UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS TRINDADE
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO ENGENHARIA DE PRODUÇÃO CIVIL

Arthur Barretto Leite de Barros

Título: DESENVOLVIMENTO DE UM MODELO DE PREVISÃO DO SUCESSO ACADÊMICO DE GRADUANDOS DE ENGENHARIA VISANDO O AUMENTO DE EFICIÊNCIA DA INSTITUIÇÃO DE ENSINO SUPERIOR

Florianópolis
2020

Arthur Barretto Leite de Barros

Título: DESENVOLVIMENTO DE UM MODELO DE PREVISÃO DO SUCESSO ACADÊMICO DE GRADUANDOS DE ENGENHARIA VISANDO O AUMENTO DE EFICIÊNCIA DA INSTITUIÇÃO DE ENSINO SUPERIOR

Trabalho Conclusão do Curso de Graduação em Engenharia de Produção Civil do Centro de Tecnologia da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel em Engenharia de Produção Civil
Orientador: Prof. Dr. Ricardo Faria Giglio

Florianópolis
2020

Ficha de identificação da obra

BARRETTO LEITE DE BARROS, ARTHUR
DESENVOLVIMENTO DE UM MODELO DE PREVISÃO DO SUCESSO ACADÊMICO DE GRADUANDOS DE ENGENHARIA VISANDO O AUMENTO DE EFICIÊNCIA DA INSTITUIÇÃO DE ENSINO SUPERIOR / ARTHUR BARRETTO LEITE DE BARROS ; orientador, RICARDO FARIA GIGLIO, 2020.
95 p. Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Civil, Florianópolis, 2020. Inclui referências. 1. Engenharia de Produção Civil. 2. Aprendizado de máquina. 3. Gestão do ensino. 4. Previsão de performance. I. FARIA GIGLIO, RICARDO. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Civil. III. Título.

Arthur Barretto Leite de Barros

Título: DESENVOLVIMENTO DE UM MODELO DE PREVISÃO DO SUCESSO ACADÊMICO DE GRADUANDOS DE ENGENHARIA VISANDO O AUMENTO DE EFICIÊNCIA DA INSTITUIÇÃO DE ENSINO SUPERIOR

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de "Engenheiro Civil com habilitação em Engenharia de Produção" e aprovado em sua forma final pelo Curso Engenharia de Produção Civil

Local, 29 de novembro de 2020.

_____

Prof. Dr. Guilherme Ernani Vieira
Coordenador do Curso

Banca Examinadora:

_____

Prof. Dr. Ricardo Faria Giglio
Orientador
Instituição UFSC

_____

Prof. Dr. Sérgio Peters
Avaliador
Instituição UFSC

_____

Prof. Dr. Eduardo Ferreira da Silva
Avaliador
Instituição UFSC

Este trabalho é dedicado aos meus colegas de classe e aos meus queridos pais.

AGRADECIMENTOS

A todos os quais suas ações e palavras refletiram na minha condição enquanto pessoa e cidadão, cada qual com sua forma e intensidade de relação.

“A pessoa conscientizada tem uma compreensão diferente da história e de seu papel nela. Recusa acomodar-se, mobiliza-se, organiza-se para mudar o mundo.” (FREIRE, 1994)

RESUMO

O Brasil apresenta historicamente, e o mesmo quadro não consta diferentemente na atualidade, indicadores bastante problemáticos no que tange à produtividade e desigualdade. Quando se fala de desenvolvimento econômico, a educação superior desempenha um papel fundamental. Dentro dela, destaca-se não a falta investimento, conforme se insistiu nas políticas públicas recentes, mas na boa alocação desses recursos. Direcionado à geração de ações de melhoria nas políticas adotadas, o presente trabalho desenvolveu uma ferramenta de previsão da performance dos alunos de engenharia da UFSC, capaz de prever com assertividade e confiabilidade o resultado dos alunos do CTC quanto à aprovação ou não nas disciplinas cursadas, para uma gestão mais efetiva do ensino. Por meio de técnicas de aprendizado de máquina, elaborou-se dois modelos de previsão com os algoritmos de florestas aleatórias e regressão logística. Pela grande diferença de funcionamento interno dos algoritmos, a leitura dos resultados de ambos os modelos permitiu uma discussão de diagnóstico referente às características que influenciam a performance dos alunos, favorecendo futuras investigações sobre o tema. Dois objetivos principais nortearam a elaboração e ajuste dos modelos: a identificação de casos críticos e o mapeamento de todas as reprovações. Em ambos os casos o modelo de florestas aleatórias teve um desempenho melhor, maior precisão e maior especificidade. Com precisão de 86%, o modelo se demonstrou capaz de prever casos críticos e, para o objetivo de previsão de todas as reprovações, o modelo indicou uma sensibilidade de 79% e 36% de falsos positivos, concluindo que ele pode ser útil para aplicação geral, mas não confiável para casos de discentes específicos. Dessa maneira, o ferramental produzido e apresentado neste trabalho gera resultados confiáveis e constituem um avanço na capacidade de leitura da performance acadêmica dos alunos de engenharia, além de um exemplo para demais projetos similares que possuem demandas ainda não preenchidas por várias universidades e escolas no cenário de educação brasileira.

Palavras-chave: Aprendizado de máquina. Gestão do ensino. Previsão de performance.

ABSTRACT

Brazil has historically struggled with respect to economic productivity and inequality. By looking at recent indicators one can come to the conclusion that both remain pressing problems to solve. In terms of economic development, superior education plays a key role. Public policy of the last couple of decades has aimed improving superior education through increasing investments, however the main obstacle has been the correct allocation of those investments. The project documented on this paper aims to aid the guidance of future policies through the creation of a model that can predict student performance as to assertively and reliably predict which students will pass or fail in the subjects they enroll as a tool for a better teaching management. Using machine learning methods, two prediction models have been developed, where one is a random forest model, and another is a logistic regression model. Both algorithms' inner mechanisms differ in such a way that permitted a discussion of the characteristics important for each pointing to possible performance diagnosis hypotheses that could be explored in future investigations. Two main objectives guided the development and adjustment of both models: the identification of critical cases and the mapping of all student failures. In both cases the random forest model has shown better results, higher precision, and specificity. With 86% precision, the model demonstrated itself capable of predicting critical cases, as for the identification of all failures it has indicated sensitivity of 79% and 36% false positive rate, concluding that it may be useful for general application, but not reliable for specific cases. Therefore, the toolset produced and presented in the present paper generates reliable results and constitute an advancement in the capacity of academic performance interpretation as well as an example for other similar demands not yet attended in other universities and schools in the Brazilian educational landscape.

Keywords: Machine Learning. Teaching management. Performance prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

OCDE Organização para Cooperação Econômica Europeia

SUMÁRIO

1 INTRODUÇÃO

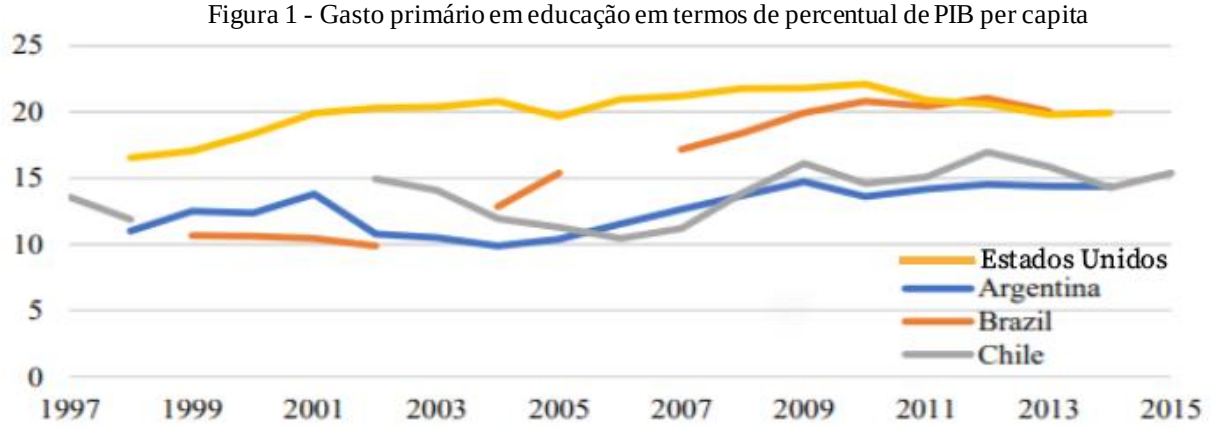
1.1 CONTEXTUALIZAÇÃO

O Brasil apresenta historicamente, e o mesmo quadro não consta diferentemente na atualidade, indicadores bastante problemáticos no que tange a produtividade e desigualdade. Um dos aparentes consensos na discussão de panorama mundial das últimas décadas é que o investimento e sucesso no aprimoramento da educação da população produz efeitos significativos de melhoria econômica. Conforme coloca Pieri (2018):

"A teoria do capital humano desenvolvida nos últimos quarenta anos estabelece haver uma relação entre investimentos individuais em educação e retornos futuros no mercado de trabalho. Essa teoria vem sendo empiricamente testada e diversos experimentos e quase-experimentos em diversas economias têm comprovado haver uma relação causal entre educação e sucesso no mercado de trabalho, medido como maiores salários, menor probabilidade e mais curtos períodos de desemprego."

Desde a década de 70, o governo brasileiro tem aumentado o investimento público na educação e assim gerado uma aceleração na democratização do ensino. Especificamente desde a virada do século XXI, desde o governo de Fernando Henrique Cardoso, o dispêndio médio por aluno cresceu muito e apresentou um salto bastante alto se comparado a países vizinhos. O gráfico a seguir ilustra tal afirmação:

Figura 1 - Gasto primário em educação em termos de percentual de PIB per capita

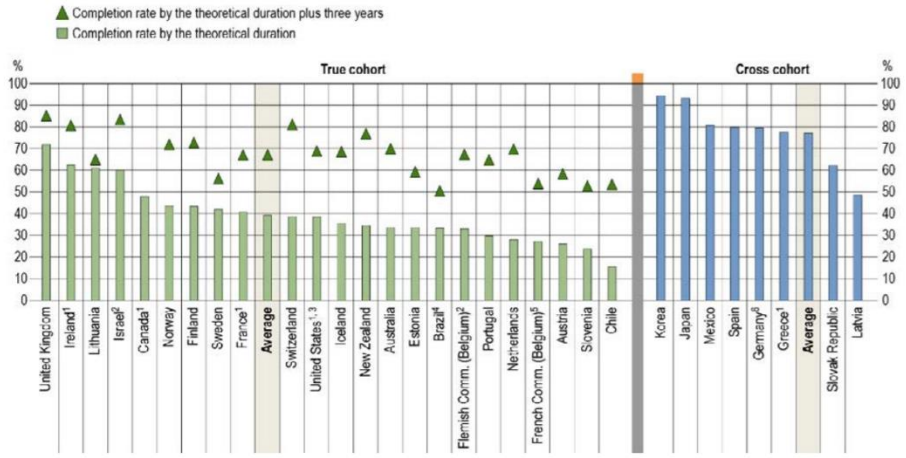


Fonte: (FEHER 2017)

Diferentemente do que se espera com esse grande aumento do investimento público na educação, a evolução dos níveis de produtividade, empregabilidade e desenvolvimento de pesquisa do país foi muito baixa. A maior parcela desse aumento de investimento nas últimas duas décadas foi concentrada na educação superior, onde existem grandes desafios a se enfrentar para que realmente se efetive em resultados visíveis em termos de nação. Alguns pontos relevantes foram evidenciados pelos comentários da pesquisa "*Education at a glance*" (OCDE, 2019):

1) Aproximadamente 33% dos alunos que se matriculam em uma graduação no Brasil se formam dentro do período estipulado, abaixo da média de 39% dos demais países do grupo. Após 3 anos adicionais, esse número sobe para 50%, ainda abaixo da média de 67%. O gráfico abaixo apresenta o percentual de alunos que se formam dentro do período estipulado para formatura assim como do período estipulado somado de 3 anos, para cada país no estudo, ilustrando o ponto a questão levantada:

Figura 2 - Taxa de formatura de estudantes de período integral que se matricularam em cursos de graduação (2017)



Fonte: (Education at a glance 2019)

2) Aumento de resultados na educação terciária demanda ambos amplificação de acesso quanto maiores taxas de finalização, ambos relativamente baixos no Brasil. Em 2017, apenas 15% dos jovens de 20-29 anos se encontravam matriculados na educação superior, comparados a 22% da média da OCDE.

