**Support Vector Machine**			**Logistic Regression**			**Voting Classifier**		
	Precisão	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***
1. Evasão até 4 anos	0,782864	0,768137	0,774714	0,740215	0,787893	0,762921	0,722906	0,787231	0,753342	0,772908	0,800000	0,786221
2. Evasão após 4 anos	0,722222	0,696986	0,708138	0,747148	0,608064	0,669709	0,711017	0,562611	0,627885	0,729064	0,651982	0,688372
3. Formação	0,777180	0,816187	0,796068	0,738828	0,827386	0,780523	0,728156	0,814222	0,768624	0,760349	0,815421	0,786922
Média	0,760755	0,760437	0,759640	0,742064	0,741114	0,737718	0,720693	0,721354	0,716617	0,754107	0,755801	0,753838
Acurácia	0,760430			0,741103			0,721341			0,755669		

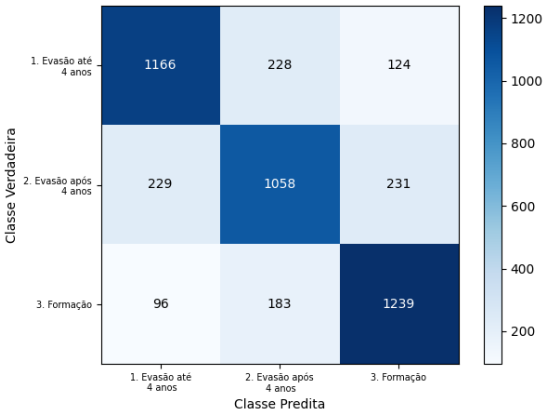

Figura 40 – Matriz de confusão - XGBoost - Exp 2. Fonte: o autor (2024).

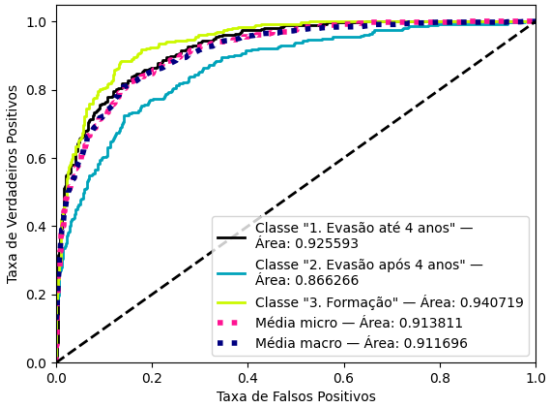

Figura 41 – AUC/ROC - XGBoost - Exp 2. Fonte: o autor (2024).

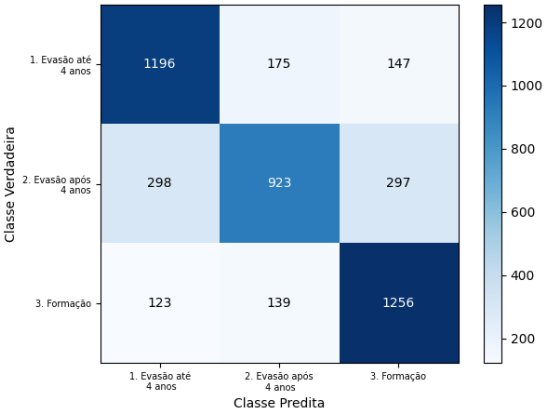

Figura 42 – Matriz de confusão - SVC - Exp 2. Fonte: o autor (2024).

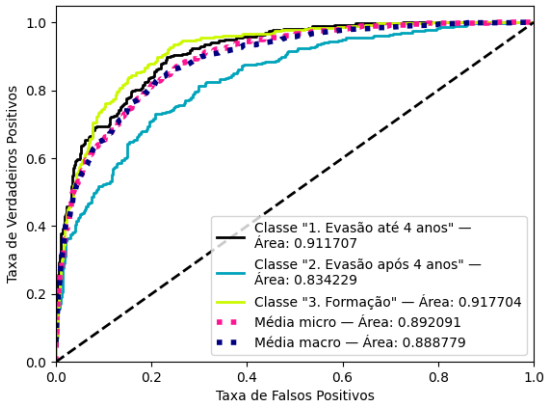

Figura 43 – AUC/ROC - SVC - Exp 2. Fonte: o autor (2024).

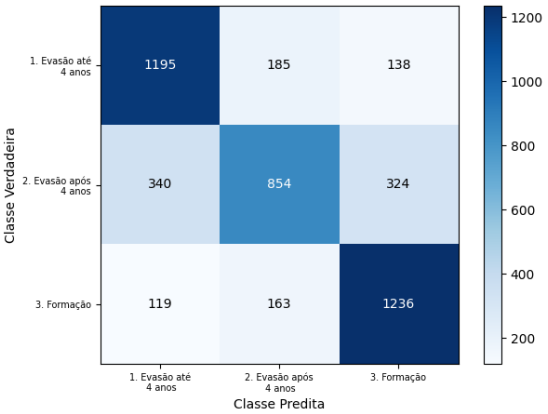

Figura 44 – Matriz de confusão - Log. Regress. - Exp 2. Fonte: o autor (2024).

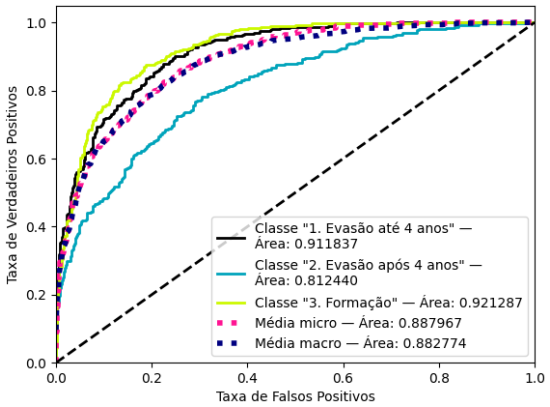

Figura 45 – AUC/ROC - Log. Regress. - Exp 2. Fonte: o autor (2024).

Quanto aos resultados dos modelos individuais, no segundo cenário, o XGBoost apresentou o melhor desempenho entre os três modelos avaliados por meio da validação

cruzada de cinco grupos. Com uma acurácia de 76,04%, o XGBoost foi capaz de realizar classificações com precisão, mantendo uma boa proporção entre precisão, *recall* e *f1-score*. Sua precisão de 76,08% indica a capacidade de classificar corretamente a maioria das instâncias, enquanto o *recall* de 76,04% sugere que o modelo é capaz de identificar corretamente a maioria das instâncias no conjunto de dados. Embora o SVC tenha alcançado uma acurácia de 74,11%, próxima da acurácia do XGBoost, seu desempenho ficou abaixo do XGBoost em termos de precisão, *recall* e *f1-score*. Ainda assim, a precisão de 74,21% e o *recall* de 74,11% indicam uma capacidade decente de distinguir entre as classes. Por fim, o Logistic Regression apresentou o desempenho mais baixo entre os três modelos, com uma acurácia de 72,13%. Suas métricas de precisão, *recall* e *f1-score* também foram inferiores às dos outros modelos.

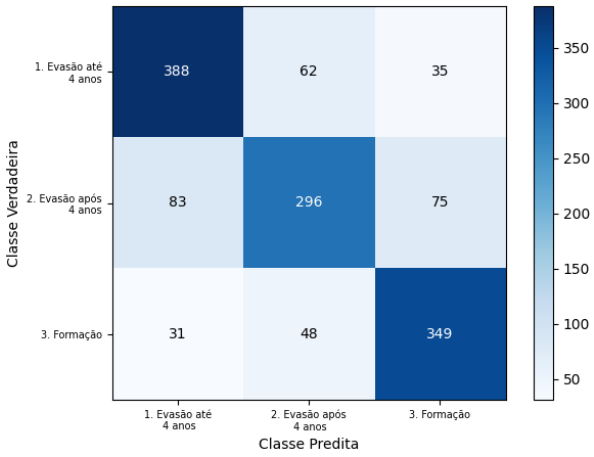

Figura 46 – Matriz de confusão - Voting Classifier - Exp 2. Fonte: o autor (2024).

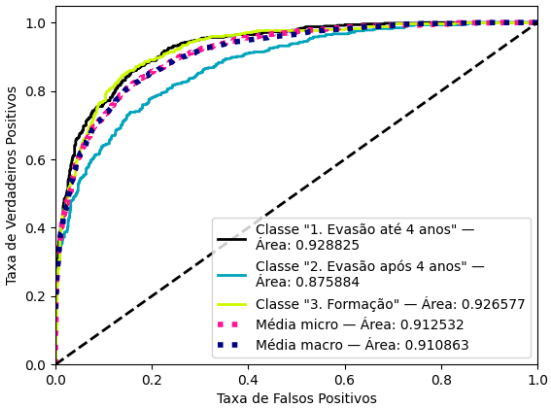

Figura 47 – AUC/ROC - Voting Classifier - Exp 2. Fonte: o autor (2024).

Tratando do *ensemble* Voting Classifier, para a classe *"Evasão até 4 anos"*, o modelo apresentou uma precisão de 77,29% e um *recall* de 80%. Esses resultados indicam que a maioria dos exemplos classificados como *"Evasão até 4 anos"* pelo modelo realmente pertencem a essa classe, considerando a alta precisão, e que o modelo conseguiu capturar uma proporção significativa dos exemplos verdadeiros de *"Evasão até 4 anos"* presentes no conjunto de dados, tendo em vista o *recall* elevado. O valor de *f1-score* para essa classe foi de 78,62%, sugerindo um equilíbrio entre precisão e *recall*, o que é fundamental para garantir uma boa performance em tarefas de classificação para essa classe específica.

Para a classe *"Evasão após 4 anos"*, o *ensemble* obteve uma precisão de 72,91% e um *recall* de 65,20%. Isso indica que uma parte considerável dos exemplos classificados como *"Evasão após 4 anos"* pelo modelo realmente pertencem a essa classe, observando a

precisão razoável, e que o modelo conseguiu capturar uma proporção significativa, embora menor, dos exemplos verdadeiros de *"Evasão após 4 anos"* presentes no conjunto de dados, tendo em vista o *recall* razoável. O *f1-score* para essa classe foi de 68,84%, indicando um equilíbrio aceitável entre precisão e *recall*, embora o *recall* seja mais baixo do que o desejado.

Para a classe *"Formação"*, o *ensemble* apresentou uma precisão de 76,03% e um *recall* de 81,54%. Esses resultados indicam que a maioria dos exemplos classificados como *"Formação"* pelo modelo realmente pertencem a essa classe, por conta da alta precisão, e que o modelo conseguiu capturar a maioria dos exemplos verdadeiros de *"Formação"* presentes no conjunto de dados, diante do *recall* elevado. O valor de *f1-score* para essa classe foi de 78,69%, sugerindo um bom equilíbrio entre precisão e *recall*.

Observa-se que o *ensemble* apresenta resultados sólidos para as classes *"Evasão até 4 anos"* e *"Formação"*, com precisão e *recall* relativamente altos e *f1-score*s equilibrados. Para a classe *"Evasão após 4 anos"*, embora a precisão seja aceitável, o *recall* é um pouco mais baixo, o que pode indicar uma necessidade de melhorias para capturar uma maior proporção dos exemplos verdadeiros dessa classe. Entretanto, deve ser levado em conta o fato de que o cenário em questão trata da construção de um classificador multiclasse, sendo esperada uma maior dificuldade de previsão. Dessa forma, o *ensemble* demonstra uma boa capacidade de realizar a classificação eficaz de classes específicas de evasão.

A Figura 48 exibe as 15 variáveis mais importantes para o *ensemble*. Por meio da figura, nota-se a importância de cada variável para cada classe considerada. Para a classe *"Evasão até 4 anos"*, as variáveis mais importantes foram *media_final*, *duracao_vinculo*, *ch_aprovacoes_acum* e *ch_reprovacoes_acum*, nessa ordem. Para a classe *"Evasão após 4 anos"*, as variáveis mais importantes foram *duracao_vinculo*, *ch_aprovacoes_acum*, *ch_reprovacoes_acum* e *media_final*, nessa ordem. Para a classe *"Formação"*, as variáveis mais importantes foram *media_final*, *ch_aprovacoes_acum* e *ch_reprovacoes_acum*, nessa ordem.

Quanto às classes *"Evasão até 4 anos"* e *"Formação"*, é interessante notar que ambas compartilham semelhanças significativas em relação às variáveis mais importantes para a previsão. Em ambas as classes, *media_final*, *ch_aprovacoes_acum* e *ch_reprovacoes_acum* surgem como fatores determinantes. Isso sugere que o desempenho acadêmico e o histórico de aprovações e reprovações têm um impacto substancial tanto na decisão de permanecer

no curso até a formação quanto na evasão nos primeiros anos do curso. No entanto, uma distinção crucial surge com a inclusão da variável *duracao_ vinculo* como mais importante para a classe *"Evasão até 4 anos"*. Isso implica que a quantidade de períodos de vínculo com a instituição pode desempenhar um papel mais relevante na evasão durante os estágios iniciais do curso, possivelmente indicando uma janela crítica de tempo em que intervenções específicas podem ser mais eficazes para reter os alunos.