Ao se tratar de resultados de educação terciária, um ponto relevante a ser considerado é o nível de produção tecnológica e científica do país, pilar importante para o

desenvolvimento econômico. É por meio do uso de pesquisas, produtos e inovações condizentes às necessidades ditadas pelo mercado que se consegue aprimorar seus processos e produção. Em outras palavras, a produtividade possui uma relação direta com o desenvolvimento tecnológico e científico. Quando se trata de Brasil, 95% da pesquisa científica provém das universidades públicas, onde os recursos humanos são disputados entre os âmbitos de ensino e pesquisa.

Conforme fica ilustrado no quadro apresentado acima, tem-se que, quando se fala de desenvolvimento econômico, a educação superior desempenha um papel fundamental. Dentro dela, destaca-se não a falta investimento, conforme se insistiu nas políticas públicas recentes, mas na boa alocação desses recursos, na qualidade desses investimentos para que se revertessem em melhores resultados. Dentro desse contexto, inúmeros trabalhos têm sido desenvolvidos nos âmbitos de identificação de causas, soluções e controles para contenção dos índices de evasão no país (Almeida, Godoy, 2016). Um dos problemas associados a esse tipo de abordagem, é que a pesquisa sobre evasão ainda é bastante divergente. Para citar dois motivos recorrentes, têm-se que existem variadas conceituações possíveis de serem aplicadas, além de ser um fenômeno que varia muito conforme a especificidade do grupo de estudo. Seja por se considerar diferentes cursos, instituições ou regiões.

Com um viés mais quantitativo e principalmente, mais direcionado à geração de ações de melhoria nas políticas adotadas dentro das instituições, alguns autores têm investido no desenvolvimento de ferramentas que possam mensurar e direcionar os esforços dos gestores enquanto se observa o reflexo das ações no desempenho dos alunos em seus respectivos cursos. Artigos têm sido publicados a fim de realizar previsões de sucesso dos alunos na finalização da graduação ou mesmo do seu sucesso nas matérias cursadas (SHINYASHIKI, BACCARO, 2014; BUENAÑO-FERNÁNDEZ, GIL, LUJÁN-MORA, 2019; ABUT, YUKSEL, AKAY, DANESHVAR, 2018; CASEY, AZCONA, 2017) e pretendem o desenvolvimento de um ferramental para que os gestores possam diagnosticar, avaliar e atuar, até mesmo prever e simular melhorias no desempenho dos alunos dentro do sistema educacional.

É nesse contexto que se insere o presente trabalho, tendo como objetivo o desenvolvimento de um ferramental construído especificamente para os cursos de engenharia da Universidade Federal de Santa Catarina (UFSC). Têm-se também a intenção de que a

proposição metodológica adotada no presente trabalho possa ser testada em outros cursos e instituições a fim de contribuir com os avanços já apresentados no tópico.

A literatura aponta que uma das melhores fontes de dados para a previsão de sucesso dos alunos em uma determinada matéria são os resultados obtidos ao decorrer da disciplina. Dados como os acessos de alunos ao material disponibilizado no Ambiente Virtual de Aprendizagem (AVA) e exercícios executados na plataforma, conferem ao modelo uma adaptabilidade mais real sobre o aluno na disciplina específica, proporcionando uma precisão muito elevada. Essa é a abordagem utilizada por CASEY e AZCONA (2017), ao utilizarem uma abordagem pedagógica de utilização e assimilação de conceitos ao decorrer do semestre.

1.2 JUSTIFICATIVA

O tópico de contextualização enfatizou a importância do avanço da educação superior quanto a seus índices de eficiência. OLIVEIRA (2016) quantificou quanto seria o desperdício por aluno causado pela evasão em determinadas Instituições Federais de Ensino Superior (IFES) brasileiras consideradas de boa qualidade de ensino e chegou ao número de 45% dos gastos por aluno. Esse número demonstra como o sistema de ensino vigente não é sustentável.

Como tentativa de compartimentalização do tema a fim de poder abordá-lo, coloca-se que dois dos maiores indicadores para mensurar a eficiência de um curso de graduação são o sucesso dos alunos nas matérias, pois define a sua taxa de evolução em direção ao diploma, e a taxa de evasão. Muitos autores associam o mal desempenho como uma das maiores causas das altas taxas de evasão, principalmente nos cursos de engenharia (Almeida, Godoy, 2016). Assim, pode-se dizer que ao se endereçar o problema de desempenho dos alunos da graduação de engenharia, serão direcionados esforços que atuarão direta ou indiretamente sobre dois dos principais problemas de eficiência nesses respectivos cursos.

Tratando-se especificamente dos cursos do Centro Tecnológico (CTC) da UFSC, centro dos cursos de engenharia, existe uma dificuldade de mensuração e por conseguinte de acompanhamento, dos níveis de eficiência e desempenho dos alunos na graduação. Muitos dados são coletados, mas não são feitas análises voltadas à tomada de decisão, o que dificulta o desenho e acompanhamento de políticas que sejam efetivas para lidar com os problemas

enfrentados no cotidiano. Apesar disso, a quantidade de dados armazenados no sistema é abundante.

Uma das maneiras de prover um maior esclarecimento sobre o tema é a previsão de performance de alunos. Ela permite com que se identifique locais sensíveis a determinação das próprias variáveis relevantes para a previsão de performance são parâmetros para a avaliação da situação atual. Conforme apontam Belachew e Gobena (2017):

"*Prever a performance futura dos alunos com acurácia baseado em seu histórico acadêmico é crucial para a aplicação efetiva de intervenções pedagógicas para se assegurar a finalização do curso dos alunos em no tempo previsto e de maneira satisfatória"*

Nessa linha, corroborado pelo interesse demonstrado pela diretoria do centro, o presente trabalho visa o desenvolvimento de uma ferramenta de previsão da performance dos alunos de engenharia para a gestão do ensino. A ferramenta foi desenvolvida em conjunto e aliada aos interesses do grupo de gestão e se propõe a orientar e interpretar os resultados obtidos pelo modelo produzido.

1.3 OBJETIVOS

Nas seções abaixo estão descritos o objetivo geral e os objetivos específicos deste TCC.

1.3.1 Objetivo Geral

O objetivo geral do presente trabalho é desenvolver um modelo de previsão que seja capaz de prever com assertividade e confiabilidade o resultado dos alunos do CTC quanto à aprovação ou não nas disciplinas cursadas a fim de que sua utilização abra a possibilidade de ações que melhorem a eficiência dos cursos em questão.

1.3.2 Objetivos Específicos

Com o propósito de se chegar ao objetivo geral, são propostos objetivos específicos que devem ser alcançados no processo:

1) Limpar a base de dados e ajustar variáveis existentes para que estejam adaptadas aos modelos
2) Criar variáveis que aumentem a capacidade de previsão do modelo
3) Aplicar modelos de previsão para prever os resultados de aprovação dos alunos do CTC
4) Avaliar, parametrizar e selecionar os modelos de acordo com as métricas adequadas
5) Discutir as características mais relevantes do modelo em torno do que significa em termos de leitura do desempenho acadêmico.
6) Propor uma utilização do modelo desenvolvido para diagnosticar e direcionar esforços voltados a um aumento de eficiência do centro

2 FUNDAMENTAÇÃO TEÓRICA

O presente trabalho procura desenvolver um modelo de previsão do sucesso dos alunos nas matérias cursadas a partir de seus dados históricos e socioeconômicos. Para fins de confiabilidade e assertividade do modelo, foram escolhidas técnicas disponíveis no leque de ferramentas do aprendizado de máquina. Assim, para possibilitar o embasamento necessário às etapas subsequentes, será feita uma explanação sobre os conceitos do aprendizado de máquina, das formas de desenvolvimento dos modelos de previsão e das formas de avaliação e seleção dos modelos desenvolvidos.

2.1 APRENDIZADO DE MÁQUINA

De acordo com Arthur Samuel (1959), a área de aprendizado de máquina é o "campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados". Se trata de uma área proveniente da engenharia e das ciências da computação que foi incorporada como subcampo da inteligência artificial. Mais especificamente, se trata de uma nova abordagem dentro do estudo de estatística que permite o aproveitamento de bases de dados de forma a tornar análises e tomadas de decisão mais assertivas, úteis e eficientes. De acordo com (SARKAR, 2018) a necessidade de se utilizar o aprendizado de máquina se resume no seguinte: "Para se fazer decisões guiadas por dados em escala".

O aprendizado de máquina se ocupa do estudo de reconhecimento de padrões e desenvolvimento de previsões a partir de dados. Conforme o modelo se desenha, ele evidencia características sobre a forma como a realidade se estrutura a partir de resultados empíricos que pode se provar interessante para a indução de características de maneira mais relevante do que a criação de uma teoria específica que se aplique no contexto específico estudado. Conforme aponta Ofori (2020):

"O aprendizado de máquina oferece uma vantagem em relação a formas tradicionais de análises estatísticas, colocando ênfase em performance preditiva acima de propriedades teóricas e pressupostos da população a priori."

Os problemas de aprendizado de máquina podem ser divididos em 2 tipos: o aprendizado supervisionado e não-supervisionado. Para ambos se exige a definição de um

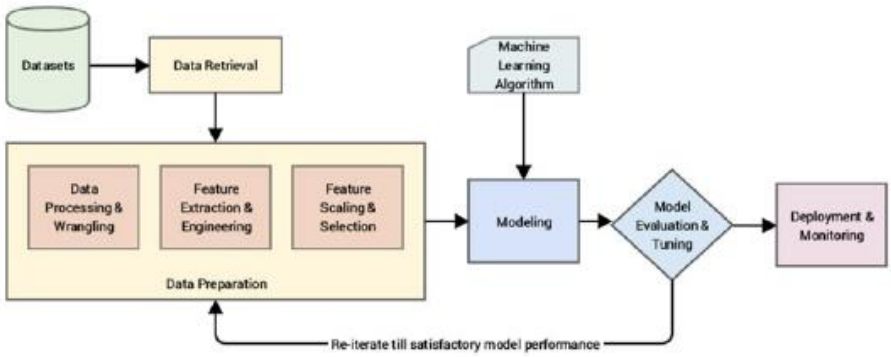
problema a ser investigado onde se buscará encontrar padrões que auxiliem a compreender, interpretar ou prever os resultados do problema definido a partir de do comportamento e inter-relações dos dados. A diferença entre eles consiste no fato de que no primeiro, aprendizado supervisionado, já se possui a classificação ou valor esperado, ou seja, os resultados da amostra da base de dados, enquanto o segundo, aprendizado não-supervisionado, pretende fazer a análise sem uma definição prévia dos resultados da amostra.

De acordo com a literatura (FRIEDMAN; TIBSHIRANI; HASTIE, 2001), as variáveis independentes, utilizadas como os inputs do modelo, serão chamadas no presente trabalho de preditores, *features* ou variáveis independentes, enquanto os resultados, outputs do modelo, serão denominados de respostas, variáveis dependentes ou preditas.

2.1.1 Pipeline de aprendizado de máquina

SARKAR (2018) define um esquema genérico de passos para a aplicação do aprendizado de máquina:

Figura 3 - Pipeline de modelagem de aprendizado de máquina



Fonte: Adaptado de SARKAR (2018)

SARKAR (2018) divide a sua aplicação nos seguintes passos:

- Coleta de dados: extração e geração de dados relevantes ao problema a ser resolvido.
- Preparação de dados: preparação dos dados para a sua inserção no modelo. Passa desde o entendimento dos dados por meio de análise exploratória à transformação dos dados para melhoria do modelo.

  - Processamento de dados: entendimento das características dos dados, limpeza e processamento dos dados em termos de alinhamento entre as diferentes *features* e sua usabilidade no modelo.
  - *Feature engineering*: criação de *features* ou atributos explícitos ou não na base de dados.
  - *Feature selection*: Deve ocorrer um processo de seleção das *features* que são úteis e aprimorem o resultado para um modelo mais adequado.

- Modelagem: inserção das *features* para treinamento do modelo. Aplica-se assim o modelo para resolver um objetivo específico como reduzir os erros nas bases de treino e proporcionar boa generalização dos dados.

- Avaliação e tunagem do modelo: avaliação do modelo por meio de métricas e comportamento do modelo em bases de validação. Um segundo momento é a regulagem de parâmetros não mais dos dados, mas do modelo em si, chamados de hiperparâmetros, para chegar ao modelo mais adequado em termos de resultados de aplicação.

- Implementação em monitoramento: modelos selecionados são aplicados em produção e monitorados a partir de suas predições e resultados.

Conforme mencionado na introdução, a UFSC possui uma base de dados com muita informação bastante subutilizada. Este trabalho tem seu ponto de partida na união das bases de dados de vestibular, cadastro e histórico escolar dos alunos. Essa fase de coleta foi realizada com apoio do SETIC (centro de informática).

O presente trabalho se propõe a gerar o modelo e propor um plano para sua aplicação. Dessa maneira, a etapa final de implementação não será abordada.

2.1.2 Preparação de dados

Modelos de aprendizado de máquina são desenvolvidos a partir do uso de dados. Esse termo inclui, porém, uma grande variedade de elementos. Existem várias formas de se

armazenar dados, cada qual com sua forma de coletar e organizar os mesmos. Sendo assim, dados crus, provenientes diretamente das bases de informação, não estão em formatos adequados para resolver um problema específico. Além disso, as bases de dados são alimentadas com uma série de falhas de armazenamento, organização e integração que fazem com que existam lacunas ou demais questões que precisam ser resolvidas. Dessa maneira, existe uma macro etapa de processamento dos dados frequentemente denominada como um todo de *feature engineering*. Englobados no conceito se encontram as atividades de limpeza e ajuste dos atributos, ou pré-processamento de dados; o processo de extração ou criação de novas *features* com valor de fornecer maior capacidade de previsão ao modelo; seleção das melhores *features* com maior valor para o desenvolvimento do modelo, ou seja, descarte das *features* que não contribuem para a melhoria do modelo. Diversos autores apontam que essa é a parte mais dispendiosa do trabalho com aprendizado de máquina. Conforme consta em Sarkar (2018):

"Coming up with features is difficult, time-consuming, requires expert knowledge. Applied machine learning" is basically feature engineering."

Andrew Ng

Por motivos didáticos, essa etapa do presente trabalho vai se referir para *feature engineering* apenas a parte de criação e refinamento de *features*. A macro etapa será dividida nas 3 atividades que o compõem, denominadas de pré-processamento, *feature engineering* e *feature selection*.

2.1.3 Trade-off viés-variância

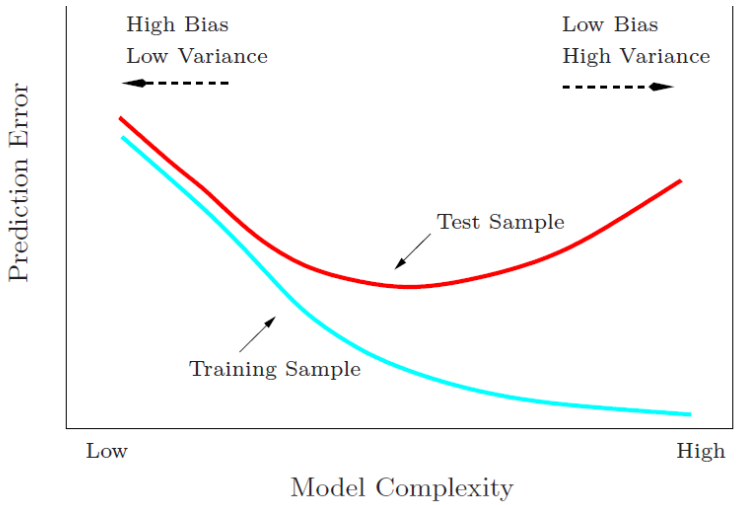
Ao se conduzir análises preditivas a partir de uma amostra de dados contendo os resultados desejados, uma das formas de avaliá-lo seria pela distância entre os valores previstos pelo mesmo e os valores de uma dada base de dados.

A qualificação do modelo pode, portanto, ser avaliada pela distância de suas previsões e a base de dados analisada. Agora, o objetivo do desenvolvimento de modelos de previsão recai sobre a sua capacidade de reproduzir resultados semelhantes em diferentes situações, contextos e por conseguinte, bases de dados. Essa capacidade se traduz melhor dizendo no quanto o modelo é generalizável para o resto da população.

Com fins de compartimentalizar essa questão, podemos utilizar dois conceitos importantes: o conceito de viés, que representa o quanto o modelo consegue captar as diferenciações e padrões internamente na amostra de desenvolvimento do modelo; o conceito de variância mensura a variabilidade da aplicação do modelo em amostras distintas, ou seja, da consistência do modelo.

Existe uma importante relação, um *trade-off*, entre o viés e a variância no desenvolvimento de modelos. Esse trade-off é mediado pela complexidade do modelo. O gráfico da figura 4 ilustra o comportamento dessa questão à luz de como modelos, dependendo de sua complexidade, variam em termos de erros e adequação às amostras de treino e alternativas. A título de compreensão do gráfico, os termos *bias*, *variance*, *e sample*, significam viés, variância e amostra, respectivamente.

Figura 4 - Erro dos modelos nas bases de treino e de teste em função da complexidade do modelo



Fonte: (FRIEDMAN; TIBSHIRANI; HASTIE 2008)

O objetivo do desenvolvimento de um modelo é que seu resultado permita uma boa diferenciação entre os indivíduos analisados, onde se capte bem os padrões que condicionam o problema e que seja bastante estável e consistente. Nos termos de viés e variância isso significa modelos com baixo viés e baixa variância. Sendo assim, o equilíbrio entre o viés e a variância configura um problema estrutural no desenvolvimento de algoritmos de predição e nas próximas seções serão apresentadas formas analisar e tratar do mesmo.

2.1.4 *Underfitting* e *overfitting*

O equilíbrio do viés e variância, apresentado na última seção, ocasiona em duas situações problemáticas: o *underfitting* e o *overfitting*. Em uma amostra de dados usual, temos uma certa variabilidade entre os inputs identificados nas diferenças de comportamento das variáveis independentes. Procura-se explicar essa variabilidade por meio das features e a interação do modelo, ilustrando os padrões de comportamento entre os indivíduos. Ocorre que se exige uma certa complexidade de tratamento dos padrões para que a variável dependente seja correspondente a essas da base de dados. Conforme se aumenta a complexidade do modelo, abre-se a possibilidade de explicar com grande acurácia os resultados. Ao mesmo tempo, quanto mais se aumenta a complexidade do algoritmo, mais se explicam fatores que são parte da variabilidade aleatória da base de dados onde se treinará o algoritmo, fazendo com que ele se torne viciado àquela amostra específica. Essas condições levam a uma situação de pouco viés e alta variância, situação essa que é denominada de *overfitting*.

De maneira oposta, um modelo muito simples, não capta muito bem os padrões a fim de predizer o resultado da variável dependente com grande acurácia, mas tendem a ter boa estabilidade e manter a consistência de acertos e erros entre diferentes bases de dados. Explicitamente, a situação de alto viés e baixa variabilidade é denominada de *underfitting*.

Ambas as situações apresentadas nessa seção são extremos que não são positivos no desenvolvimento de modelos. Nas seções a seguir serão apresentadas formas de evitá-las.

2.1.5 Florestas aleatórias

O modelo de florestas aleatórias é derivado de um modelo mais simples chamado de árvores de decisão, portanto o primeiro passo é compreender este último. Relembrando, o objetivo geral de qualquer modelo de classificação supervisionado, como o do presente trabalho, é de classificar as instâncias de maneira correta, a partir das *features* disponíveis. Para tanto, necessita-se extrair conclusões provenientes das *features* a fim de gerar uma conclusão sobre a sua relação com as classes existentes.

A maneira pela qual as árvores de decisão é a segmentação das instâncias por características das *features*. Para cada segmentação, existe uma parte das variáveis independentes que são explicadas e outra parte, não. Para exemplificar, digamos que a intenção seja especificar alunos que serão aprovados ou não em uma disciplina. Uma das variáveis dependentes é o sexo do discente. A primeira segmentação será, portanto, a divisão de sexo. No grupo feminino existem alunas aprovados e desaprovadas e alunos masculinos aprovados e desaprovados. Se o grupo de alunas tiver uma relação de alunos aprovados de 80%, enquanto o grupo de alunos tiver uma aprovação de 50%, poderíamos dizer de maneira geral que o sexo feminino é bom para diferenciar se uma pessoa foi aprovada e o sexo masculino não nos diz muito sobre a aprovação do discente. Em seguida, serão utilizadas outras maneiras de segmentar as instâncias, como idade ou nota no vestibular, para chegar a uma conclusão final sobre cada uma delas quanto à variável independente.

As divisões podem ser feitas com variáveis categóricas, ordinais ou contínuas e a determinação de quais serão as segmentações e qual será a sua ordem por meio de um critério de qualificação das segmentações que normalmente é o grau de impurezas, distância ou dependência. Em geral, a mais utilizada é a entropia, utilizada para calcular o ganho de informação fornecido por aquela divisão.

$$Entropia\ (S) = \sum_{i=1}^{c} -pi \log_2 pi \quad (1)$$

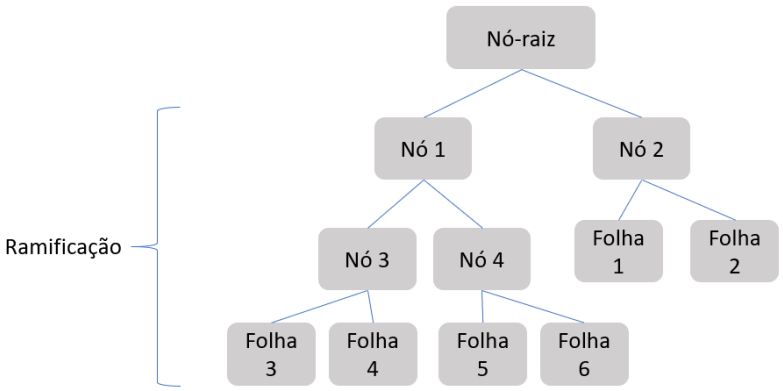
S = Conjunto de dados

pi = proporção de dados em S que pertencem à respectiva classe i

c = conjunto de classes existentes em S

Conforme o seu nome, a árvore de decisão é estruturada de maneira similar a uma árvore, onde cada segmentação é denominada de nó, que dividem as instâncias em 2 caminhos possíveis, denominados de ramos, que serão por sua vez segmentados por outros nós até chegarem a um grupo de instâncias com características similares e mesma classificação, denominados de folhas. O primeiro nó é denominado de nó raiz A figura 6 ilustra uma árvore de decisão genérica:

Figura 5 - Árvore de decisão genérica geral de árvore de decisão



Fonte: Elaborado pelo autor

Assim como no apresentado pela figura 6, o número de ramos não necessita ser não necessita ser par, uma mesma variável dependente pode ser segmentar um ramo mais de uma vez, o grau de hierarquia entre folhas não precisa ser idêntico e o número de folhas também é variável. Todas essas características podem ser modificadas para se obter um melhor modelo com melhor resultado.