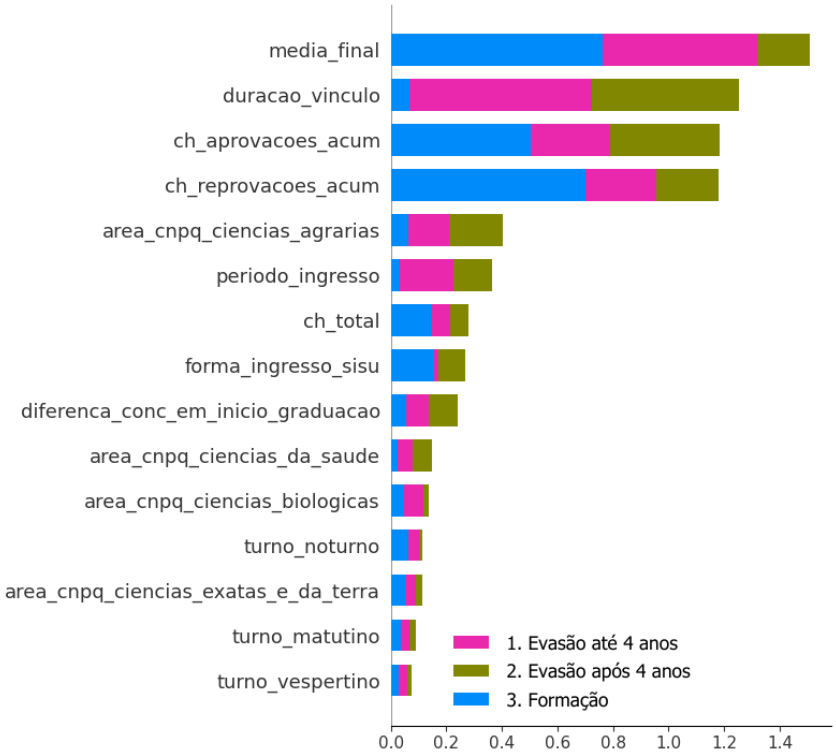

Figura 48 – Importância das variáveis - Voting Classifier - Exp 2. Fonte: o autor (2024).

A classe *"Evasão após 4 anos"* se destaca por uma mudança na ordem de importância das variáveis em comparação com as outras classes. Enquanto *ch_ aprovacoes_ acum* e *ch_ reprovacoes_ acum* permanecem relevantes, a variável *duracao_ vinculo* assume uma posição de destaque, indicando que, à medida que os alunos avançam em seus estudos, a quantidade de períodos de permanência na instituição torna-se um fator ainda mais crítico para a evasão. Essa distinção sugere que a gestão eficaz do tempo de permanência dos alunos pode desempenhar um papel crucial em sua retenção em estágios mais avançados do curso, possivelmente requerendo estratégias específicas para lidar com as necessidades dos alunos em diferentes momentos de sua jornada acadêmica.

6.3 Evasão até dois anos, Evasão entre dois e quatro anos, Evasão após quatro anos ou Formação

A Tabela 24 apresenta a distribuição de frequências para cada classe empregada no terceiro cenário. A Tabela 25 exibe informações estatísticas sobre as variáveis quantitativas utilizadas. As figuras 49 e 50 apresentam *boxplots* das variáveis em questão.

Tabela 24 – Distribuição de frequências por classe do terceiro cenário. Fonte: o autor (2024).

Classe	**Quantidade de Ocorrências**
1. Evasão até 2 anos	1518
2. Evasão entre 2 e 4 anos	1518
3. Evasão após 4 anos	1518
4. Formação	1518

Tabela 25 – Variáveis quantitativas utilizadas no terceiro cenário. Fonte: o autor (2024).

Variável	**Min.**	**Máx.**	**Média**	**Desv. Pad.**	**Mediana**
duracao_vinculo	1	18	3,93	3	3
diferenca_conc_em_inicio_graduacao	1	45	5,93	6,21	5
media_final	0	10	3,90	3,23	4,06
ch_total	15	4200	357,06	259,60	315
ch_aprovacoes_acum	0	4095	785,64	854	450
ch_reprovacoes_acum	0	2895	323,59	329,16	255
ch_cancelamentos_acum	0	1425	17,82	87,51	0

A Tabela 26 exibe o resumo dos resultados obtidos pelos modelos construídos com validação cruzada de cinco grupos (*5-fold CV*), além dos resultados do *ensemble* Voting Classifier construído com *hold-out* 70/30, observando as métricas de acurácia, precisão, *recall* e *f1-score*. Para os modelos construídos com validação cruzada, os valores mostrados por classe indicam a média dos resultados obtidos para cada classe, por modelo, em cada iteração do processo.

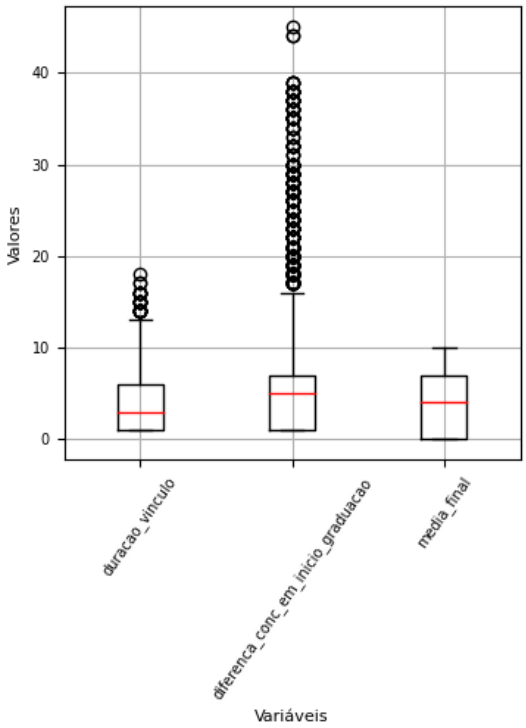

Figura 49 – *Boxplots* - Cen. 3 (parte 1).
Fonte: o autor (2024).

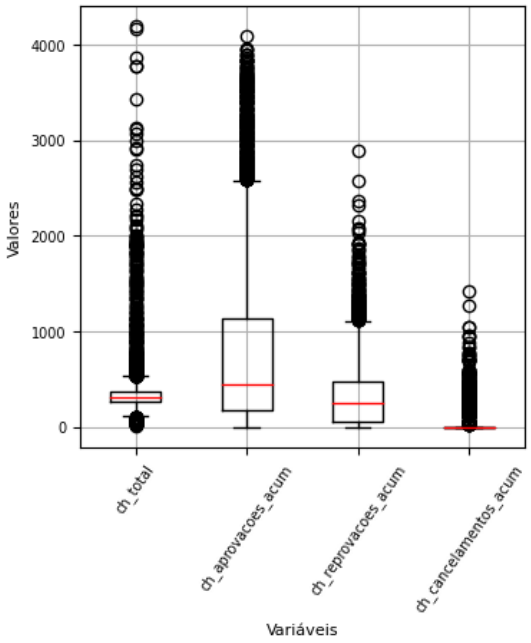

Figura 50 – *Boxplots* - Cen. 3 (parte 2).
Fonte: o autor (2024).

As figuras 51, 53 e 55 mostram as matrizes de confusão dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 52, 54 e 56 apresentam os gráficos AUC/ROC dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 57 e 58 mostram, respectivamente, a matriz de confusão e o gráfico AUC/ROC do Voting Classifier.

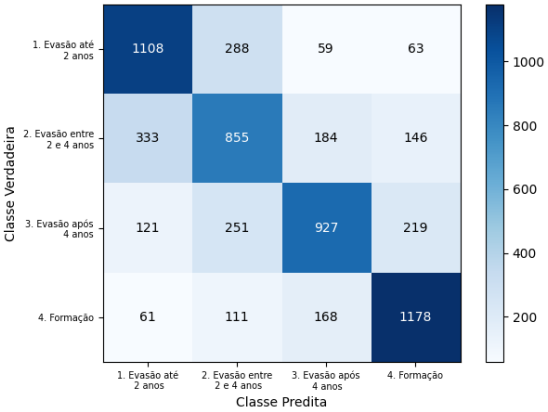

Figura 51 – Matriz de confusão - XGBoost - Cen. 3. Fonte: o autor (2024).

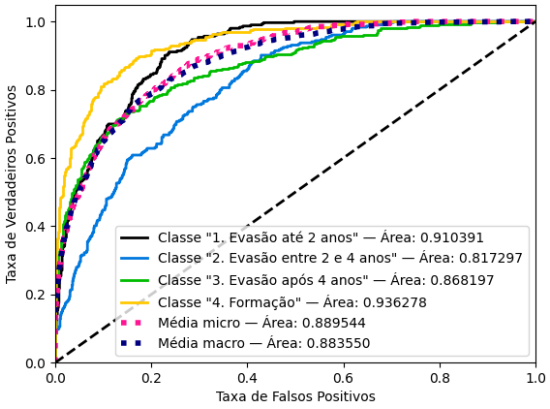

Figura 52 – AUC/ROC - XGBoost - Cen. 3. Fonte: o autor (2024).

Tabela 26 – Resultados dos modelos construídos no terceiro cenário. Fonte: o autor (2024).

	5-fold CV									***Hold-out 70/30***		
	XGBoost			**Support Vector Machine**			**Logistic Regression**			**Voting Classifier**		
	Precisão	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***
1. Evasão até 2 anos	0,683791	0,729868	0,705587	0,668275	0,761519	0,711318	0,666481	0,766133	0,712611	0,705653	0,746392	0,725451
2. Evasão entre 2 e 4 anos	0,569164	0,563264	0,565639	0,541452	0,553400	0,546971	0,496904	0,473671	0,484795	0,546485	0,536748	0,541573
3. Evasão após 4 anos	0,692900	0,610694	0,649082	0,718154	0,488162	0,581041	0,661297	0,491437	0,563832	0,674033	0,572770	0,619289
4. Formação	0,733498	0,776034	0,754089	0,692943	0,801057	0,742956	0,682295	0,786586	0,730645	0,705534	0,772727	0,737603
Média	0,669838	0,669965	0,668600	0,655206	0,651035	0,645572	0,626744	0,629457	0,622971	0,657926	0,657159	0,655979
Acurácia	0,669960			0,651019			0,629446			0,660812		

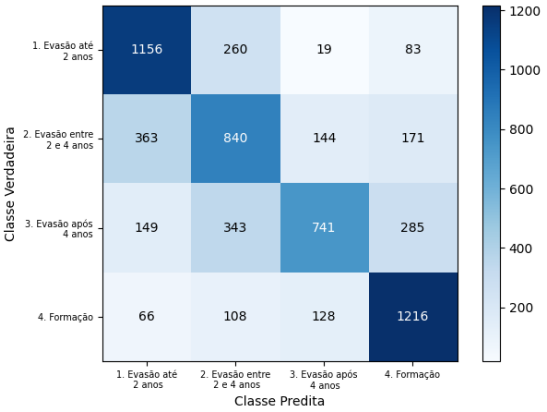

Figura 53 – Matriz de confusão - SVC - Cen. 3. Fonte: o autor (2024).

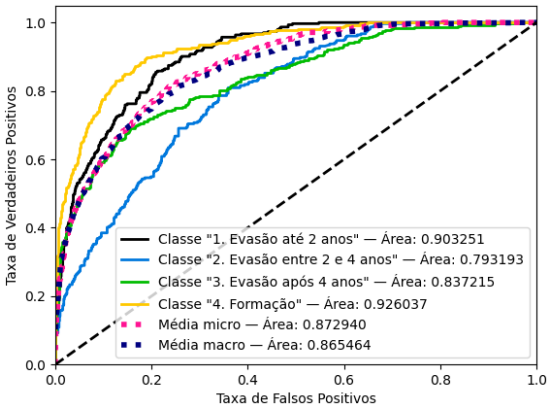

Figura 54 – AUC/ROC - SVC - Cen. 3. Fonte: o autor (2024).

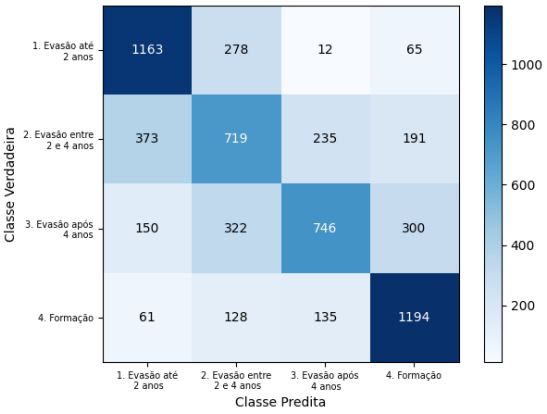

Figura 55 – Matriz de confusão - Log. Regress. - Cen. 3. Fonte: o autor (2024).

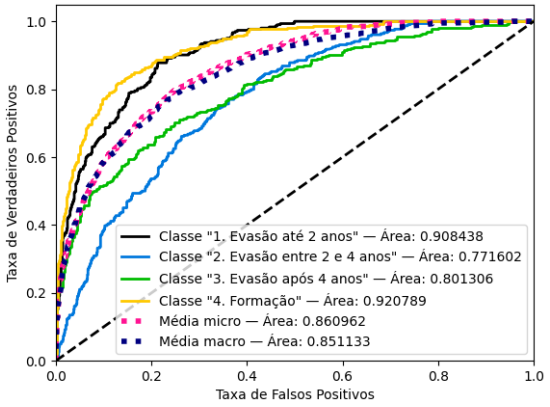

Figura 56 – AUC/ROC - Log. Regress. - Cen. 3. Fonte: o autor (2024).

Considerando os resultados dos modelos individuais, o XGBoost apresenta uma acurácia de aproximadamente 67%, o que indica que cerca de dois terços das previsões estão corretas. Suas métricas de precisão, *recall* e *f1-score* estão próximas da acurácia, sugerindo um desempenho balanceado em termos de classificação correta das classes. O SVC alcançou uma acurácia em torno de 65%, com métricas de precisão, *recall* e *f1-score* também nessa faixa. O Logistic Regression obteve uma acurácia em torno de 63%, com métricas de precisão, *recall* e *f1-score* também nessa faixa.

Quanto aos resultados do *ensemble* Voting Classifier, o modelo apresentou uma precisão razoável para a classe *"Evasão até 2 anos"*, indicando que a maioria dos exemplos classificados como *"Evasão até 2 anos"* pelo modelo realmente pertencem a essa classe. O *recall* também é relativamente alto, sugerindo que o modelo conseguiu capturar a maioria dos exemplos de *"Evasão até 2 anos"* presentes no conjunto de dados. O valor de *f1-score*,

que combina precisão e *recall*, também é satisfatório.

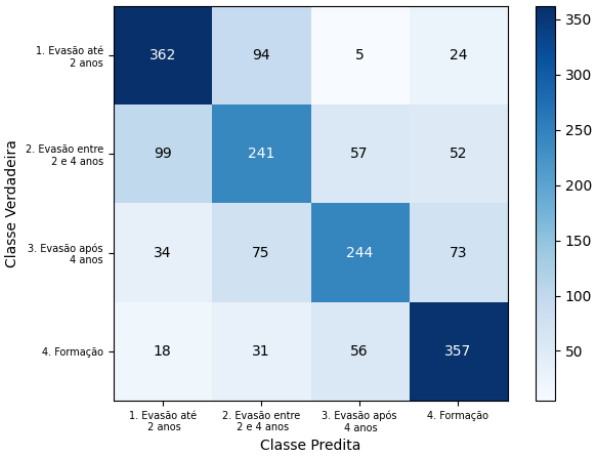

Figura 57 – Matriz de confusão - Voting Classifier - Cen. 3. Fonte: o autor (2024).

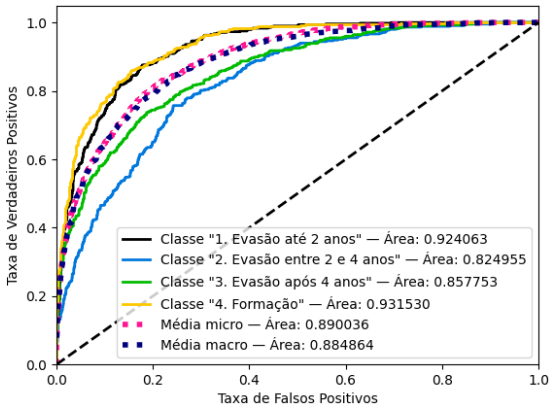

Figura 58 – AUC/ROC - Voting Classifier - Cen. 3. Fonte: o autor (2024).

Para a classe *"Evasão entre 2 e 4 anos"*, o *ensemble* apresentou uma precisão e *recall* mais baixos em comparação com as outras classes. O valor de *f1-score* também é mais baixo. O *ensemble* teve um desempenho intermediário para a classe *"Evasão após 4 anos"*, com uma precisão decente, mas um *recall* um pouco mais baixo. Isso sugere que o modelo pode ter dificuldade em capturar todos os exemplos de *"Evasão após 4 anos"*, resultando em um *f1-score* moderado. Para a classe *"Formação"*, o *ensemble* apresentou uma precisão razoável, indicando que a maioria dos exemplos classificados como *"Formação"* realmente pertencem a essa classe. O *recall* é mais alto, sugerindo que o modelo conseguiu capturar a maioria dos exemplos de *"Formação"* presentes no conjunto de dados. O valor de *f1-score* também é satisfatório.

No geral, o *ensemble* apresenta um desempenho variável para cada classe, com resultados mais sólidos para as classes *"Evasão até 2 anos"* e *"Formação"*, e desempenho inferior para as classes *"Evasão entre 2 e 4 anos"* e *"Evasão após 4 anos"*. A acurácia geral do modelo é de 66,08%, sugerindo uma performance moderada. Considerando que se trata de um modelo multiclasse com quatro classes que tem o objetivo de prever a evasão em momentos específicos do curso, o desempenho inferior para determinadas classes era esperado. Diante disso, pode-se afirmar que os resultados obtidos são satisfatórios.

A Figura 59 apresenta as 15 variáveis mais importantes para o *ensemble*. Por meio da figura, percebe-se a importância de cada variável para cada classe. Para a classe *"Evasão até 2 anos"*, as variáveis mais importantes foram *duracao_vinculo*, *media_final*, *ch_aprovacoes_acum* e *ch_reprovacoes_acum*, nessa ordem. Para a classe *"Evasão entre*

2 e 4 anos", as variáveis mais importantes foram *media_final*, *ch_aprovacoes_acum* e *duracao_vinculo*, nessa ordem. Para a classe *"Evasão após 4 anos"*, as variáveis mais importantes foram *ch_aprovacoes_acum*, *media_final* e *duracao_vinulo*, nessa ordem. Para a classe *"Formação"*, as variáveis mais importantes foram *ch_aprovacoes_acum*, *media_final* e *ch_reprovacoes_acum*, nessa ordem.

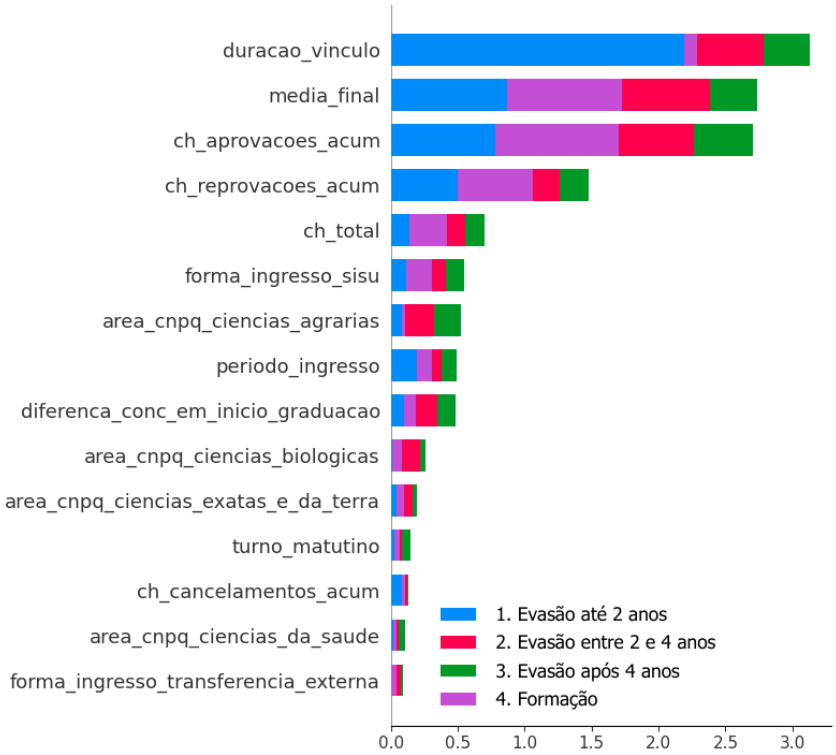

Figura 59 – Importância das variáveis - Voting Classifier - Cen. 3. Fonte: o autor (2024).

Ao analisar as classes *"Evasão até 2 anos"* e *"Evasão entre 2 e 4 anos"*, é evidente que ambas compartilham algumas semelhanças em relação às variáveis mais importantes para prever a evasão. Tanto *media_final* quanto *ch_aprovacoes_acum* surgem como fatores críticos em ambas as classes. No entanto, uma distinção importante surge com a inclusão da variável *duracao_vinculo* como a mais relevante na classe *"Evasão até 2 anos"*, indicando que a quantidade de períodos de permanência na instituição é especialmente crucial para análise da evasão nos estágios iniciais do curso.

Por outro lado, a comparação entre as classes *"Evasão entre 2 e 4 anos"* e *"Evasão após 4 anos"* revela semelhanças na importância de *media_final* e *duracao_vinculo*, sugerindo que esses fatores continuam sendo relevantes à medida que os alunos avançam em sua trajetória acadêmica. No entanto, as diferenças na ordem de importância dessas variáveis podem indicar uma mudança na dinâmica da evasão à medida que o tempo

avança, com *media_final* sendo mais determinante para *"Evasão entre 2 e 4 anos"* e *duracao_vinculo* assumindo maior importância para *"Evasão após 4 anos"*.

Por fim, ao comparar as classes *"Evasão após 4 anos"* e *"Formação"*, observa-se uma convergência na importância de *ch_aprovacoes_acum* e *media_final*. No entanto, a diferença na ordem de importância das variáveis, com *duracao_vinculo* sendo mais relevante para *"Evasão após 4 anos"* e *ch_reprovacoes_acum* assumindo mais importância para *"Formação"*, destaca nuances nos fatores que influenciam a evasão e a formação.

6.4 Evasão precoce ou Não evasão precoce

A Tabela 27 apresenta a distribuição de frequências para cada classe empregada no quarto cenário. A Tabela 28 exibe informações estatísticas sobre as variáveis quantitativas utilizadas. As figuras 60 e 61 apresentam *boxplots* das variáveis em questão.

Tabela 27 – Distribuição de frequências por classe do quarto cenário. Fonte: o autor (2024).

Classe	**Quantidade de Ocorrências**
1. Evasão	4691
2. Não evasão	4691

Tabela 28 – Variáveis quantitativas utilizadas no quarto cenário. Fonte: o autor (2024).

Variável	**Min.**	**Máx.**	**Média**	**Desv. Pad.**	**Mediana**
duracao_vinculo	1	4	2	1,07	2
diferenca_conc_em_inicio_graduacao	1	50	5,63	6,04	4
media_final	0	10	3,88	3,39	4,08
ch_total	15	4170	378,66	269,78	330
ch_aprovacoes_acum	0	1800	400,87	410,75	300
ch_reprovacoes_acum	0	1500	212,61	204,62	180
ch_cancelamentos_acum	0	750	1,66	23,95	0

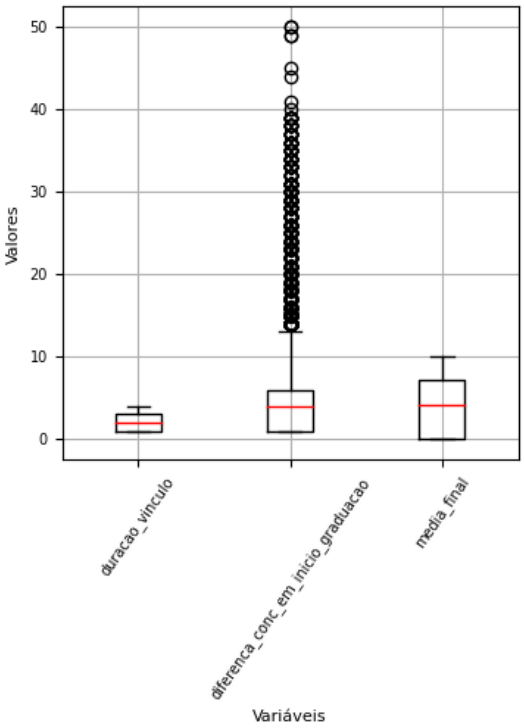

Figura 60 – *Boxplots* - Cen. 4 (parte 1).
Fonte: o autor (2024).

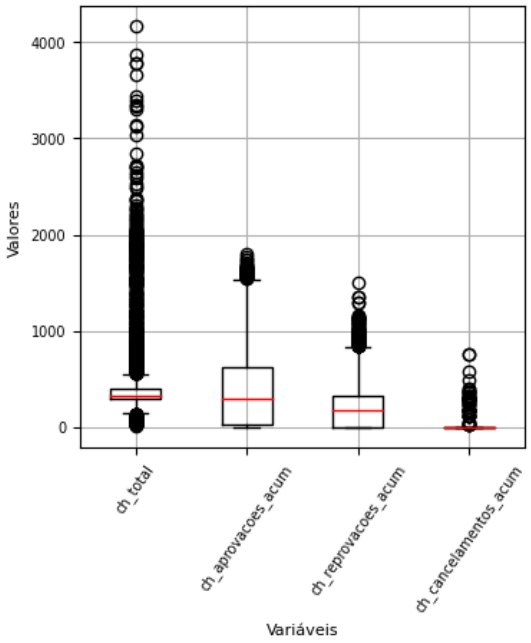

Figura 61 – *Boxplots* - Cen. 4 (parte 2).
Fonte: o autor (2024).

A Tabela 29 exibe o resumo dos resultados obtidos pelos modelos construídos com validação cruzada de 5 grupos (*5-fold CV*), além dos resultados do *ensemble* Voting Classifier construído com *hold-out* 70/30, observando as métricas de acurácia, precisão, *recall* e *f1-score*. Tratando dos modelos construídos com validação cruzada, os valores mostrados por classe indicam a média dos resultados obtidos para cada classe, por modelo, em cada iteração do processo.

As figuras 62, 64 e 66 mostram o resumo das matrizes de confusão dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 63, 65 e 67 exibem o resumo dos gráficos AUC/ROC dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 68 e 69 mostram, respectivamente, a matriz de confusão e o gráfico AUC/ROC do Voting Classifier.

Tabela 29 – Resultados dos modelos construídos no quarto cenário. Fonte: o autor (2024).

	5-fold CV									***Hold-out 70/30***		
	XGBoost			**Support Vector Machine**			**Logistic Regression**			**Voting Classifier**		
	Precisão	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***
1. Evasão	0,830479	0,822640	0,826454	0,848264	0,809639	0,828460	0,829827	0,827117	0,828439	0,844316	0,821705	0,832857
2. Não evasão	0,824304	0,831808	0,827951	0,817919	0,855045	0,836035	0,827710	0,830314	0,828979	0,823570	0,845989	0,834629
Média	0,827391	0,827224	0,827203	0,833091	0,832342	0,832248	0,828769	0,828716	0,828708	0,833943	0,833847	0,833743
Acurácia	0,827224			0,832340			0,828716			0,833748		

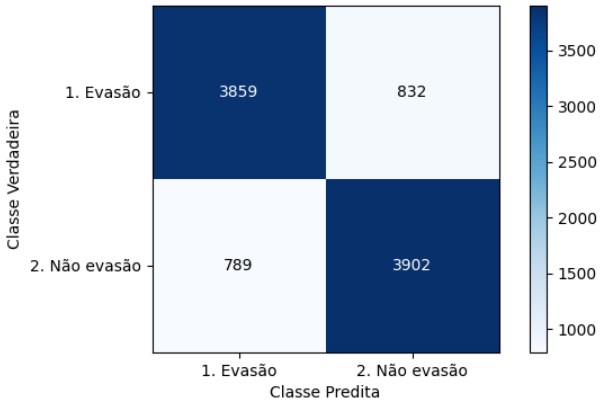

Figura 62 – Matriz de confusão - XGBoost - Cen. 4. Fonte: o autor (2024).