A árvore de decisão é muito prática e de uso muito simples, sendo uma vantagem a diferenciação o fato dela capturar muito bem as diferenciações do *set* de treino, porém um dos seus maiores problemas é a sua alta variância. Essa instabilidade é em grande medida por conta da natureza hierárquica do processo, o que faz com que o efeito do erro no topo da árvore seja propagado conforme se desce até as folhas (FRIEDMAN; TIBSHIRANI; HASTIE 2001). Uma maneira de lidar com essa questão é o *bagging*. De acordo com Breiman (1996):

"Preditores bagging *é o método de gerar múltiplas versões do preditor e usar os mesmos para consolidar um preditor agregado. As múltiplas versões são formadas ao fazer réplicas* bootstrap *do set de treino e utilizá-las como novos sets de treino. Testes em dados*

reais e simulados usando árvores de classificação e regressão mostram que bagging *pode prover ganhos substanciais em acurácia."*

Réplicas *bootstrap* são partições de mesmo tamanho da base de treino que servirão para construir diferentes modelos, no caso analisado, árvores. No caso específico das árvores de decisão, enquanto uma árvore tem muito ruído em uma característica específica, outra terá ruído em outra segmentação, fazendo com que um preditor agregado amenize esses ruídos, aprimorando a acurácia e diminuindo a variância das classificações. Com a aplicação do método *bagging* em árvores de decisão, obtém-se o modelo de florestas aleatórias (*random forests*). Conforme apontado por Breiman (2001): "florestas aleatórias são a combinação de árvores preditores tal que cada árvore dependa de valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para cada árvore da floresta". O livro de Friedman, Tishbirani e Hastie (2001) descreve as etapas de construção de uma floresta aleatória:

1) Para b = 1 para B:
    a. Desenhe uma amostra *bootstrap* Z* de tamanho N a partir dos dados de treino.
    b. Cresça uma árvore de uma floresta aleatória com os dados *bootstrapped*, por recursivamente repetir os passos a seguir para cada nó da árvore até que o nó mínimo de tamanho Nmin é alcançado:
        i. Selecionar m variáveis randomicamente dentre as p variáveis.
        ii. Selecionar a melhor relação variável e ponto de divisão de variável dentre os valores de m.
        iii. Separar os nós em outros 2 nós-filhas.
2) Retornar o agregado de árvores
3) Classificar as instâncias teste por meio do voto de cada árvore.

Significa que no final cada uma das árvores dentre as criadas na floresta aleatória classificará de maneira separada cada instância e a soma desses votos constituirá a avaliação final da classificação da instância. Esse método tende a ser muito eficaz para uma série de problemas e é um dos modelos mais utilizados atualmente.

2.1.6 Regressão logística

O modelo de regressão logística é um modelo alternativo utilizado para classificação. Ele é em alguns aspectos análogo ao modelo de regressão linear, mas como o próprio nome sugere, ele não é linear. A regressão linear, resumidamente, utiliza as instâncias da base de treino para calcular a correlação entre as variáveis, gerando um coeficiente para cada variável independente em relação à variável independente. Esses coeficientes são ajustados para minimizar os erros entre as variáveis dependentes reais e as variáveis dependentes previstas pelo modelo, segundo a seguinte fórmula:

$$y_i = \alpha + \beta X_i + \varepsilon_i \quad (2)$$

Onde *yi* = variável dependente

α = constante que representa o ponto onde a reta intercepta com o eixo vertical

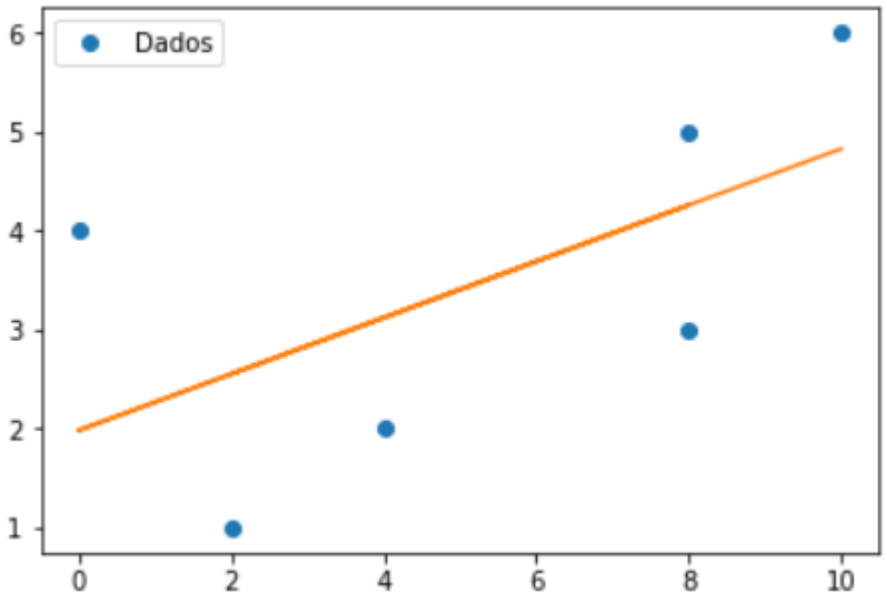
β = coeficiente da respectiva variável independente

Xi = variável independente

$\mathscr{E}$i = valor residual de y entre a base de dados e a regressão

A aplicação da regressão linear pode ser ilustrada pelo gráfico na figura 7, onde os pontos são as instâncias e a linha representa o modelo de regressão linear.

Figura 6 - Regressão linear genérica



Fonte: elaborado pelo autor

O modelo de regressão logística, por sua vez, é construído de maneira bastante similar, porém com algumas especificidades. A principal diferença é no tocante do seu objetivo, pois ela retorna um valor categórico. A maior similaridade é que ela segue os mesmos passos de calcular os coeficientes e posteriormente ajustar iterativamente para o melhor resultado conforme uma fórmula, porém tanto o cálculo dos coeficientes quanto a fórmula de otimização são diferentes. Os coeficientes de uma regressão logística são derivados do cálculo da probabilidade de uma determinação da variável dependente em relação à variável independente. O seu cálculo é o logaritmo da probabilidade de classificação positiva da instância, dada a característica independente. A otimização da curva, por sua vez, é a maximização das probabilidades de classificação da instância, dadas as instâncias e seus coeficiente, conforme a fórmula:

$$p_i = \frac{1}{1 + e^{-(\beta_0 - \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}. \quad (3)$$

Onde *pi* é a probabilidade de classificação correta, variando entre 0 e 1
βi são as variáveis independentes
xi são os coeficientes das respectivas variáveis independentes

Esse é um modelo de uso bastante simples e amplamente utilizado pela sua praticidade e obtenção de resultados razoavelmente bons por um baixo custo. Em muitas aplicações de classificação por aprendizado de máquina ele consiste em um modelo de baixa complexidade e possível aplicação geral consistente, além de ser muito útil como parâmetro para o teste de outros modelos mais sofisticados.

2.1.7 Tunagem de hiperparâmetros (*Hyperparameter Tuning*)

Cada modelo de aprendizado de máquina opera para obter seus resultados de formas muito distintas. Todos possuem uma forma de leitura dos dados para extrair informações e padrões de comportamento das variáveis independentes a fim de apontar um valor ou classe para a variável dependente. Algo comum a vários métodos é que eles são todos parametrizados por um grupo de hiperparâmetros λ, que devem ser configurados apropriadamente pelo usuário para maximizar a utilidade da abordagem de aprendizagem (CLAESEN, DE MOOR, 2015). Atualmente existem várias formas, várias fórmulas para

maximizar a utilidade e ela depende de cada objetivo do modelo, assim como ela pode ser aplicada a cada um dos hiperparâmetros. Existem algoritmos que buscam o melhor valor possível de hiperparâmetro, mas por conta da praticidade, optou-se no presente trabalho por aplicar testes com valores arbitrários para cada hiperparâmetro realizados iterativamente e comparação dos resultados nas métricas de avaliação do modelo.

2.1.8 Avaliação de modelos de classificação

2.1.8.1 Validação Cruzada

Na seção 2.1.4 foi apresentado o trade-off de viés e variância. Ele se encontra dentro da discussão de como tornar um modelo mais generalizável, mais bem sucedido situações similares à apresentada nos dados, mas distintas, permitindo alguma análise sobre outro grupo dentro dessa população. Uma maneira simples e eficaz é a validação cruzada. Esse método consiste na segmentação da base de dados em 2 grupos: a base de treino e a base de teste. O objetivo com isso é a realização de um exercício de teste do desempenho do modelo em um grupo de dados diferente, ou seja, um experimento no qual o modelo é gerado a partir de um grupo e aplicado em outro. Ao aplicar o modelo na base de treino, obtém-se predições sobre o resultado de cada uma instância. Dessa maneira, ambos os valores previstos e o resultado real dessas instâncias estão disponíveis, o que significa que se pode calcular o erro das previsões. O cálculo desse erro segue conforme a fórmula genérica:

$$Err = \sum L(Y, f(X)) \quad (4)$$

Y = y real da base de dados

f(X) = valor predito pelo modelo com os inputs X

L(Y,f(X)) = função que determina o erro de cada valor predito

O desempenho do modelo em outro *set* de dados, nomeadamente o *set* de teste, corresponde a um uma forma de analisar o quão generalizável o modelo é, o quão confiável ele será em uma situação real de aplicação. O erro, conforme apontado na fórmula, é um indicador importante e, conforme será visto nas seções seguintes, existem muitos indicadores que podem ser derivados desse erro e utilizados para refinar o modelo. Um ponto crítico para se ressaltar é sobre o trade-off viés-variância, a relação entre o tamanho da base de treino e a base de teste, assim como a escolha de como será a divisão dos dois é essencial para poder

criar um modelo que seja robusto no sentido de conseguir explicar os padrões relevantes para uma boa diferenciação entre as instâncias, assim como uma boa estabilidade para previsão em outras aplicações.

Existem formas de trabalhar com a segmentação das bases para se aprimorar a qualidade do modelo, como a *K-Fold Cross-Validation*. Nesse tipo de aplicação, refina-se o problema segmentando a base várias partes, as quais são utilizadas para treinar o modelo e testar cada um dos modelos treinados para prever em uma parte específica, sendo que o erro é avaliado para todos os modelos treinados e realiza-se uma otimização do modelo para além da simples separação entre uma base de treino e de validação (FRIEDMAN, 2001). Esse tipo de técnica aprimora aumenta a complexidade do modelo, captando de maneira mais assertiva a variabilidade da amostra, mas pode dificultar a generalização para aplicar em demais amostras mais distintas.

2.1.8.2 Matriz de confusão e métricas associadas

A matriz de confusão é uma ferramenta amplamente utilizada para a avaliação de modelos por conta da sua simplicidade e praticidade. Consiste em uma representação comparativa dos resultados reais e previstos na base de teste. Em termos práticos, uma matriz de confusão de tamanho n x n associada a um classificador mostra os as classificações previstas e reais, onde n é o número de classes diferentes (S VISA, B RAMSAY, AL RALESCU, E VAN DER KNAAP - MAICS, 2011). Considerando que o problema do presente trabalho busca predizer uma resposta binária, aprovado ou não aprovado, o n utilizado é igual a 2, apontando uma matriz conforme a Tabela 1:

Tabela 1 - Matriz de confusão genérica

	Negativo previsto	*Positivo Previsto*
Negativo Real	VN	FP
Positivo real	FN	VP

Fonte: elaborado pelo autor

A matriz é formatada por linhas e colunas, onde as linhas indicam os resultados reais e as colunas os resultados previstos. As células onde estão os números indicam o somatório da intersecção entre instâncias previstas e reais. A seguir estão apresentados os significados das de intersecção:

1) Verdadeiros negativos (VN) – somatório de instâncias corretamente previstas como negativas.

2) Verdadeiros positivos (VP) – somatório de instâncias corretamente previstas como positivas.

3) Falsos negativo (FN) – somatório de instâncias equivocadamente previstas como negativas.

4) Falsos positivos (FP) – somatório de instâncias equivocadamente previstas como positivas.

A matriz permite uma visualização interessante por conta de algumas peculiaridades de fácil extração. A soma das células de cada coluna representa a quantidade de negativos e positivos reais. A soma das células de cada linha representa a quantidade de negativos e positivos previstos. A soma de cada diagonal representa os valores corretamente previstos e equivocadamente previstos.