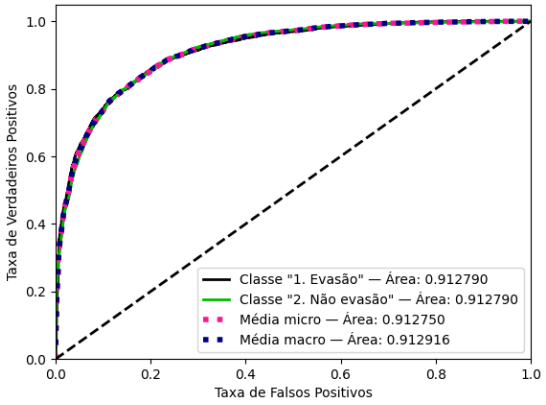

Figura 63 – AUC/ROC - XGBoost - Cen. 4. Fonte: o autor (2024).

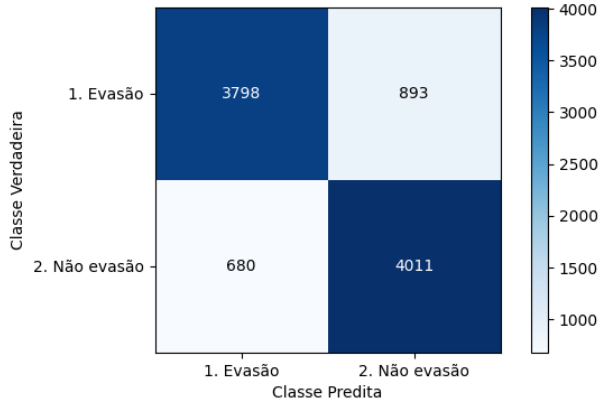

Figura 64 – Matriz de confusão - SVC - Cen. 4. Fonte: o autor (2024).

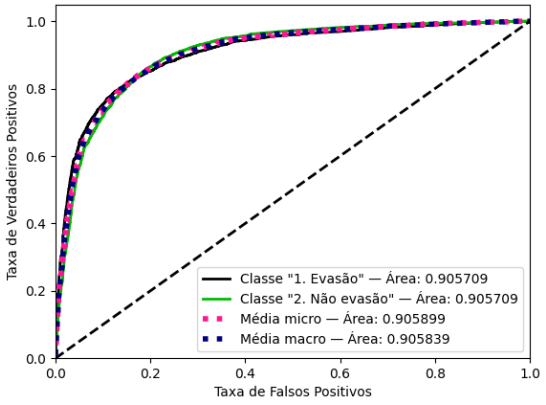

Figura 65 – AUC/ROC - SVC - Cen. 4. Fonte: o autor (2024).

Sobre os resultados dos modelos individuais, o XGBoost obteve uma acurácia de 82,72%, o que indica a proporção de previsões corretas em relação ao total de previsões. A precisão, *recall* e *f1-score* também foram de aproximadamente 82,74%, sugerindo um bom equilíbrio entre a capacidade do modelo de prever corretamente as instâncias de *"Evasão"* e *"Não evasão"*, considerando as análises de evasão precoce. O SVC obteve uma acurácia ligeiramente superior ao XGBoost, alcançando 83,23%. As métricas de precisão, *recall* e *f1-score* também foram consistentemente altas, em torno de 83,31%. O Logistic Regression teve um desempenho semelhante ao XGBoost e ao SVC, com uma acurácia de 82,87% e métricas de precisão, *recall* e *f1-score* em torno de 82,88%.

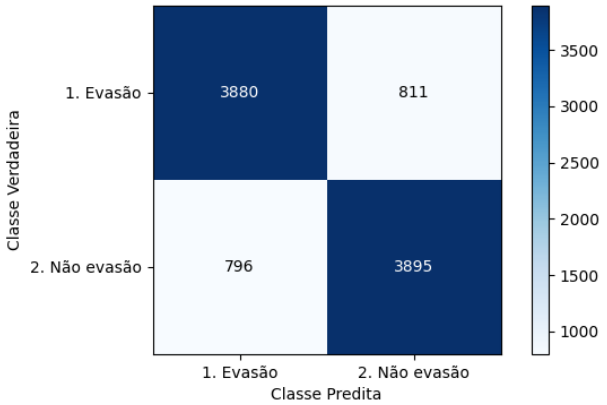

Figura 66 – Matriz de confusão - Log. Regress. - Cen. 4. Fonte: o autor (2024).

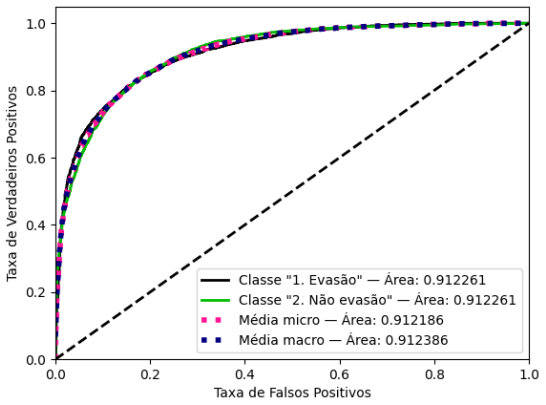

Figura 67 – AUC/ROC - Log. Regress. - Cen. 4. Fonte: o autor (2024).

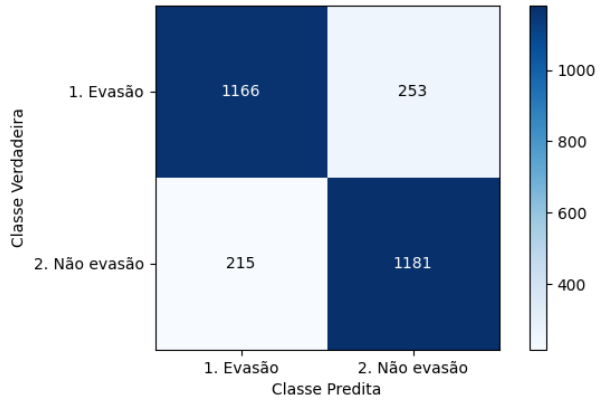

Figura 68 – Matriz de confusão - Voting Classifier - Cen. 4. Fonte: o autor (2024).

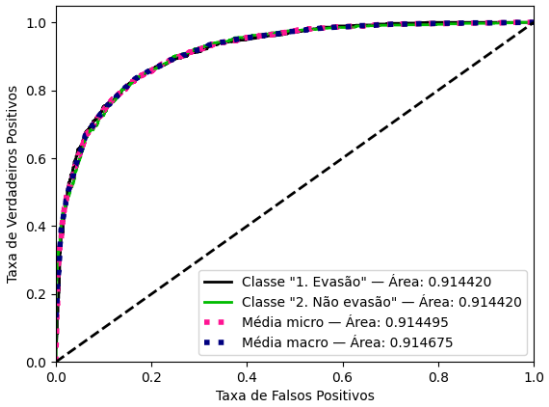

Figura 69 – AUC/ROC - Voting Classifier - Cen. 4. Fonte: o autor (2024).

Quanto aos resultados do *ensemble* Voting Classifier, para a classe *"Evasão"*, o modelo obteve uma precisão de 84,43%, indicando que a grande maioria dos exemplos classificados como evasão pelo modelo realmente pertencem a essa classe. Isso sugere que o modelo minimiza os falsos positivos, ou seja, casos em que prevê evasão quando não há. O *recall* foi de 82,17%, o que significa que o modelo capturou a maioria dos exemplos de *"Evasão"* presentes no conjunto de dados. Isso indica que o modelo é eficaz em identificar casos de evasão. O valor de *f1-score* para essa classe foi de 83,29%. Isso sugere um equilíbrio adequado entre a capacidade do modelo de identificar evasões e sua habilidade de evitar falsos positivos.

Para a classe *"Não evasão"*, o *ensemble* apresentou uma precisão de 82,36%, o que significa que a grande maioria dos exemplos classificados como *"Não evasão"* pelo modelo realmente pertencem a essa classe. Isso indica uma capacidade razoável de evitar classificar

casos de evasão quando não se deve. O *recall* foi de 84,60%, indicando que o modelo capturou a maioria dos exemplos de *"Não evasão"* presentes no conjunto de dados. Isso sugere que o modelo é eficaz em identificar casos em que os estudantes não evadiram. O valor de *f1-score* para a classe foi de 83,46%, sugerindo um equilíbrio entre a capacidade do modelo de identificar casos de não evasão e de evitar classificar erroneamente a evasão.

No geral, o *ensemble* apresenta um desempenho equilibrado para ambas as classes, com precisão e *recall* bastante próximos. Isso sugere que o *ensemble* é capaz de realizar uma classificação confiável tanto para *"Evasão"* quanto para *"Não evasão"* nas análises da evasão precoce. A Figura 70 exibe as 15 variáveis mais importantes para o *ensemble*, considerando as classes utilizadas. Como este cenário trata da construção de classificadores binários, a importância de cada variável é tipicamente a mesma para cada classe. Por meio da figura, nota-se que *media_final*, *ch_reprovacoes_acum*, *ch_aprovacoes_acum*, *duracao_vinculo* e *periodo_ingresso* se destacam como os principais fatores influenciadores para a evasão precoce, que ocorre nos dois primeiros anos do curso de graduação.

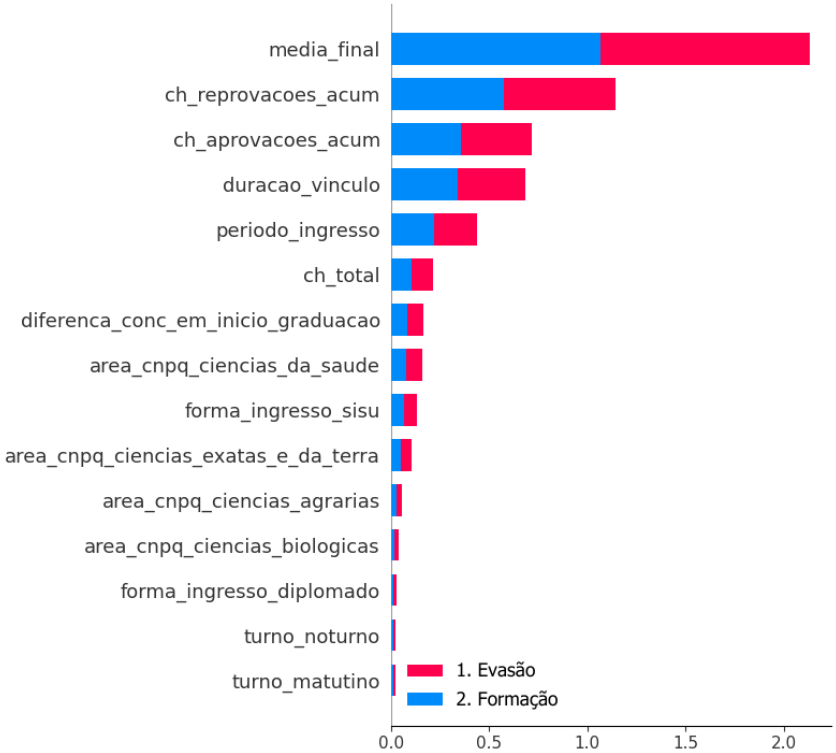

Figura 70 – Importância das variáveis - Voting Classifier - Cen. 4. Fonte: o autor (2024).

A presença de *media_final* como uma variável importante sugere que o desempenho acadêmico global continua sendo um preditor significativo da evasão, corroborando o que foi percebido nos cenários anteriores. Além disso, tanto *ch_reprovacoes_acum* quanto

ch_ aprovacoes_ acum indicam a importância do histórico acadêmico na tomada de decisão dos alunos sobre a continuidade de seus estudos.

A variável *duracao_ vinculo* ressalta a influência da quantidade de períodos de permanência na instituição para a evasão precoce, identificando que alunos com maior quantidade de períodos de vínculo podem estar mais propensos a completar seus estudos com sucesso. Por fim, a inclusão de *periodo_ ingresso* como uma variável importante pode indicar que fatores contextuais, como mudanças em políticas educacionais ao longo do ano, podem desempenhar um papel na decisão dos alunos de evadir ou permanecer no curso.

6.5 Evasão no primeiro ano, Evasão no segundo ano ou Não evasão até o segundo ano

A Tabela 30 apresenta a distribuição de frequências para cada classe empregada no quinto cenário. A Tabela 31 exibe informações estatísticas sobre as variáveis quantitativas utilizadas. As figuras 71 e 72 apresentam *boxplots* das variáveis em questão.

Tabela 30 – Distribuição de frequências por classe do quinto cenário. Fonte: o autor (2024).

Classe	**Quantidade de Ocorrências**
1. Evasão no primeiro ano	2234
2. Evasão no segundo ano	2234
3. Não evasão até o segundo ano	2234

Tabela 31 – Variáveis quantitativas utilizadas no quinto cenário. Fonte: o autor (2024).

Variável	**Min.**	**Máx.**	**Média**	**Desv. Pad.**	**Mediana**
duracao_vinculo	1	4	1,88	1,02	2
diferenca_conc_em_inicio_graduacao	1	50	5,75	6,02	5
media_final	0	9,83	3,22	3,34	2,15
ch_total	15	3870	367,87	249,22	330
ch_aprovacoes_acum	0	1785	323,83	379,43	210
ch_reprovacoes_acum	0	1500	241,55	204,79	240
ch_cancelamentos_acum	0	585	1,22	19,50	0

A Tabela 32 exibe o resumo dos resultados obtidos pelos modelos construídos com

validação cruzada de cinco grupos (*5-fold CV*), além dos resultados do *ensemble* Voting Classifier construído com *hold-out* 70/30, observando as métricas de acurácia, precisão, *recall* e *f1-score*. Para os modelos construídos com validação cruzada, os valores mostrados por classe indicam a média dos resultados obtidos para cada classe, por modelo, em cada iteração do processo.

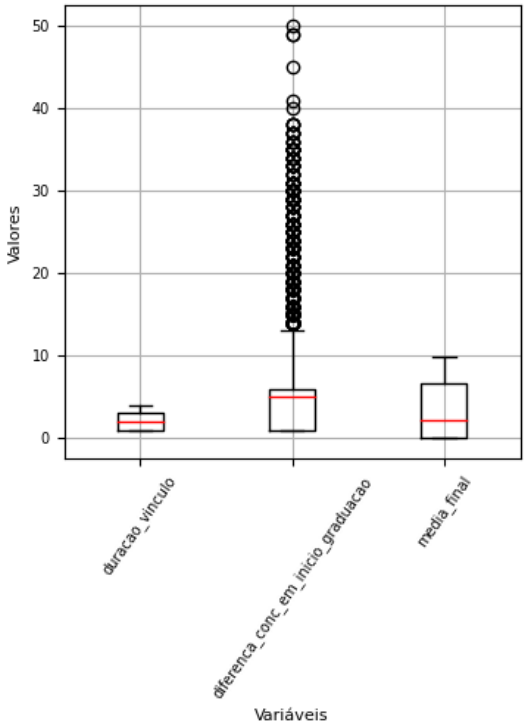

Figura 71 – *Boxplots* - Cen. 5 (parte 1).
Fonte: o autor (2024).

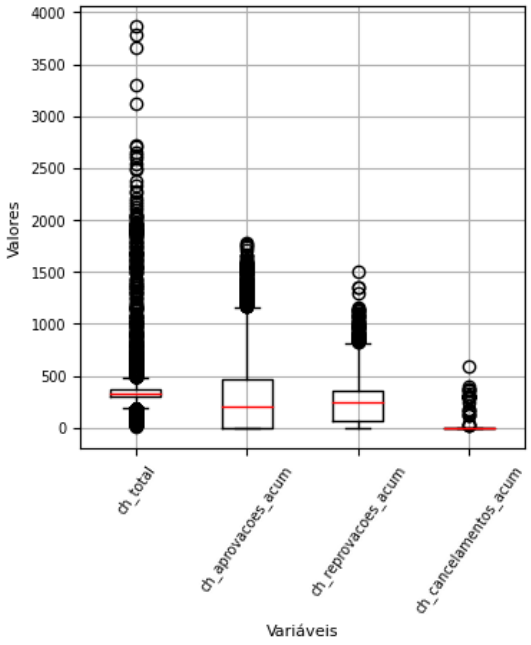

Figura 72 – *Boxplots* - Cen. 5 (parte 2).
Fonte: o autor (2024).

As figuras 73, 75 e 77 mostram o resumo das matrizes de confusão dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 74, 76 e 78 exibem o resumo dos gráficos AUC/ROC dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 79 e 80 mostram, respectivamente, a matriz de confusão e o gráfico AUC/ROC do Voting Classifier.

Tabela 32 – Resultados dos modelos construídos no quinto cenário. Fonte: o autor (2024).

	5-fold CV									***Hold-out 70/30***		
	XGBoost			**Support Vector Machine**			**Logistic Regression**			**Voting Classifier**		
	Precisão	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***
1. Evasão no primeiro ano	0,765689	0,794986	0,779895	0,774778	0,788714	0,781587	0,735898	0,794096	0,763632	0,787293	0,821326	0,803949
2. Evasão no segundo ano	0,648554	0,615044	0,631228	0,656918	0,597130	0,625194	0,657792	0,583242	0,617713	0,654514	0,576453	0,613008
3. Não evasão até o segundo ano	0,743037	0,752462	0,747633	0,730545	0,782458	0,755439	0,746533	0,770816	0,758287	0,722925	0,775264	0,748180
Média	0,719093	0,720831	0,719585	0,720747	0,722767	0,720740	0,713408	0,716051	0,713211	0,721577	0,724347	0,721713
Acurácia	0,720829			0,722770			0,716054			0,726504		

Tratando dos resultados dos modelos individuais, o XGBoost obteve uma acurácia de 72,08%, com precisão, *recall* e *f1-score* em torno de 71% a 72%. Isso indica uma performance consistente do modelo. Quanto ao SVC, observam-se resultados muito semelhantes aos do XGBoost. O SVC alcançou uma acurácia ligeiramente superior, atingindo 72,28%, com precisão, *recall* e *f1-score* em torno de 72%. O Logistic Regression apresentou resultados ligeiramente inferiores em comparação com os outros dois modelos. Com uma acurácia de 71,61% e precisão, *recall* e *f1-score* em torno de 71%, o Logistic Regression mostrou-se um pouco menos preciso.

Considerando os resultados do *ensemble* Voting Classifier, para a classe *"Evasão no primeiro ano"*, o modelo apresentou uma precisão de 78,73%, um *recall* de 82,13% e um *f1-score* de 80,39%. Isso indica que o modelo é capaz de identificar corretamente uma proporção significativa dos casos de evasão no primeiro ano, com uma precisão bastante alta em relação aos exemplos classificados como evasão.

Já para a classe *"Evasão no segundo ano"*, observa-se que o *ensemble* tem uma precisão de 65,45%, um *recall* de 57,65% e um *f1-score* de 61,30%. Esses números indicam que o modelo tem mais dificuldade em identificar corretamente os casos de evasão no segundo ano em comparação com o primeiro ano. A precisão é significativamente menor, o que sugere que uma proporção considerável dos exemplos classificados como evasão no segundo ano são, na verdade, falsos positivos. Além disso, o *recall* mostra que o modelo deixa de capturar uma parcela substancial dos casos reais de evasão no segundo ano.

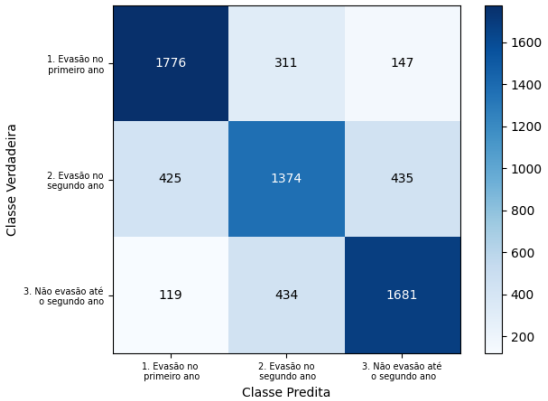

Figura 73 – Matriz de confusão - XGBoost - Cen. 5. Fonte: o autor (2024).

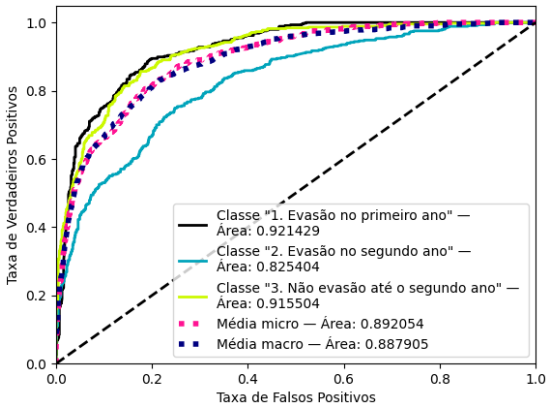

Figura 74 – AUC/ROC - XGBoost - Cen. 5. Fonte: o autor (2024).

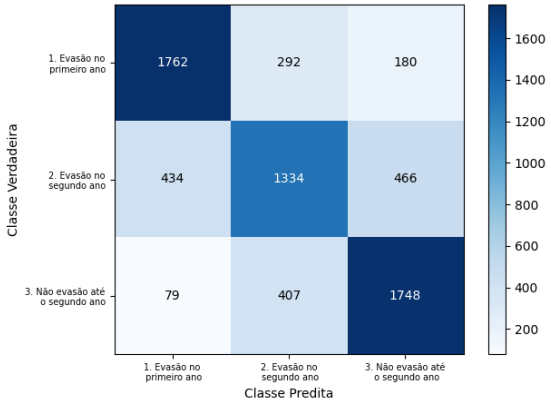

Figura 75 – Matriz de confusão - SVC - Cen. 5. Fonte: o autor (2024).

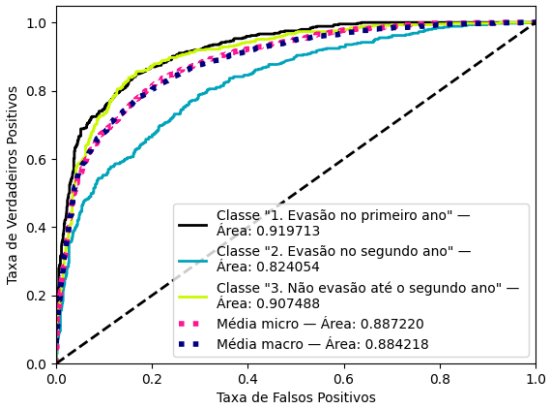

Figura 76 – AUC/ROC - SVC - Cen. 5. Fonte: o autor (2024).

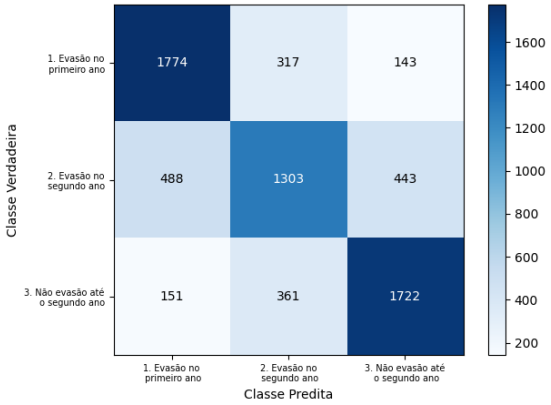

Figura 77 – Matriz de confusão - Log. Regress. - Cen. 5. Fonte: o autor (2024).

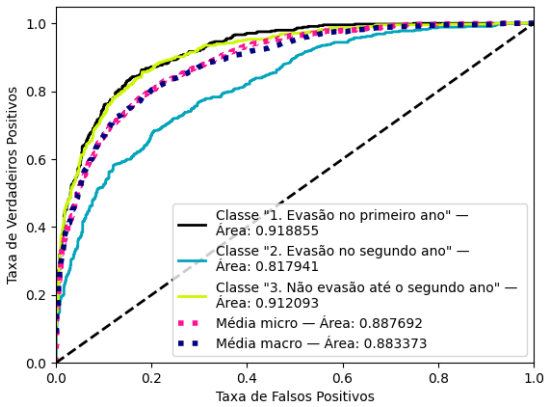

Figura 78 – AUC/ROC - Log. Regress. - Cen. 5. Fonte: o autor (2024).

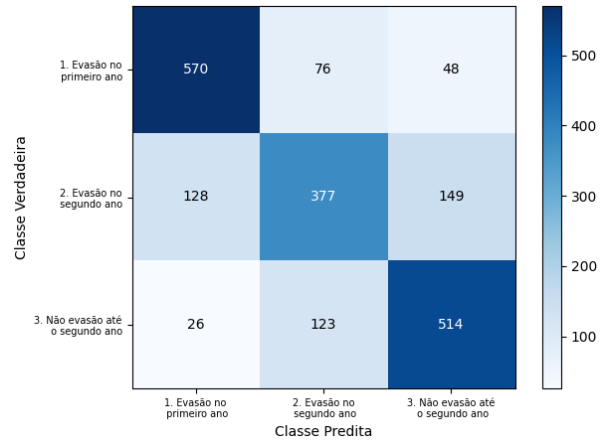

Figura 79 – Matriz de confusão - Voting Classifier - Cen. 5. Fonte: o autor (2024).

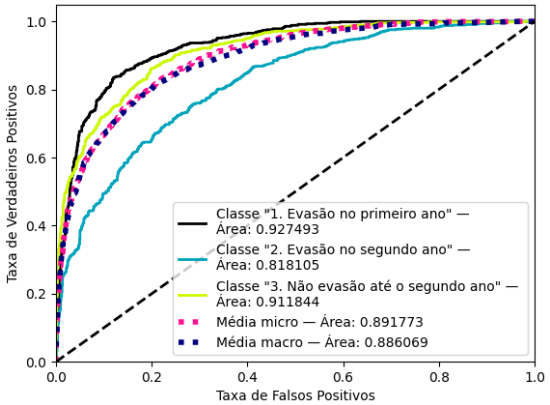

Figura 80 – AUC/ROC - Voting Classifier - Cen. 5. Fonte: o autor (2024).

Para a classe *"Não evasão até o segundo ano"*, o *ensemble* obteve uma precisão de

72,29%, um *recall* de 77,53% e um *f1-score* de 74,82%. Isso indica que o modelo é mais eficaz na identificação dos casos de não evasão até o segundo ano em comparação com os casos de evasão. A precisão é relativamente alta, sugerindo que a maioria dos exemplos classificados como não evasão até o segundo ano são corretos. O *recall* também é bastante elevado, o que indica que o modelo captura a grande maioria dos casos reais de não evasão até o segundo ano.

Ao considerar as médias macro das métricas de precisão, *recall* e *f1-score*, nota-se que o *ensemble* apresenta um desempenho geral equilibrado para as três classes, com pontuações em torno de 72%. Isso sugere que o modelo é capaz de realizar uma classificação razoavelmente precisa e abrangente para todas as classes consideradas. A acurácia geral do modelo foi de 72,65%, o que indica a proporção de predições corretas em relação ao total de predições. É válido ressaltar que se trata de um modelo multiclasse que objetiva prever a evasão precoce, que ocorre nos dois primeiros anos do curso, de forma detalhada, por ano, sendo esperado que haja maior dificuldade em classificar.

A Figura 81 apresenta as 15 variáveis mais importantes para o *ensemble*. Por meio da figura, percebe-se a importância de cada variável para cada classe considerada durante a construção dos classificadores. Para a classe *"Evasão no primeiro ano"*, as variáveis mais importantes foram *duracao_vinculo*, *media_final* e *periodo_ingresso*, nessa ordem. Para a classe *"Evasão no segundo ano"*, as variáveis mais importantes foram *ch_aprovacoes_acum*, *duracao_vinculo* e *media_final*, nessa ordem. Para a classe *"Não evasão até o segundo ano"*, as variáveis mais importantes foram *duracao_vinculo*, *media_final*, *ch_aprovacoes_acum* e *ch_reprovacoes_acum*, nessa ordem.