Existem métricas menos imediatas derivadas da matriz de confusão bastante úteis no sentido de apontar para uma qualificação dos resultados provenientes do modelo. A tabela 2 apresenta algumas delas:

Tabela 2 - Métricas derivadas da matriz de confusão

Métrica	Descrição	Cálculo
Acurácia	Representa qual foi a taxa de acerto geral do modelo. Em geral serve para avaliar o modelo de maneira genérica, sem diferenciação perante o tipo de classificação	$\frac{VP + VN}{(\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN})}$
Taxa de erro	Representa a taxa de erro geral do modelo. É análoga, porém oposta à precisão.	$\frac{FP + FN}{(\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN})}$
Precisão	Representa a taxa de acerto dos preditos positivos, ou seja, qual a proporção de acertos das instâncias que foram preditas como positivas	$\frac{VP}{(\mathrm{VP} + \mathrm{FP})}$
Sensibilidade	Representa a taxa de acerto dos positivos, ou seja, qual a proporção de acertos das instâncias que eram de fato negativas	$\frac{VP}{(\mathrm{VP} + \mathrm{FN})}$
Especificidade	Representa a taxa de acerto dos negativos, ou seja, qual a proporção de acertos das instâncias que eram de fato negativas	$\frac{VN}{(\mathrm{VN} + \mathrm{FP})}$

Fonte: elaborado pelo autor

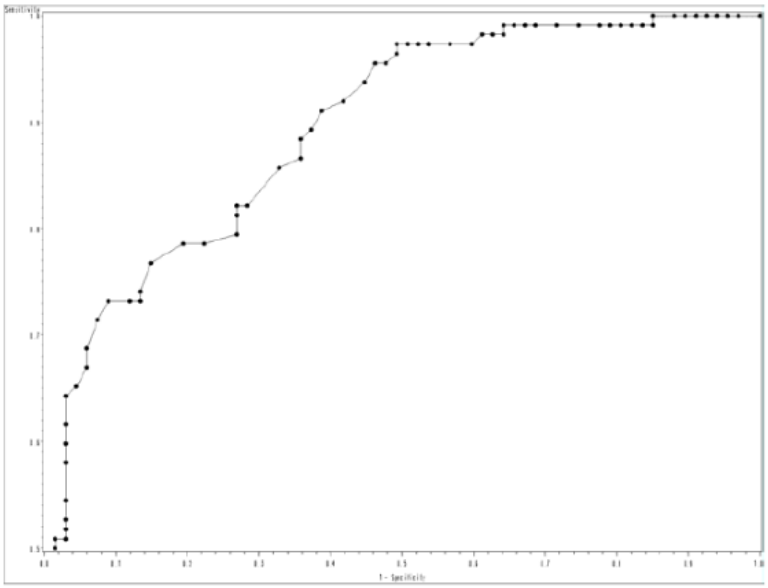
Cada uma das métricas aponta uma informação bastante específica e nenhuma é definitiva para a conclusão da utilidade do modelo. A utilidade do modelo depende das características dos dados e da utilização, do objetivo da modelagem. Dessa maneira, existem situações em que se optará por um modelo com menor precisão, mas uma melhor sensibilidade, pois o custo de um falso positivo pode vir a ser maior do que o custo de um falso negativo.

2.1.8.3 Curva AUC-ROC

A utilização das métricas como precisão, sensibilidade de maneira isolada pode, conforme brevemente apontado na última seção, levar a uma dificuldade de avaliação de modelos, seja para uma comparação entre modelos ou a calibragem de um modelo específico. Para auxiliar nessa tarefa, uma ferramenta eficaz é a construção de uma curva *Receiver Operating Curve* (ROC). De acordo com Fawcett (2006): "em adição a ser um método de grafia de performance geralmente útil ela tem propriedades que fazem dela especialmente úteis para domínios que possuem classes assimetricamente distribuídas e erros de classificação com custos desiguais".

Uma curva ROC é a junção dos pontos obtidos variando o limiar de decisão do modelo, responsável por calibrar a classificação do modelo. Variando o limiar de decisão, pode-se observar uma mudança de comportamento sobre a taxa de acerto dos positivos e a taxa de acerto dos negativos, respectivamente a sensibilidade e especificidade. É uma calibragem da classificação do modelo pela visualização de um gráfico no qual o eixo y representa a sensibilidade e o eixo x representa a especificidade. Apesar disso, como as duas métricas são inversamente proporcionais, um jeito mais interessante de apresentar a curva é modificando o eixo x para 1 – especificidade (GONEN, 2007), conforme ilustrado pela figura 8:

Figura 7 – Exemplo de curva ROC



Fonte: (GONEN, 2007)

Ao se construir um modelo a intenção é aumentar a sensitividade, taxa de acertos positivos, e a especificidade, taxa de acertos negativos, de maneira que uma no gráfico acima apresentado, quanto mais à esquerda do gráfico, melhor será o resultado obtido. Em um problema de classificação bidimensional, como o do presente projeto, o problema classificatório binário sempre terá a solução aleatória, que consolida um patamar de 50% de acertos. Como exemplos pode se pegar os extremos do problema, como classificar todas as instâncias como positivas ou negativas, em cada caso obtendo sensibilidade 1 e especificidade 0, no caso 1, ou sensibilidade 0 e especificidade 1, no caso 2. Dessa maneira, os parâmetros de escolha dentro da curva ROC são claros, sendo no máximo igual a 1 e no mínimo igual a 0,5.

Conforme apontado pela citação de Fawcett, um elemento relevante para a avaliação é a diferença de custo e desbalanço de certas classes. Para exemplificar, digamos que exista um modelo para a previsão de quebra de uma barragem, onde as barragens classificadas com risco de rompimento seriam revisadas. O risco de um falso positivo é a revisão de uma barragem enquanto o risco de um falso negativo é o rompimento da barragem. Nessa situação, o custo entre as classes é muito discrepante o que faz com que o modelo deva ser calibrado para uma alta especificidade, possivelmente em detrimento da sensitividade. O mesmo pode ocorrer no caso de desbalanço muito grande de classes, fazendo com que o volume aumente consideravelmente o custo do erro em alguma das classes.

A partir da curva ROC, existem métricas que buscam simplificar e resumir a visualização e escolha entre limiar de decisão ou entre modelos concorrentes. A mais amplamente utilizada é a *area under the curve* (AUC). Não levando em conta questões como diferença de custo ou outras peculiaridades do problema em questão, ela aponta com clareza qual ponto, qual calibragem possui o melhor resultado em questão de otimização de sensibilidade e especificidade, sendo o seu ideal o valor 1, sensibilidade e especificidade iguais a 1, e 0,5, área abaixo da linha de pontos de sensibilidade e especificidade provenientes da classificação aleatória.

2.1.8.4 Precision at K

Certos tipos de aplicação de modelos de aprendizagem de máquina exigem uma grande precisão para elementos mais relevantes dentro da amostra prevista. Para casos de classificação com classes raras, como a identificação de *spams*, devem detectar anomalias para serem efetivos. Para tanto, uma série de métricas foram desenvolvidas como a precisão em K (*Precision at K*), precisão média e *Non discounted cumulative gain* (NDCG), cada qual especialmente aplicável para cada domínio de aplicações (NARASIMHAN et. al, 2015). Esse tipo de métrica é muito interessante para a identificação de elementos relevantes, mas não representa uma boa maneira de classificar os elementos em questão de relevância entre si. Apesar disso, na aplicação do presente projeto, a precisão de identificação de elementos com maior probabilidade de reprovação se torna importante para medidas que tenham restrição de recursos ou direcionadas a certos grupos de alunos para com suas necessidades específicas. A maneira como a métrica funciona é a aplicação da métrica de precisão, ou taxa de acertos, dentro do grupo de K elementos mais relevantes da base de teste.

3 METODOLOGIA

A metodologia deste trabalho é composta por 3 partes. A primeira se refere à descrição da base de dados utilizada e o cuidado exigido durante a pesquisa e documentação por conta do sigilo das informações. A segunda se refere ao trabalho de adequação da leitura e interpretação dos dados que levou à construção das *features* utilizadas para a construção do modelo. A terceira corresponde à efetiva construção, avaliação e aprimoramento do modelo.

3.1 DESCRIÇÃO DOS DADOS

3.1.1 Sigilo dos dados

Os dados utilizados para o presente trabalho são provenientes da base de dados de cadastro e histórico de alunos do CTC desde o ano de 2000 até o ano de 2019. Eles compreendem, de maneira genérica, informações socioeconômicas dos alunos matriculados, assim como de seu desempenho nas matérias cursadas. Ambas as informações individuais são protegidas por lei de maneira que se exige uma autorização direta dos alunos para que sejam divulgadas. Como o presente trabalho não possui a intenção de divulgar tais dados, eles não serão apresentados no presente documento a fim de resguardar os interessados. Assim, todos os levantamentos quantitativos constarão de maneira agregada, impossibilitando a associação de qualquer aluno a o que for apresentado. A elucidação metodológica e de apresentação dos resultados do modelo, todavia não depende do fornecimento de tais informações particulares.

3.1.2 Variáveis

Dada a condição de instituição pública e o porte da UFSC, as exigências legais, assim como as demandas de dados feita por diferentes atores externos e internos à universidade, a variedade de bases de dados e diferentes informações armazenadas dos alunos é enorme. Sendo assim, inicialmente foi feito um levantamento de quais variáveis seriam de importância para a construção do modelo proposto. Esse levantamento inicial foi feito por meio de uma análise qualitativa de outros trabalhos da literatura (SHINYASHIKI, BACCARO, 2014) (BUENAÑO-FÉRNANDEZ,GIL,LUJÁN-MORA, 2019) (ABUT, YUKSEL, AKAY, DANESHVAR, 2018) (CASEY, AZCONA, 2017) e professores

especialistas sobre o tema nos âmbitos da instituição e nacional. Essa base inicial passou por alterações, ao passo que hipóteses intermediárias de desenvolvimento do modelo necessitavam de dados que não constavam no apanhado inicial de variáveis. Assim, sucessivamente, conforme o ganho de maturidade sobre o tema e do universo disponível para análise, a base de dados utilizada cresceu.

A primeira etapa é, portanto, caracterizar os dados extraídos das bases de dados e entender suas características e limitações. É importante ressaltar que algumas das variáveis tem seu preenchimento provavelmente incompleto e outras começaram a ser contabilizadas no meio do período em análise. A tabela a seguir apresenta todas as variáveis extraídas das bases de dados da universidade, indicando também a quantidade de dados contidas em cada uma:

Tabela 3 - Tabela de variáveis extraídas das bases de dados

Nº	Variável	Descrição	Dados não nulos
1	Periodo Vest	Período no qual o discente ingressou na universidade	1.229.891
2	Ead	Indica se a matéria é ou não de modalidade à distância	567.008
3	Matricula	Código de matrícula do discente	1.229.891
4	CPF	Código do comprovante de situação cadastral do discente	1.219.324
5	Sexo	Gênero pelo qual o aluno se identificou no registro do vestibular	1.229.891
6	Nacionalidade	Nacionalidade do pai do discente	1.229.891
7	Naturalidade Pai	Nacionalidade do pai do discente	1.204.295
8	Estado	Estado de origem do discente	1.204.295
10	Nome Curso	Nome do curso no qual o discente está matriculado	1.229.891
11	Turno	Turno do dia no qual o aluno se matriculou	1.229.891
12	Pontos Vest	Pontuação total do discente na prova de vestibular	1.171.151
13	Ano Vest	Ano no qual o discente se prestou o vestibular	1146118
14	Ano Form Colegio	Ano no qual o discente se formou no ensino médio	1.222.826
15	Colegio EM	Colégio no qual o discente cursou o ensino médio	1.226.459
16	Cidade EM	Cidade no qual o discente cursou o ensino médio	1.166.201
17	Estado EM	Estado no qual o discente cursou o ensino médio	1.166.201

18	Data Nasc	Data de nascimento do discente	1.229.891
19	Idade Vest	Idade com a qual o discente ingressou na universidade	1.146.118
20	Etnia	Categoria étnica pela qual o aluno se identificou no cadastro do vestibular	1.080.163
21	Forma de Ingresso	Instrumento avaliativo pelo qual o discente concorreu para ingressar no curso	1.229.891
22	Categoria de Ingresso	Categoria pela qual o aluno se matriculou no curso	567.008
23	SemestreMax	Número máximo de semestres que o discente pode demorar para completar seu currículo	1.229.891
24	Semestre Histórico	Semestre no qual o discente ou o docente participaram daquela turma	1.229.891
25	Disciplina	Nome da disciplina cursada pelo discente ou ministrada pelo docente no respectivo semestre	1.229.891
26	Nota	Nota obtida pelo aluno na respectiva matéria	1.229.891
27	Horas Aula	Carga horário da matéria em questão	1.229.891
28	Frequência	Indica se o aluno teve frequência suficiente para qualificar para potencial aprovação	1.204.656
29	TipoCurriculo	Categoria pela qual a matéria é classificada no currículo do discente	1.194.193
30	Revalida	Corresponde a se o aluno cursou ou validou a respectiva matéria no respectivo semestre	59.060
31	Turma	Código da turma no qual os respectivos discentes e docentes referentes à uma determinada matéria em um determinado semestre	1.136.035
32	Matricula Professor	Número de cadastro de matrícula do docente	1.099.274
33	SIAPE Professor	Número de cadastro nacional de servidor público do docente	1.070.230
34	Nome Professor	Nome de registro em cartório do docente	1.099.262
36	Situacão Aluno	Categoria referente à situação de matrícula do aluno na data de extração do banco de dados	1.229.891

Fonte: elaborado pelo autor

É necessário explicitar informações que caracterizam a base extraída.

a) A base de dados possui o histórico dos alunos no compreendido entre o primeiro semestre do ano de 1990 ao segundo semestre do ano de 2018.

b) Os dados apresentados na tabela 4 possuem 2 tipos de dados: dados do perfil de alunos e professores; dados do histórico de disciplinas cursadas.

Quantidade de alunos e professores na base:

Tabela 4 - Quantidade de matrículas e cadastros sociais de alunos e professores

	ALUNOS	PROFESSORES
MATRÍCULAS	26.882	3.162
CADASTRO	23.227	3.112

Fonte: elaborado pelo autor

Existem mais matrículas do que CPFs de alunos. Isso ocorre, pois existem alunos que fazem transferência interna ou passam no vestibular novamente, gerando uma outra matrícula para a mesma pessoa. Da mesma maneira, existem professores que mudam de departamento ou local e criam uma nova matrícula, mas possuem o mesmo SIAPE, cadastro referente ao governo.

3.1.3 Leitura e Interpretação dos dados

A segunda etapa, conforme apontou-se no capítulo 2, é a análise descritiva dos "*inputs"* com o fim de entender com maior detalhe a situação em análise e como isso afeta o objetivo do modelo. O objetivo do presente trabalho procura explicar e sugerir mudanças para um melhor desempenho dos alunos durante o curso da sua graduação, portanto devemos definir como será avaliado o desempenho e seus critérios de sucesso. Sendo assim, o presente trabalho define o desempenho dos alunos por nota, sendo notas mais altas um melhor desempenho. A UFSC trabalha com a média 6 para determinar aprovação, portanto esse é um crivo importante para determinar se o aluno foi bem sucedido na disciplina. Para melhor compreender quais as relações de todas as variáveis frente ao sucesso de cada matéria, foram criados alguns parâmetros derivados da variável "Nota":

1- Índice de Aproveitamento (IA) – Média das notas em cada disciplina de um semestre, ponderada pela carga do curso, ou número de horas equivalentes de crédito.

$$IA = \frac{(\text{Nota Disciplina } 1 * \text{Carga horária disciplina } 1) + (\text{Nota Disciplina } 2 * \text{Carga horária disciplina } 2)}{(\text{Carga horária disciplina } 1 + \text{Carga horária disciplina } 2)} \quad (6)$$

2- Índice de Aproveitamento Acumulado (IAA) – Média dos IAs de todos semestres, ponderada pela soma das cargas das disciplinas do respectivo semestre.

$$IA = \frac{(\text{IA semestre } 1 * \text{Soma carga horária semestre } 1) + (\text{IA semestre } 1 * \text{Soma carga horária semestre } 2)}{(\text{Soma carga horária semestre } 1 + \text{Soma carga horária semestre } 2)} \quad (7)$$

3- Aprovado – Critério de sucesso na matéria. Se nota na disciplina > = 6, o aluno foi aprovado e se a nota na disciplina < 6, o aluno não foi aprovado.

Se nota > = 6 → Aprovado = 1
Se nota < 6 → Aprovado = 0

4- Taxa de aprovação: Proporção de aprovados em relação ao total de disciplinas cursadas. A taxa pode ser aplicada a alunos, professores ou disciplinas, seja em ordem semestral ou histórico.

Estes parâmetros são formas de quantificar objetivamente o desempenho por meio dos dados de histórico. Serão feitas análises dos perfis de alunos para entender a diferença de comportamento do desempenho para cada característica. A análise seguirá a mesma estrutura para cada tópico:

a. Tabela de informações gerais dos respectivos segmentos
    a. Representatividade de cada grupo na base
    b. Média de aprovação das matrículas do respectivo grupo
    c. Média de IAA das matrículas do respectivo grupo
    d. Percentual de matrículas em cada situação de formatura, ou seja, formado, desistente ou regular, sendo regular o aluno com matrícula ativa.

b. Gráficos da distribuição dos alunos de cada segmento quanto a seu IAA
    a. Como professores e disciplinas não possuem IAA, não incluirão essa etapa.

c. Gráficos da distribuição dos alunos de cada segmento quanto a sua média de aprovação das matrículas
d. Por fim uma leitura sobre o desempenho do tópico analisado

Cabe observar que, conforme apresentado pela tabela 3 (tabela de informações) existem dados faltantes em tópicos de análise e isso deve ser levado em consideração. Uma segunda questão é que 5% dos dados de disciplinas são de validações de disciplinas feitas anteriormente ao registro da matrícula do aluno. Esse efeito não será analisado pelo presente trabalho. Dentro desse efeito, um terceiro elemento é a rematrícula de alunos no mesmo curso. Esse efeito gera uma distorção para certos cursos, que possuem quase 10% de rematrícula da mesma pessoa, verificada pelo CPF, mas esse efeito novamente não será incluído na análise dos segmentos, pois em termos da base inteira, o efeito não é significativo. Por fim, nota-se uma redução no número de matrículas extraído do banco, pois foi feita uma limpeza da base, excluindo matérias especiais como intercâmbio e matrículas com menos de 4 disciplinas, que constitui um semestre normal.

3.1.3.1 Geral

3.1.3.1.1 Dados Gerais

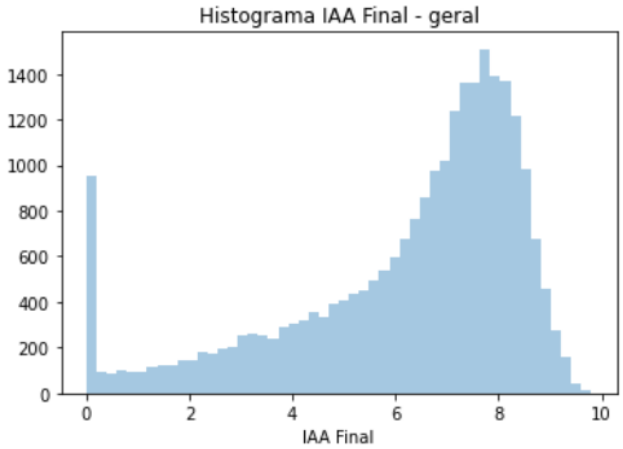
Tabela 5 - Dados gerais dos alunos da base analisada

MATRÍCULAS	APROVAÇÃO	MÉDIA IAA	FORMADO	REGULAR	DESISTENTE
25.061	76%	6,21	48%	22%	30%

Fonte: elaborado pelo autor

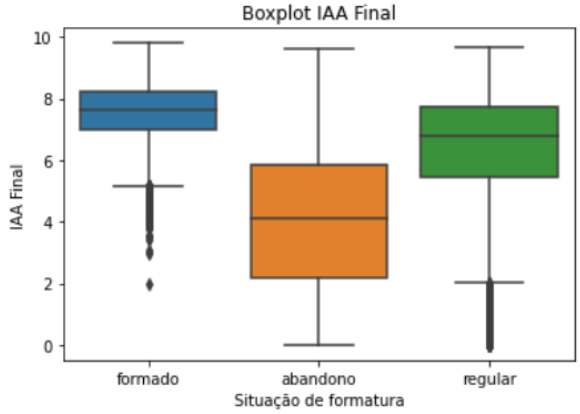
3.1.3.1.2 Análise IAA

Figura 8 - Histograma IAA – geral



Fonte: elaborado pelo autor

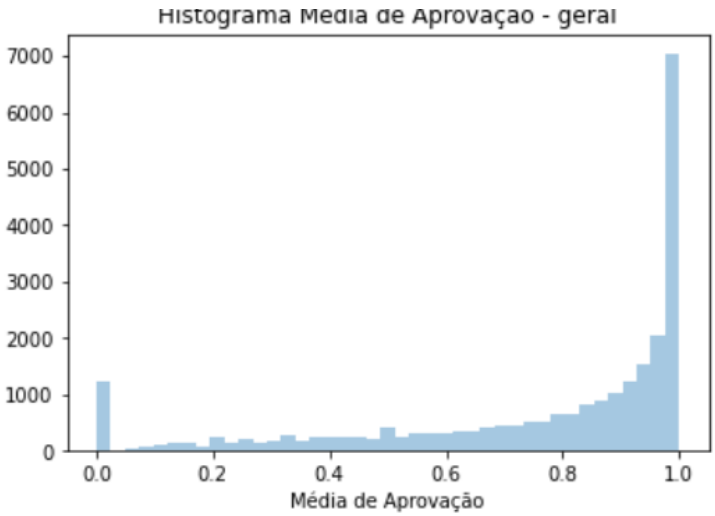
Figura 9 - Boxplot IAA Final - segmentado por situação de formatura



Fonte: elaborado pelo autor

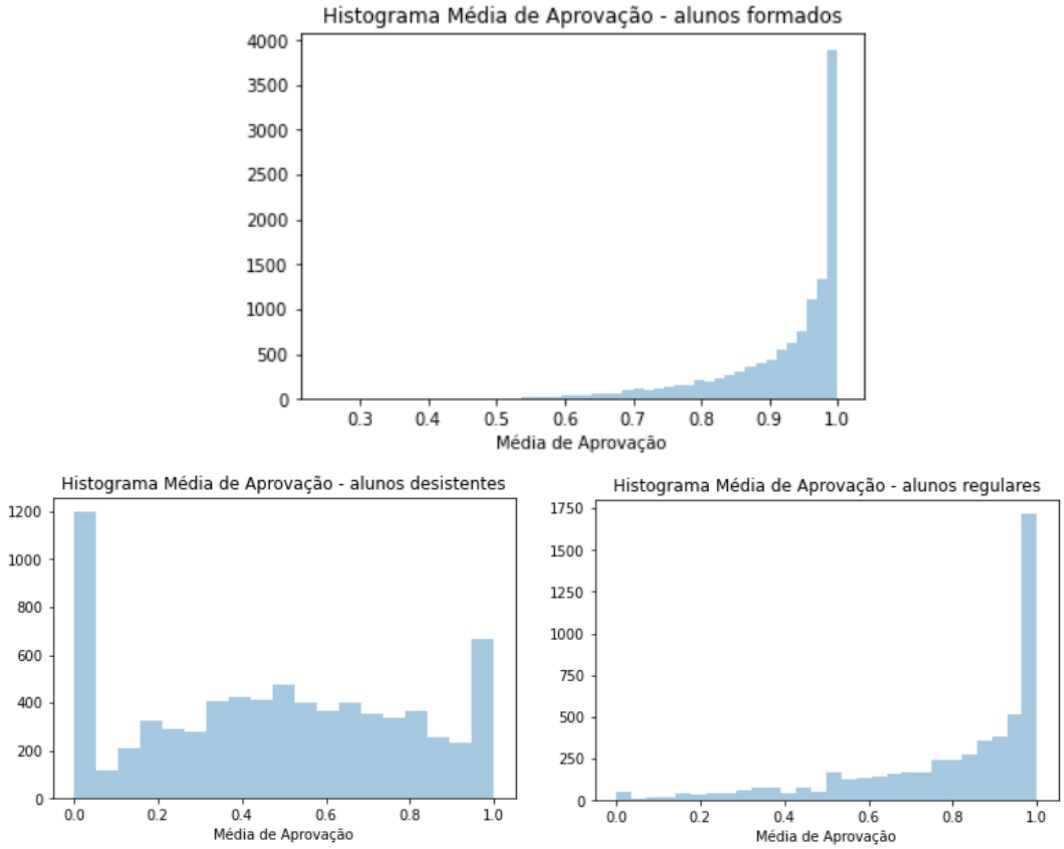
3.1.3.1.3 Análise aprovação

Figura 10 - Histograma de aprovação - geral



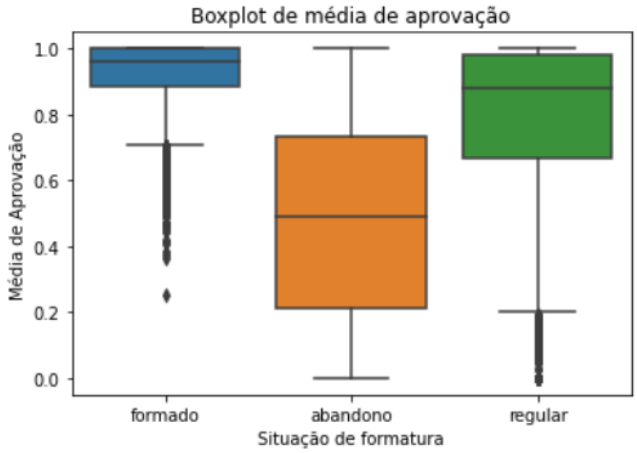
Fonte: elaborado pelo autor

Figura 11 - Histogramas de alunos por média de aprovação por situação de formatura