Primeiramente, a elevada importância de *duracao_vinculo* era esperada para todas as classes, observando que essa variável define, em quantidade de períodos, a duração da trajetória acadêmica do estudante. Além disso, notam-se *media_final* e *periodo_ingresso* como variáveis importantes para *"Evasão no primeiro ano"*, destacando a importância do desempenho acadêmico inicial e do momento de ingresso na instituição como preditores específicos para evasão no início do curso. Por fim, nota-se que a variável *ch_aprovacoes_acum* é mais relevante para *"Evasão no segundo ano"*, sugerindo que o histórico de aprovações acumuladas tem um papel crucial nessa fase específica do curso.

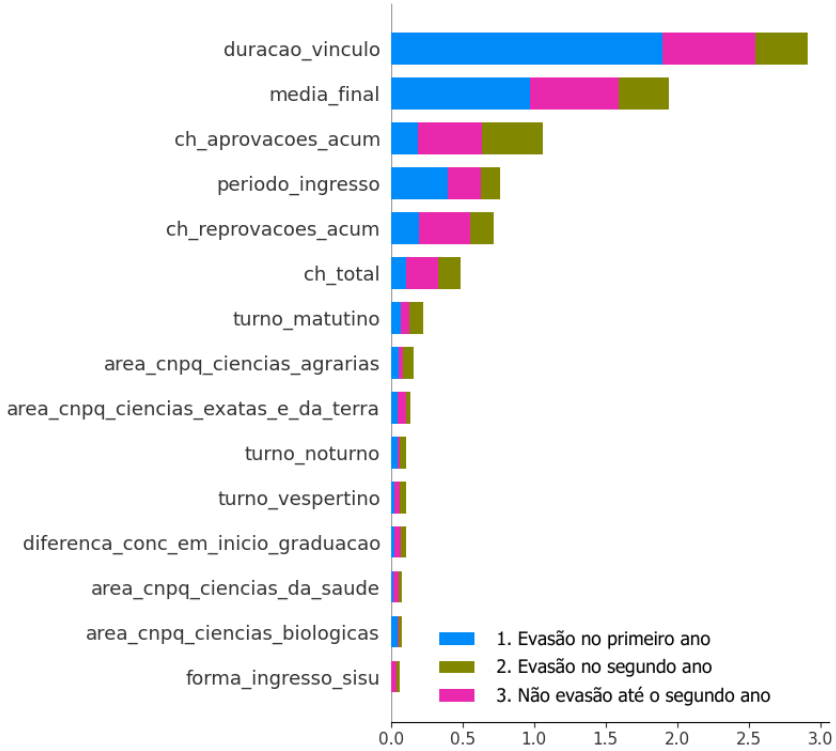

Figura 81 – Importância das variáveis - Voting Classifier - Cen. 5. Fonte: o autor (2024).

6.6 Evasão no período seguinte ou Não evasão no período seguinte

A Tabela 33 apresenta a distribuição de frequências para cada classe empregada no sexto cenário. A Tabela 34 exibe informações estatísticas sobre as variáveis quantitativas usadas. As figuras 82 e 83 têm *boxplots* das variáveis. Vale ressaltar que este cenário usa dados do primeiro período para a previsão da evasão ou da não evasão no segundo período.

A Tabela 35 exibe o resumo dos resultados obtidos pelos modelos construídos com validação cruzada de cinco grupos (*5-fold CV*), além dos resultados do *ensemble* Voting Classifier construído com *hold-out* 70/30, observando as métricas de acurácia, precisão, *recall* e *f1-score*. Para os modelos construídos com validação cruzada, os valores mostrados por classe indicam a média dos resultados obtidos para cada classe, por modelo, em cada iteração do processo.

Tabela 33 – Distribuição de frequências por classe do sexto cenário. Fonte: o autor (2024).

Classe	**Quantidade de Ocorrências**
1. Evasão	1141
2. Não evasão	1141

Tabela 34 – Variáveis quantitativas utilizadas no sexto cenário. Fonte: o autor (2024).

Variável	**Min.**	**Máx.**	**Média**	**Desv. Pad.**	**Mediana**
diferenca_conc_em_inicio_graduacao	1	44	5,95	6,05	5
media_final	0	9,38	2,71	3,25	0,71
ch_total	30	3420	443,57	371,91	330
ch_aprovacoes	0	570	124,28	146,76	30
ch_reprovacoes	0	480	194,02	145,27	240
ch_cancelamentos	0	180	0,08	3,78	0

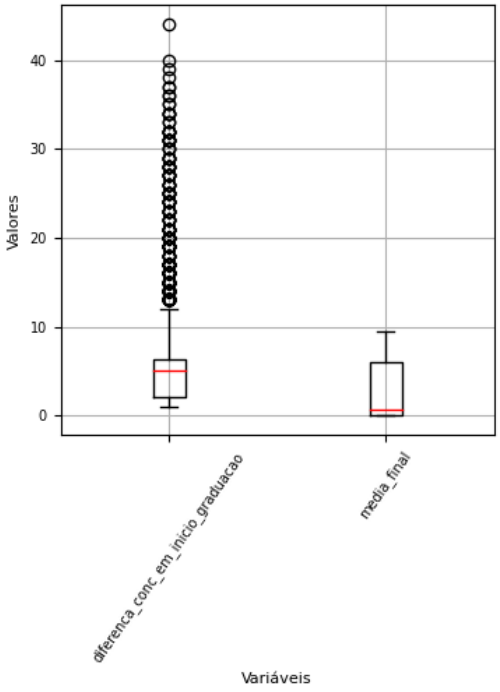

Figura 82 – *Boxplots* - Cen. 6 (parte 1).
Fonte: o autor (2024).

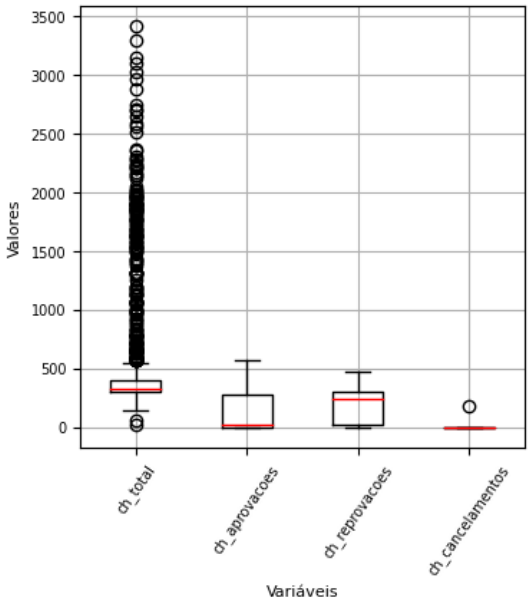

Figura 83 – *Boxplots* - Cen. 6 (parte 2).
Fonte: o autor (2024).

As figuras 84, 86 e 88 mostram as matrizes de confusão dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 85, 87 e 89 apresentam os gráficos AUC/ROC dos modelos XGBoost, SVC e Logistic Regression, respectivamente, após a validação cruzada. As figuras 90 e 91 mostram, respectivamente, a matriz de confusão e o gráfico AUC/ROC do Voting Classifier.

Tabela 35 – Resultados dos modelos construídos no sexto cenário. Fonte: o autor (2024).

	5-fold CV									***Hold-out 70/30***		
	XGBoost			**Support Vector Machine**			**Logistic Regression**			**Voting Classifier**		
	Precisão	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***	**Precisão**	***Recall***	***F1-score***
1. Evasão	0,821902	0,819440	0,820191	0,832608	0,847483	0,839760	0,810013	0,872907	0,839959	0,832432	0,865169	0,848485
2. Não evasão	0,820228	0,821194	0,820261	0,844966	0,829081	0,836714	0,862227	0,794005	0,826277	0,847619	0,811550	0,829193
Média	0,821065	0,820317	0,820226	0,838787	0,838282	0,838237	0,836120	0,833456	0,833118	0,840026	0,838359	0,838839
Acurácia	0,820332			0,838303			0,833478			0,839416		

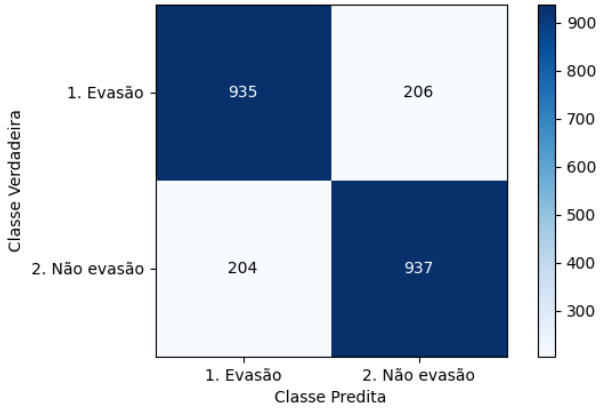

Figura 84 – Matriz de confusão - XGBoost - Cen. 6. Fonte: o autor (2024).

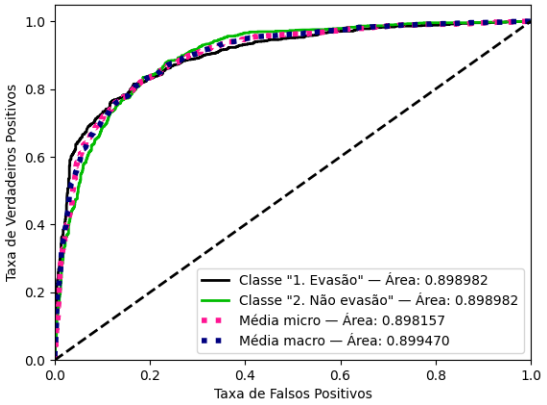

Figura 85 – AUC/ROC - XGBoost - Cen. 6. Fonte: o autor (2024).

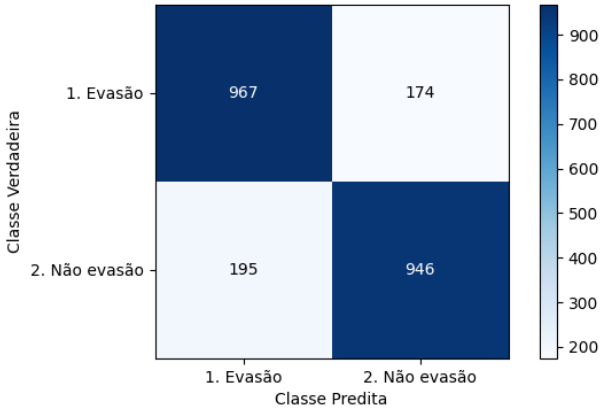

Figura 86 – Matriz de confusão - SVC - Cen. 6. Fonte: o autor (2024).

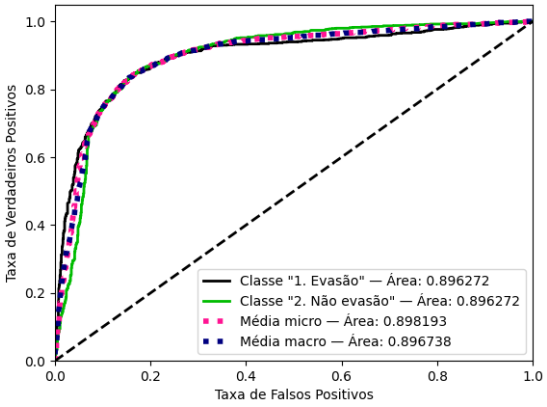

Figura 87 – AUC/ROC - SVC - Cen. 6. Fonte: o autor (2024).

Observando os resultados dos modelos individuais, o XGBoost obteve uma acurácia de 82,03%, com precisão, *recall* e *f1-score* em torno de 82%. Esses resultados indicam uma performance consistente e equilibrada do modelo na classificação das amostras. Tratando do SVC, observa-se uma acurácia ligeiramente superior em relação ao XGBoost, atingindo 83,83%. As métricas de precisão, *recall* e *f1-score* também giram em torno de 83%. Isso indica que o modelo SVC conseguiu realizar classificações precisas. Por fim, o Logistic Regression também apresentou resultados bastante competitivos, com uma acurácia de 83,35% e precisão, *recall* e *f1-score* em torno de 83%.

Tratando dos resultados do *ensemble* Voting Classifier, para a classe *"Evasão no semestre seguinte"*, o modelo obteve uma precisão de 83,24%, *recall* de 86,52% e *f1-score* de 84,85%, o que indica que o modelo conseguiu identificar corretamente a maioria dos casos de evasão previstos para o período seguinte (o segundo período do curso, uma vez

que, neste cenário, o ponto de partida da análise é o primeiro período), mantendo uma proporção alta de previsões corretas em relação ao total de exemplos classificados como evasão. A alta precisão e *recall* sugerem um desempenho robusto do modelo nessa classe.

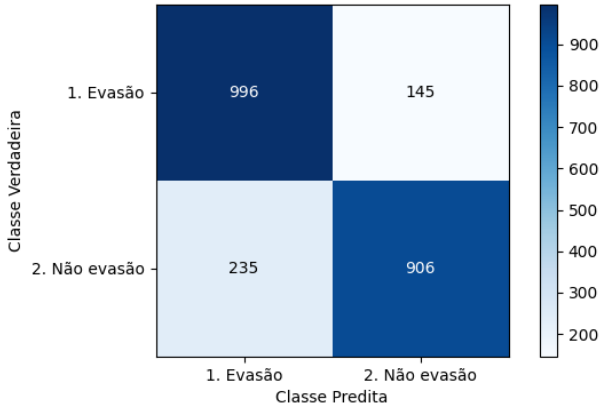

Figura 88 – Matriz de confusão - Log. Regress. - Cen. 6. Fonte: o autor (2024).

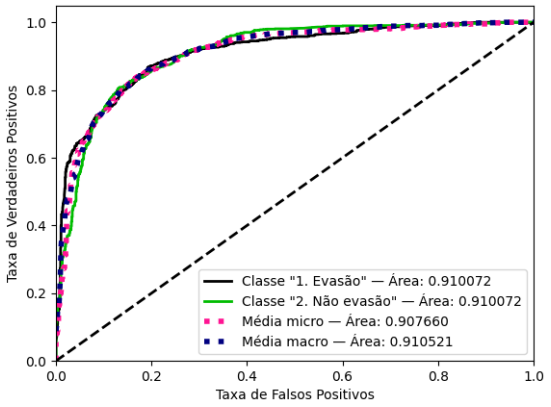

Figura 89 – AUC/ROC - Log. Regress. - Cen. 6. Fonte: o autor (2024).

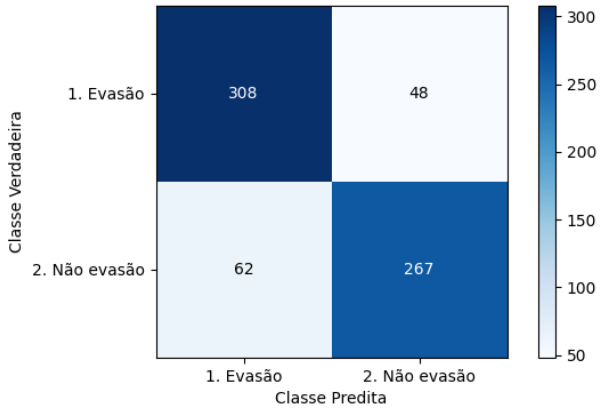

Figura 90 – Matriz de confusão - Voting Classifier - Cen. 6. Fonte: o autor (2024).

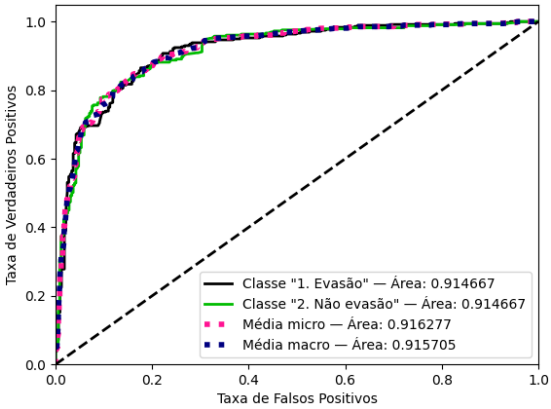

Figura 91 – AUC/ROC - Voting Classifier - Cen. 6. Fonte: o autor (2024).

Para a classe *"Não evasão no semestre seguinte"*, o modelo apresentou uma precisão de 84,76%, *recall* de 81,16% e *f1-score* de 82,92%. Isso indica que o modelo conseguiu identificar corretamente a maioria dos casos de não evasão previstos para o período seguinte (o segundo período do curso), mantendo uma proporção alta de previsões corretas em relação ao total de exemplos classificados como não evasão. Novamente, a alta precisão e *recall* sugerem um desempenho sólido do modelo nessa classe.

A média macro das métricas de precisão, *recall* e *f1-score* foi de aproximadamente 84%, indicando um bom desempenho geral do modelo. A acurácia geral do modelo foi

de 83,94%, indicando a proporção de previsões corretas em relação ao total de previsões. Esses resultados indicam que o *ensemble* obteve um desempenho bastante satisfatório na tarefa de previsão de evasão e não evasão no segundo período do curso com o uso de dados do primeiro período, tendo alta precisão, *recall* e *f1-score* para ambas as classes, além de boa acurácia.

A Figura 92 apresenta as 15 variáveis mais importantes para o *ensemble*, observando as classes utilizadas. Como este cenário trata da construção de classificadores binários, a importância de cada variável é tipicamente a mesma para cada classe. Por meio da figura, percebe-se que as variáveis *ch_ reprovacoes*, *media_ final*, *ch_ aprovacoes* e *periodo_ ingresso* foram as mais importantes para a previsão tanto da evasão quanto da não evasão no segundo período do curso.

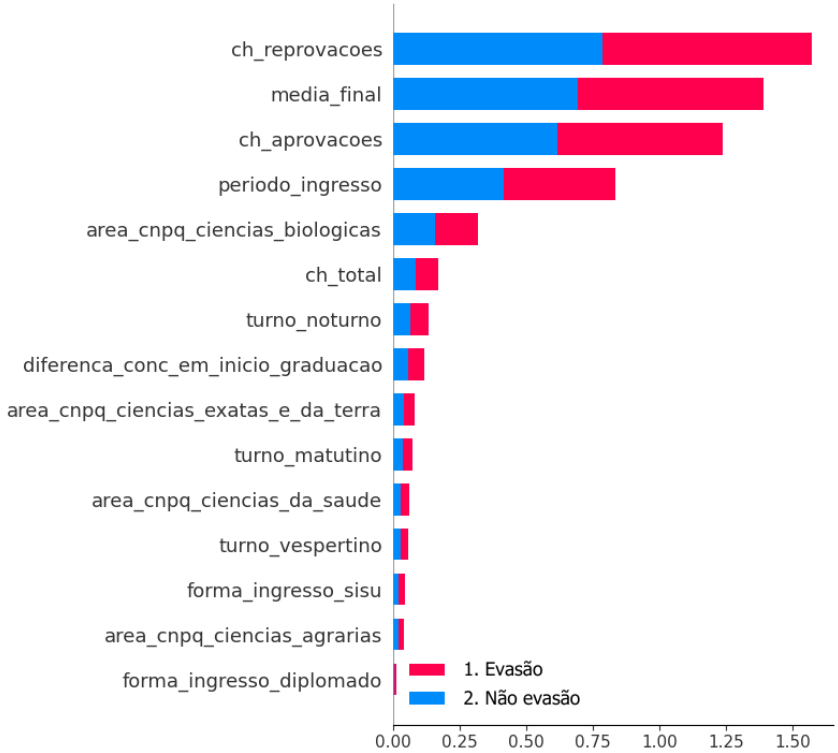

Figura 92 – Importância das variáveis - Voting Classifier - Cen. 6. Fonte: o autor (2024).

As variáveis *ch_ reprovacoes*, *media_ final*, *ch_ aprovacoes* e *periodo_ ingresso* foram identificadas como as mais importantes neste cenário. A presença de *ch_ reprovacoes* destaca a relevância da quantidade de reprovações no primeiro período do curso como um indicador significativo para prever a evasão no segundo período. Além disso, a presença de *ch_ aprovacoes* sugere que a quantidade de aprovações também desempenha um papel crucial na previsão da evasão. Consequentemente, *media_ final* é um fator importante,

refletindo a importância do desempenho acadêmico na decisão dos alunos de evadir ou permanecer no curso. Por fim, *periodo_ingresso* pode ressaltar a importância do momento em que o aluno ingressa na instituição como um fator influente na decisão de evasão. Esses resultados enfatizam a complexidade dos fatores que influenciam a evasão estudantil no início do curso de graduação e destacam a possibilidade de considerar uma variedade de variáveis que tratam do rendimento acadêmico ao desenvolver estratégias de retenção nos primeiros semestres do curso.

6.7 Discussão dos Resultados

Antes da discussão dos resultados de classificação, deve-se considerar a Questão de Pesquisa 2: *"É possível prever a evasão em momentos específicos do curso de graduação, considerando em conjunto amostras de estudantes de um grupo de cursos semelhantes e observando essencialmente o rendimento acadêmico, de forma a evitar o uso de informações socioeconômicas e a afastar vieses éticos?"*.

Este trabalho trata de construir modelos de classificação para a previsão da evasão estudantil no ensino superior por meio de seis cenários. Em cada cenário, há a construção de modelos com o uso dos algoritmos XGBoost, SVC e Logistic Regression, a partir do método de validação cruzada estratificada de cinco grupos. Além disso, há a construção de um modelo *ensemble* definitivo por cenário, observando os modelos construídos individualmente, com *hold-out* 70/30. A importância das variáveis utilizadas no treinamento do *ensemble* é observada com o emprego da técnica SHAP.

Os resultados dos modelos *ensemble* nos diferentes cenários são robustos e eficazes, apresentando altas métricas de precisão e *recall*. No primeiro cenário, o *f1-score* equilibrado para as classes *"Evasão"* e *"Formação"* demonstra um bom desempenho na identificação de ambas as situações. As variáveis mais importantes incluem *ch_aprovacoes_acum*, *media_final*, *ch_reprovacoes_acum*, *ch_total* e *duracao_vinculo*, refletindo aspectos críticos do desempenho acadêmico e trajetória do estudante. No segundo cenário, o modelo tem bom desempenho para *"Evasão até 4 anos"* e *"Formação"*, mas uma performance inferior para *"Evasão após 4 anos"*, com *recall* mais baixo, sugerindo dificuldade na identificação dessa classe, o que era esperado, pois se trata de um modelo multiclasse. As variáveis mais importantes permanecem consistentes, sendo as mesmas do cenário anterior,

destacando a relevância do desempenho acadêmico na previsão da evasão.

No terceiro cenário, o modelo apresenta desempenho variável entre as diferentes classes de evasão por período, sendo mais eficaz para *"Evasão até 2 anos"* e *"Formação"* e tendo dificuldades esperadas com *"Evasão entre 2 e 4 anos"* e *"Evasão após 4 anos"*. Considerando que o terceiro cenário tem um classificador multiclasse com quatro classes, para prever a evasão de forma detalhada, os resultados obtidos são satisfatórios. Variáveis como *duracao_vinculo*, *media_final*, *ch_aprovacoes_acum* e *ch_reprovacoes_acum* continuam a demonstrar a importância do desempenho acadêmico e quantidade de períodos cursados para a evasão. No quarto cenário, o desempenho é equilibrado para as classes *"Evasão"* e *"Não evasão"*, com precisão e *recall* próximos. As variáveis importantes incluem *media_final*, *ch_reprovacoes_acum*, *ch_aprovacoes_acum*, *duracao_vinculo* e *periodo_ingresso*, enfatizando o papel do desempenho acadêmico na previsão da evasão.

No quinto cenário, o modelo tem altos índices de precisão e *recall* para *"Evasão no primeiro ano"*, mas um desempenho menor para *"Evasão no segundo ano"*. Para *"Não evasão até o segundo ano"*, o desempenho é mais consistente, com as variáveis *duracao_vinculo*, *media_final*, *periodo_ingresso* e *ch_aprovacoes_acum* se destacando. No sexto cenário, o desempenho é alto para prever a evasão no segundo período do curso, com precisão e *recall* elevados para as classes *"Evasão no semestre seguinte"* e *"Não evasão no semestre seguinte"*. As variáveis mais importantes, *ch_reprovacoes*, *media_final*, *ch_aprovacoes* e *periodo_ingresso*, ressaltam a influência do desempenho acadêmico recente e histórico de reprovações na previsão de evasão precoce.

Observando os diferentes cenários propostos, nota-se que variáveis como *media_final*, *duracao_vinculo*, *ch_aprovacoes_acum* e *ch_reprovacoes_acum* são consistentemente importantes para a previsão da evasão, refletindo aspectos fundamentais do desempenho acadêmico e do engajamento dos estudantes. A variável *duracao_vinculo* é necessária para a análise proposta, pois marca o período do curso considerado, variando de 1 a 8 para um estudante que concluiu o curso em 8 períodos. Nota-se que essa variável se torna especialmente relevante em cenários que preveem a evasão em momentos específicos, como nas classes *"Evasão até 4 anos"* e *"Evasão após 4 anos"*. É válido considerar que a elevada importância da duração do vínculo para algumas classes de determinados cenários é justificada pelo fato de que, se um estudante possui registros em semestres posteriores ao oitavo, é evidente que ele não evadiu antes desse período.

Quanto aos resultados obtidos pelos modelos, as diferenças no desempenho entre os cenários podem indicar que alguns fatores são mais críticos em diferentes momentos do curso, sugerindo a necessidade de ajustes específicos para aprimorar a identificação da evasão em determinados momentos. Em linhas gerais, os modelos *ensemble* mostram-se eficazes, com capacidade de adaptação a diferentes contextos de classificação de evasão e formação estudantil no ensino superior, o que responde de forma afirmativa à Questão de Pesquisa 2.

7 Considerações Finais

A evasão estudantil no ensino superior é um fenômeno de grande relevância que impacta diretamente a qualidade da educação e a eficiência das instituições acadêmicas. Identificar e compreender os fatores que levam à evasão é fundamental para desenvolver estratégias que possam reter os estudantes e posteriormente aumentar a taxa de conclusão dos cursos. Nesse contexto, o uso de métodos do aprendizado de máquina supervisionado surge como uma abordagem promissora para a previsão da evasão. Este trabalho propõe o desenvolvimento de modelos de classificação com o objetivo de prever a evasão em momentos específicos do curso de graduação, tratando da evasão precoce e da evasão tardia, a partir de uma estratégia de aumento de amostras de estudantes evadidos com o agrupamento não arbitrário e não supervisionado de cursos de características semelhantes, o que diferencia o presente trabalho de diversos trabalhos relacionados.

Os trabalhos relacionados de previsão de evasão geralmente seguem uma de três estratégias de seleção de amostras de estudantes: consideram dados de estudantes de um único curso, de um conjunto arbitrário de cursos ou de todos os cursos de uma instituição. Cada estratégia apresenta desafios. O baixo volume de amostras de estudantes evadidos de um único curso costuma dificultar ou inviabilizar a construção de classificadores em alguns contextos. Além disso, a análise de um único curso pode comprometer a generalização e a construção de modelos. Por outro lado, selecionar arbitrariamente um grupo de cursos para considerar os dados dos estudantes em conjunto pode introduzir vieses e limitar a representatividade da amostra devido à dificuldade de compreender e interpretar as nuances de populações acadêmicas distintas sem o auxílio de técnicas não supervisionadas. Por fim, considerar todos os cursos da instituição simultaneamente pode impedir a captação de padrões específicos de cada curso pelos classificadores, o que não garante a eficácia e confiabilidade do uso de um mesmo classificador para todos os cursos observados.