Fonte: elaborado pelo autor

Figura 12 - Boxplot de média de aprovação por situação de formatura



Fonte: elaborado pelo autor

Dos números apresentados na tabela, um pouco mais do que 5.000 matrículas ainda estão ativas. Dentre estas, algumas se formarão e outras não. Assumindo que a proporção se mantenha, significa que a taxa de formatura dos alunos é de aproximadamente 61,5% e de desistência 38,5%. Considerando que a aprovação geral de 76%, existe uma discrepância para

a taxa de formatura de 61,5%, o que pode indicar um desequilíbrio entre discentes com muitas aprovações e a maioria dos alunos tenha uma média de aprovação bastante menor. Esse raciocínio é confirmado pelo ilustrado nos gráficos de aprovações, principalmente pelo *boxplot* que indica que alunos que 75% dos alunos que se formam possuem taxa de aprovação de 90% para cima, quase 100% se concentra acima de 70% de aprovações, aparecendo apenas alguns outliers abaixo dessa faixa. Da mesma forma, vê-se que a amplitude de alunos que abandona o curso é expressivamente maior, apontando desde alunos com média de aprovação de 100% a alunos com aprovação próxima de 0%. O histograma mostra de maneira simples a maior dispersão de alunos desistentes do que o de alunos regulares ou formados. Ao que aparenta, aproximadamente 75% dos alunos regulares está dentro da faixa de 70% de aprovação, enquanto 100% dos alunos formados e apenas 25% dos alunos desistentes estão na mesma faixa. Assumindo que 25% serão alunos que serão desistentes, os demais 50% dos alunos se formarão, apontando para a mesma hipótese levantada no início do parágrafo.

Ao observar o comportamento dos gráficos de IAA, verifica-se em geral um comportamento similar ao de aprovação, mas ao mesmo tempo uma maior amplitude de dados para todas as categorias de alunos quanto a sua formatura. Apesar de que nos alunos formados, a maioria está no grupo de alunos com aprovação de mais de 70%, vemos que 50% deste grupo tem um IAA entre 7,6 e 8,2. Se aplicarmos um valor composto do IAA médio e tamanho de cada grupo, chegamos ao valor aproximado de 5,4 de IAA. Uma possibilidade para explicar o valor de 6,21 do IAA médio é que alunos formados tendem a fazer mais disciplinas do que os evadidos, sendo assim esse grupo de alunos tem um peso maior no IAA médio.

3.1.3.2 Gênero

3.1.3.2.1 Dados gerais

Distribuição de matrículas por gênero, conforme a tabela 6:

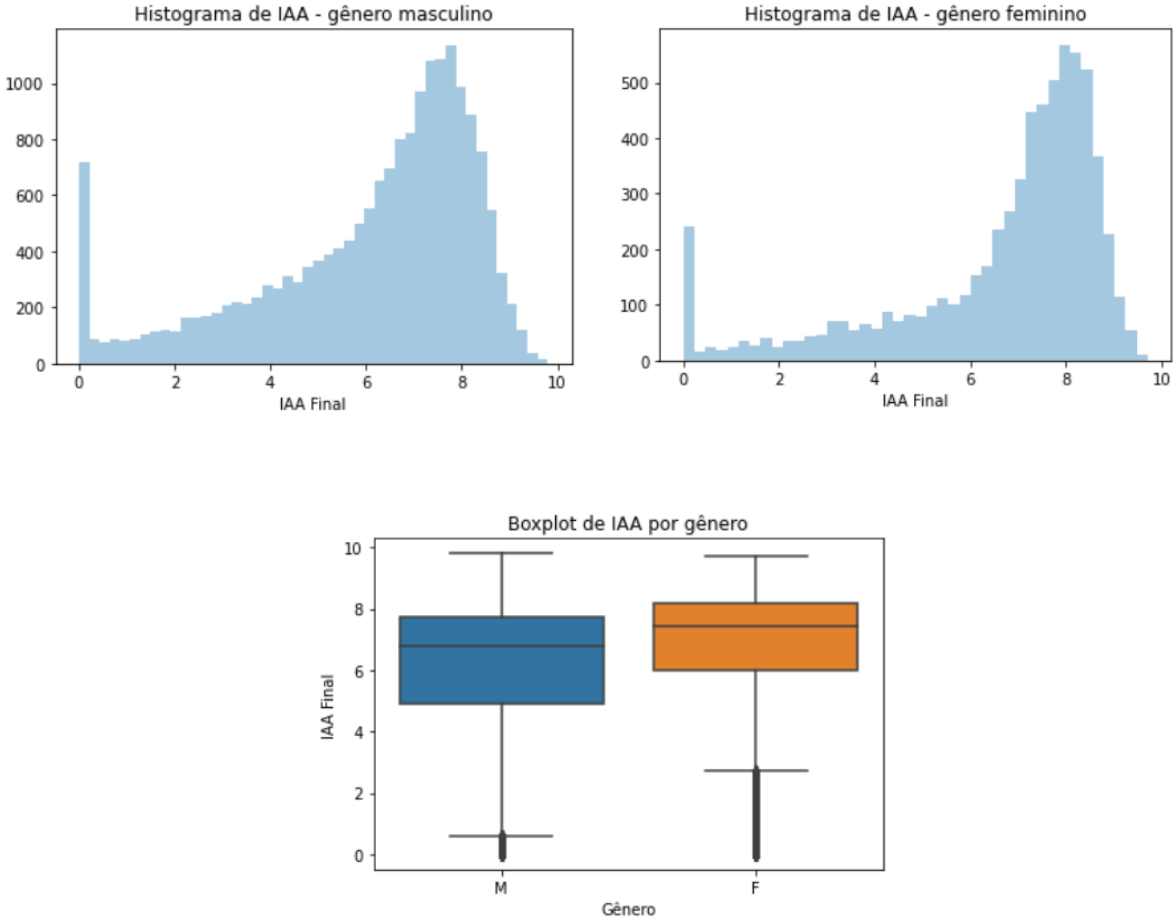
Tabela 6 - Dados gerais dos alunos por gênero

GÊNERO	MATRÍCULAS	TAXA DE APROVAÇÃO	MÉDIA IAA	FORMADO	REGULAR	DESISTENTES
MASCULINO	73%	74%	6,05	47%	21%	32%
FEMININO	27%	81%	6,65	51%	23%	26%

Fonte: Elaborado pelo autor

3.1.3.2.2 Análise IAA

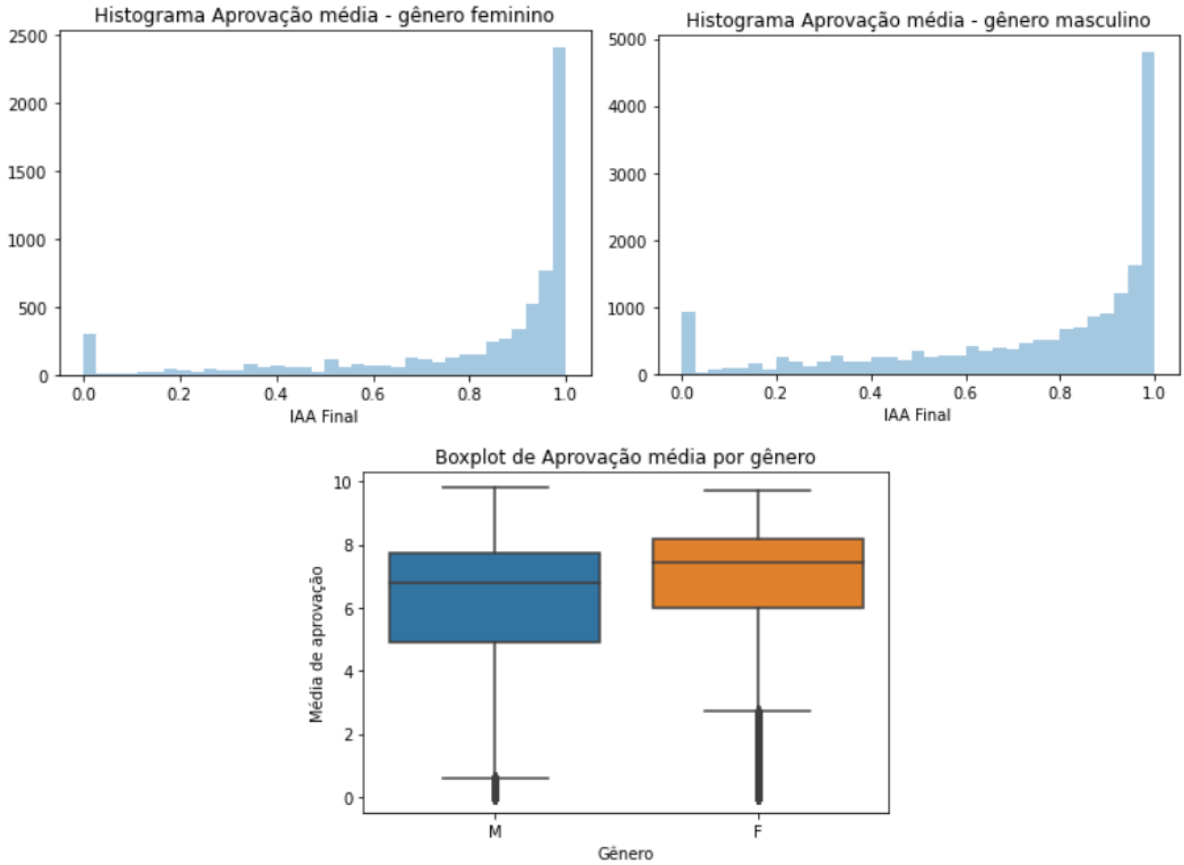
Figura 13 – Histogramas e boxplot de alunos por média de aprovação por situação de formatura



Fonte: elaborado pelo autor

3.1.3.2.3 Análise aprovação

Figura 14 - Histogramas e boxplot de aprovação média por gênero



Fonte: elaborado pelo autor

Ressalta-se imediatamente uma proporção muito maior de alunos do sexo masculino do que feminino. Como um todo, vê-se o mesmo padrão do geral tanto para o IAA quanto para as aprovações. A diferença é perceptível de um ligeiro deslocamento do desempenho à direita, ou seja, superior, para o grupo do sexo feminino, que também se verifica em todos os índices apresentados na tabela 5.

3.1.3.3 Categoria de Ingresso

3.1.3.3.1 Dados gerais

Como as cotas só foram implementadas no vestibular 2007-2008, o grupo de análise para essa comparação é apenas a partir do ano de 2008. Cabe ressaltar que as cotas as vagas disponíveis para cotas foram gradualmente aumentadas, fazendo com que boa parte do grupo de alunos de cotas seja mais recente.

É importante ressaltar a forma de funcionamento do sistema de cotas optada pela UFSC. Se um vestibulando é elegível para o sistema de cotas, mas obtém uma nota acima da nota de corte, ele é classificado na base de dados como classificação geral, tornando infactível o rastreamento de alunos nessa situação. Consequentemente, todos alunos classificados pelo sistema e cotas foram avaliados com notas abaixo do corte.