O método proposto neste trabalho trata de agrupar cursos de graduação semelhantes por meio de técnicas do aprendizado de máquina não supervisionado, o que designa uma abordagem não arbitrária de agrupamento, com vistas a aumentar o volume de amostras de estudantes evadidos. O método sugere utilizar dados de estudantes de um grupo de cursos para treinar e avaliar modelos de classificação, com algoritmos do aprendizado de máquina supervisionado. Acredita-se que, com o agrupamento não supervisionado

de cursos e com a construção de classificadores para a previsão da evasão em momentos específicos, as instituições podem implementar políticas eficazes, proativas e personalizadas para mitigar a evasão, o que pode aprimorar a gestão educacional.

Este trabalho compreendeu o uso dos algoritmos K-Means e Aglomerativo para a geração não arbitrária de grupos de cursos. O uso de algoritmos distintos permitiu a validação dos resultados, fortalecendo a confiança nos grupos identificados quando houve convergência de soluções. Durante a execução do estudo de caso proposto, foram utilizados dados de 46 cursos presenciais de uma universidade pública brasileira, oriundos do Censo da Educação Superior de 2021. Os métodos convergiram para uma quantidade de grupos em torno de 5 ou 6, indicando coesão e separação adequadas com a consideração de diversas métricas de avaliação, como a inércia, o coeficiente de silhueta e os índices de Calinski-Harabasz, de Davies-Bouldin e de Dunn, além do dendograma. Ambos os métodos revelaram três grupos idênticos em seus resultados finais. Essa consistência sugere uma relação intrínseca entre os cursos agrupados, independentemente da técnica empregada. Esses padrões fortalecem a confiabilidade dos agrupamentos, destacando a semelhança entre os cursos selecionados. Portanto, acredita-se que a seleção de grupos para análise conjunta dos dados dos estudantes pode aumentar de forma consistente o volume de amostras de estudantes para a construção de modelos de previsão de evasão.

Na etapa de classificação, foram propostos seis cenários distintos para analisar a evasão estudantil, abordando tanto a evasão precoce quanto a evasão tardia. Os algoritmos XGBoost, Support Vector Classifier e Logistic Regression foram usados para a modelagem preditiva. Tratando do estudo de caso executado, um dos grupos gerados na etapa de agrupamento foi selecionado visando à consideração de dados de estudantes em conjunto para a construção de classificadores. A validação cruzada estratificada de cinco grupos foi utilizada para avaliar a robustez e a capacidade de generalização dos modelos. Em cada cenário, o balanceamento de classes das variáveis dependentes buscou assegurar que os modelos não favorecessem as classes majoritárias. Após a avaliação individual dos modelos com validação cruzada, houve a construção de um modelo *ensemble* definitivo com o Voting Classifier e votação *soft* em cada cenário, o que combinou os pontos fortes dos diferentes algoritmos, aumentando a precisão preditiva e a generalização dos modelos. Nos seis cenários de classificação propostos, o desempenho do *ensemble* mostrou-se robusto, considerando as diferentes classes de evasão. Conclui-se que os resultados obtidos pelos

modelos construídos são satisfatórios, e que eles podem ser usados para a previsão da evasão em momentos específicos do curso, abordando tanto a evasão precoce quanto a evasão tardia.

Os modelos desenvolvidos neste trabalho possuem o propósito de fornecer *insights* aos gestores de instituições de ensino superior. No âmbito das coordenações de curso, os modelos podem desempenhar um papel crucial ao identificar e categorizar cada estudante, permitindo uma abordagem individualizada no acompanhamento dos discentes. O uso dos modelos em sistemas de gestão acadêmica permite o monitoramento contínuo e em tempo real da situação acadêmica dos estudantes com vistas à previsão da evasão, buscando intervir de forma proativa. Com isso, os modelos podem ser utilizados para a priorização de ações preventivas contra a evasão, como a implementação de programas de assistência estudantil, a indicação de participação em projetos de pesquisa e extensão e monitorias, além de outras iniciativas pertinentes. Em resumo, os modelos podem contribuir para o desenvolvimento de políticas educacionais personalizadas e para aumentar a eficácia de estratégias de retenção estudantil. Este estudo integra a pesquisa realizada no Observatório de Dados da Graduação (ODG) da Universidade Federal Rural de Pernambuco (UFRPE). Em especial, os modelos propostos serão incorporados ao System of Academic Business Intelligence and Analytics (SABIA), plataforma de gestão acadêmica da UFRPE desenvolvida pelo ODG.

Tratando da interpretação das variáveis mais importantes para a previsão da evasão a partir da técnica SHAP, considerando os diferentes cenários analisados, as variáveis *media_final*, *duracao_vinculo*, *ch_aprovacoes_acum* e *ch_reprovacoes_acum* destacam-se consistentemente como cruciais na previsão da evasão e formação estudantil. As variáveis refletem aspectos fundamentais do desempenho acadêmico e do engajamento dos estudantes. A importância dessas variáveis permanece constante, apesar das variações no desempenho dos modelos, sugerindo que o desempenho acadêmico e a trajetória do vínculo são indicadores-chave para diferentes contextos de evasão e formação.

Para trabalhos futuros, sugere-se a identificação e a coleta de outras informações relevantes para compor as bases de dados consideradas para o agrupamento de cursos e para a construção de classificadores. Também é sugerido o estudo de novos métodos de agrupamento de cursos com características semelhantes, assim como a tunagem de hiperparâmetros dos classificadores de evasão. A execução do estudo em outras instituições de ensino, visando à comparação dos grupos gerados e à interpretação dos resultados dos

classificadores, também é estimulada. Além disso, sugere-se a construção de classificadores para grupos distintos de cursos objetivando a comparação e interpretação dos resultados em diferentes contextos. A investigação de novas técnicas de explicabilidade, como LIME ou técnicas baseadas em redes neurais, também pode aprimorar a interpretação dos modelos. Estudos voltados para a inteligência artificial confiável, incluindo a avaliação de métricas de justiça e a inclusão controlada de informações socioeconômicas de estudantes, são recomendados para o aprimoramento dos modelos, garantindo a equidade e a transparência, o que pode permitir uma análise mais abrangente das causas da evasão, abordando possíveis vieses e injustiças de forma efetiva e com responsabilidade.

Referências

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

BRASIL. Constituição da República Federativa do Brasil de 1988. **Senado Federal: Centro Gráfico**, 1988.

BRUCE, A.; BRUCE, P. **Estatística Prática para Cientistas de Dados**. Alta Books, 2019. ISBN 9788550810805. Disponível em: <https://books.google.com.br/books?id=b0mvDwAAQBAJ>.

CARVALHO, L.; SANTOS, A.; NAKAMURA, F.; OLIVEIRA, E. Detecção precoce de evasão em cursos de graduação presencial em Computação: um estudo preliminar. In: **Anais do XXVII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2019. p. 233–243. ISSN 2595-6175. Disponível em: <https://sol.sbc.org.br/index.php/wei/article/view/6632>.

CHAPLOT, D.; RHIM, E.; KIM, J. Predicting Student Attrition in MOOCs using Sentiment Analysis and Neural Networks. In: . [S.l.: s.n.], 2015. v. 1432.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. CRISP-DM 1.0: Step-by-step data mining guide. In: . [S.l.: s.n.], 2000.

COLPO, M. P.; PRIMO, T. T.; AGUIAR, M. S. d.; CECHINEL, C. Mineração de Dados Educacionais na Predição da Evasão Estudantil: Tendências, Oportunidades e Desafios. **Revista Brasileira de Informática na Educação**, v. 32, p. 220–256, maio 2024. Disponível em: <https://journals-sol.sbc.org.br/index.php/rbie/article/view/3559>.

COSTA, C. T. d.; JúNIOR, M. L. L.; SANTANA, A. M. União de Dados por Clusterização para Construção de Modelos de Predição de Evasão. **Revista Novas Tecnologias na Educação**, v. 21, n. 2, p. 393–402, dez. 2023. Disponível em: <https://seer.ufrgs.br/index.php/renote/article/view/137761>.

DIGIAMPIETRI, L. A.; NAKANO, F.; LAURETTO, M. d. S. Mineração de Dados para Identificação de Alunos com Alto Risco de Evasão: Um Estudo de Caso. **Revista de Graduação USP**, v. 1, n. 1, p. 17–23, jul. 2016. Disponível em: <https://www.revistas.usp.br/gradmais/article/view/117720>.

DUNN, J. C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Taylor & Francis, 1973.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. Disponível em: <http:

//www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1>.

INEP. **Microdados de Indicadores de Qualidade da Educação Superior**. 2021. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-qualidade-da-educacao-superior>. Acesso em 13 de agosto de 2020.

_____. **Microdados do Censo da Educação Superior**. 2021. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>. Acesso em 13 de agosto de 2020.

JHA, N.; GHERGULESCU, I.; MOLDOVAN, A.-N. OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques. In: . [S.l.: s.n.], 2019. p. 154–164.

JUNIOR, S. S.; PENHOLATO, J. P.; ERLER, J. d. S.; CARNEIRO, I. J.; CRISTINA, T. Proposição de Indicadores para o Corpo Discente e Análise de Agrupamentos Aplicada aos Cursos de Graduação da UFES. **Revista Gestão Universitária na América Latina - GUAL**, 2013. Disponível em: <https://www.redalyc.org/articulo.oa?id=319327519007>.

KAPPEL, L. T. e M. A. Aplicação de Técnicas de Aprendizado de Máquina para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. **Revista Brasileira de Informática na Educação**, v. 28, n. 0, p. 838–863, 2020. ISSN 2317-6121. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/v28p838>.

KIRCH, J. L.; SCHOENHERR, R. P.; VELOSO, T. C. M. A.; HONGYU, K. Application of Principal Component and Cluster Analysis to the Performance Indicators of Federal Universities in Brazil. **Sigmae**, v. 8, n. 2, p. 55–66, Jun. 2019. Disponível em: <https://publicacoes.unifal-mg.edu.br/revistas/index.php/sigmae/article/view/920>.

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.

LUNDBERG, S. M.; ERION, G.; CHEN, H.; DEGRAVE, A.; PRUTKIN, J. M.; NAIR, B.; KATZ, R.; HIMMELFARB, J.; BANSAL, N.; LEE, S.-I. From local explanations to global understanding with explainable AI for trees. **Nature machine intelligence**, Nature Publishing Group, v. 2, n. 1, p. 56–67, 2020.

LUNDBERG, S. M.; LEE, S.-I. A Unified Approach to Interpreting Model Predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems 30**. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

MARTINS, C.; LACERDA, F.; CARMO, I.; SILVA, E.; ALVES, T.; GOMES, J.; CAMPOS, R. Modelos de previsão de evasão tardia na graduação de uma universidade pública. In: **Anais do VIII Congresso sobre Tecnologias na Educação**. Porto Alegre, RS, Brasil: SBC, 2023. p. 41–50. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/ctrle/article/view/25782>.

MARTíNEZ-PLUMED, F.; CONTRERAS-OCHANDO, L.; FERRI, C.; ORALLO, J. H.; KULL, M.; LACHICHE, N.; QUINTANA, M. J. R.; FLACH, P. A. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. **IEEE Transactions on Knowledge and Data Engineering**, 2019.

MEC. **Documento Orientador para a Superação da Evasão e Retenção na Rede Federal de Educação Profissional, Científica e Tecnológica**. 2014. Ministério da Educação. <https://rgt.ifsp.edu.br/portal/arquivos/fixos/permanencia/Doc_Orientador_Evasao_Retencao_SETEC%20(1).pdf>. Acesso em abril de 2024.

MELO, J. R. F. d.; MEDEIROS, A. K. d. A. Early school dropout in distance higher education: An analysis according to data from the degree in letters course at IFPB. **Research, Society and Development**, v. 10, n. 7, p. e31510717230, Jun. 2021. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/17230>.

MESSALAS, A.; KANELLOPOULOS, Y.; MAKRIS, C. Model-agnostic interpretability with shapley values. In: IEEE. **2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)**. [S.l.], 2019. p. 1–7.

NIYOGISUBIZO, J.; LIAO, L.; NZIYUMVA, E.; MURWANASHYAKA, E.; NSHIMYUMUKIZA, P. C. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. **Computers and Education: Artificial Intelligence**, v. 3, p. 100066, 2022. ISSN 2666-920X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666920X22000212>.

OECD. **Education at a Glance 2023**. [s.n.], 2023. 476 p. Disponível em: <https://www.oecd-ilibrary.org/content/publication/e13bef63-en>.

SANTOS, C. H.; MARTINS, S.; PLASTINO, A. É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico? In: **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 792–802. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/18107>.

SARAIVA, D.; PEREIRA, S.; BRAGA, R.; OLIVEIRA, C. Análise de Agrupamentos para Caracterização de Indicadores de Evasão. In: **Anais do XXIX Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 238–247. ISSN 2595-6175. Disponível em: <https://sol.sbc.org.br/index.php/wei/article/view/15915>.

TCU. **Acórdão 506 de 2014 do Tribunal de Contas da União**. 2014. Ministério da Educação. <https://pesquisa.apps.tcu.gov.br/documento/acordao-completo/*/KEY%253AACORDAO-COMPLETO-1250021/DTRELEVANCIA%2520desc/0/sinonimos%253Dfalse>. Acesso em abril de 2024.

VELÁSQUEZ, J. V.; VÉLEZ, E. C.; GÓMEZ, S. G.; PORTILLA, K. G. **Determinantes de la Deserción Estudiantil en la Universidad de Antioquia**. [S.l.], 2003.

XGBOOST. **XGBoost Python API Reference**. 2021. XGBoost Read the Docs. <https://xgboost.readthedocs.io/en/latest/python/python_api.html>. Acesso em 6 de novembro de 2021.