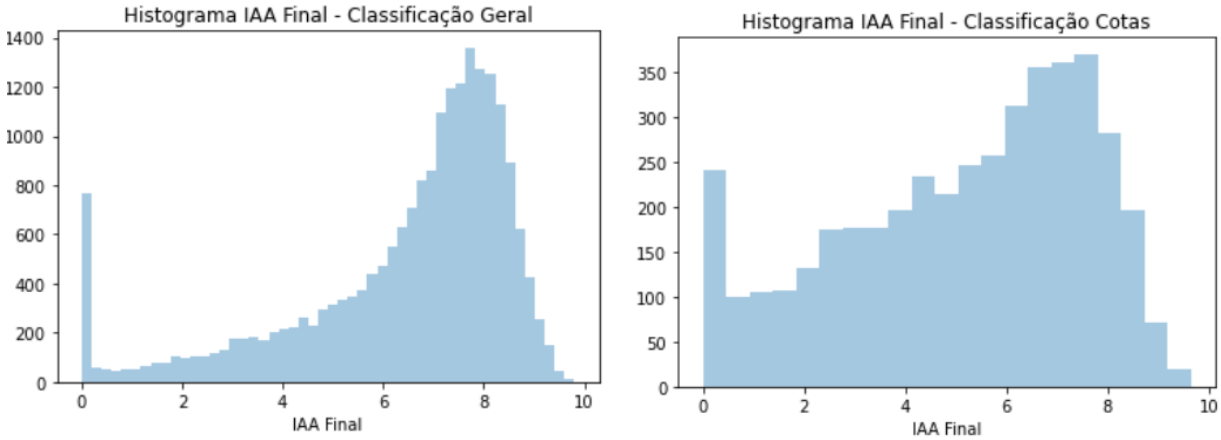
Tabela 7 - Dados gerais dos alunos por classificação

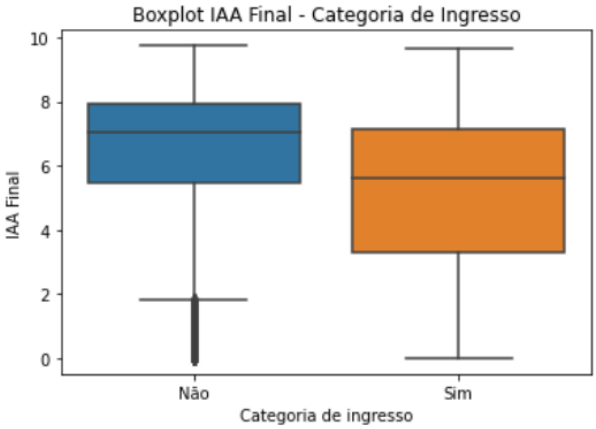
CATEGORIA DE INGRESSO	MATRÍCULAS	TAXA DE APROVAÇÃO	MÉDIA IAA	FORMADO	REGULAR	DESISTENTES
GERAL	66%	79%	6,40	33%	38%	28%
COTA	34%	62%	5,15	16%	50%	34%

Fonte: Elaborado pelo autor

3.1.3.3.2 Análise IAA

Figura 15 - Histogramas e boxplot de alunos por IAA por categoria de ingresso

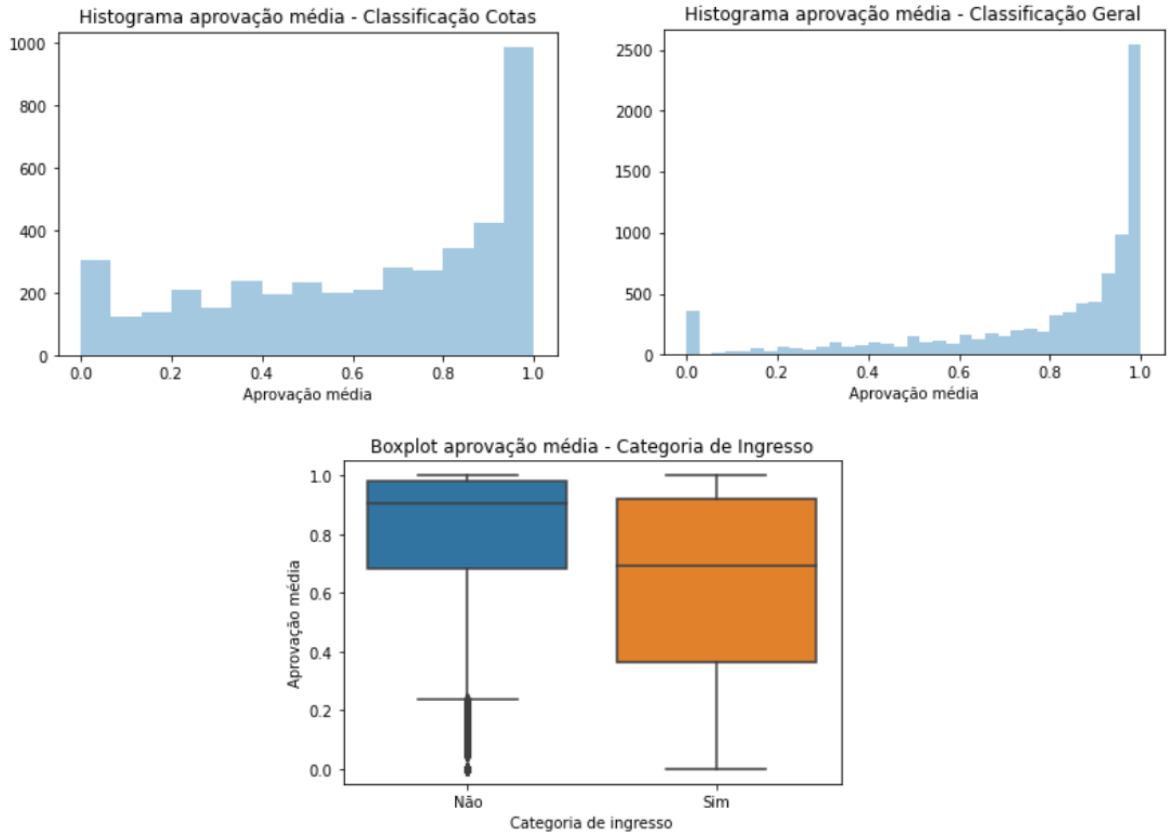




Fonte: elaborado pelo autor

3.1.3.3.3 Análise aprovação

Figura 16 - Histogramas e boxplot de alunos por aprovação por categoria de ingresso



Fonte: elaborado pelo autor

Atualmente, o sistema de cotas preenche 50% das novas vagas. Como esse valor foi incrementado durante os anos, o valor de 66% dos alunos em classificação geral provavelmente é consequente deste fato, além da maior taxa de abandono, que chega a 50% para os alunos classificados por cotas. No que tange ambos os valores de IAA e taxa de aprovação, os alunos classificados por cotas tem um desempenho em geral bastante abaixo da

classificação geral. Aproximadamente 50% dos alunos de cotas estão no grupo de abaixo de 70% de aprovação, o que de acordo com o tópico geral, quase 100% dos alunos que se formam. É forte, portanto, a probabilidade de que mais de 50% dos alunos de cotas evadirão.

A partir dos gráficos, de IAA e aprovação, vemos que o grupo de alunos classificados por cotas tem um comportamento gráfico mais similar ao de alunos que abandonaram o curso, enquanto o grupo de alunos de classificação geral é semelhante ao grupo de alunos que se formaram ou regulares.

3.1.3.4 Etnia

3.1.3.4.1 Dados gerais

A tabela 8 apresenta os dados gerais segmentados por cada categoria do campo etnia da base de dados. Destaca-se de primeira que o número de brancos é muito maior e, portanto, constitui uma base mais confiável para análises acerca de seu desempenho.

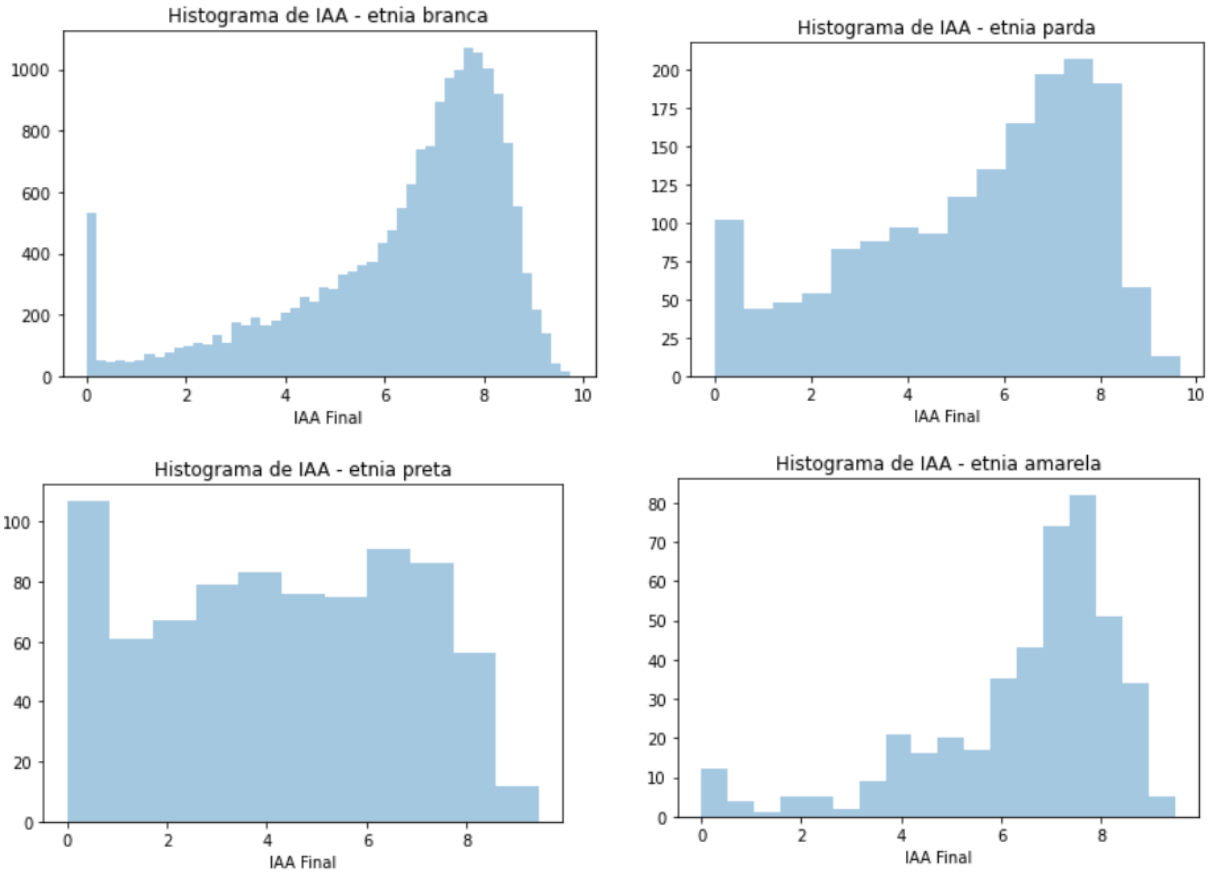
Tabela 8 - Dados gerais dos alunos por etnia

GÊNERO	MATRÍCULAS	TAXA DE APROVAÇÃO	MÉDIA IAA	FORMADO	REGULAR	DESISTENTES
BRANCA	71,5%	76%	6,27	47%	21%	32%
PARDA	6,7%	63%	5,26	30%	36%	34%
PRETA	3,2%	48%	4,05	20%	31%	49%
AMARELA	1,7%	80%	6,45	53%	22%	25%
INDÍGENA	0,2%	63%	5,28	55%	15%	30%
NÃO DECLARADA	0,1%	80%	6,62	0%	91%	9%
SEM INFORMAÇÃO	16,6%	77%	6,34	67%	1%	32%

Fonte: Elaborado pelo autor

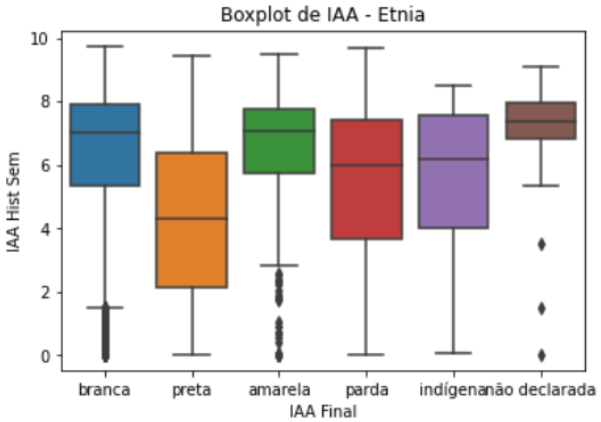
3.1.3.4.2 Análise IAA

Figura 17 - Histogramas de alunos por IAA por etnia



Fonte: elaborado pelo autor

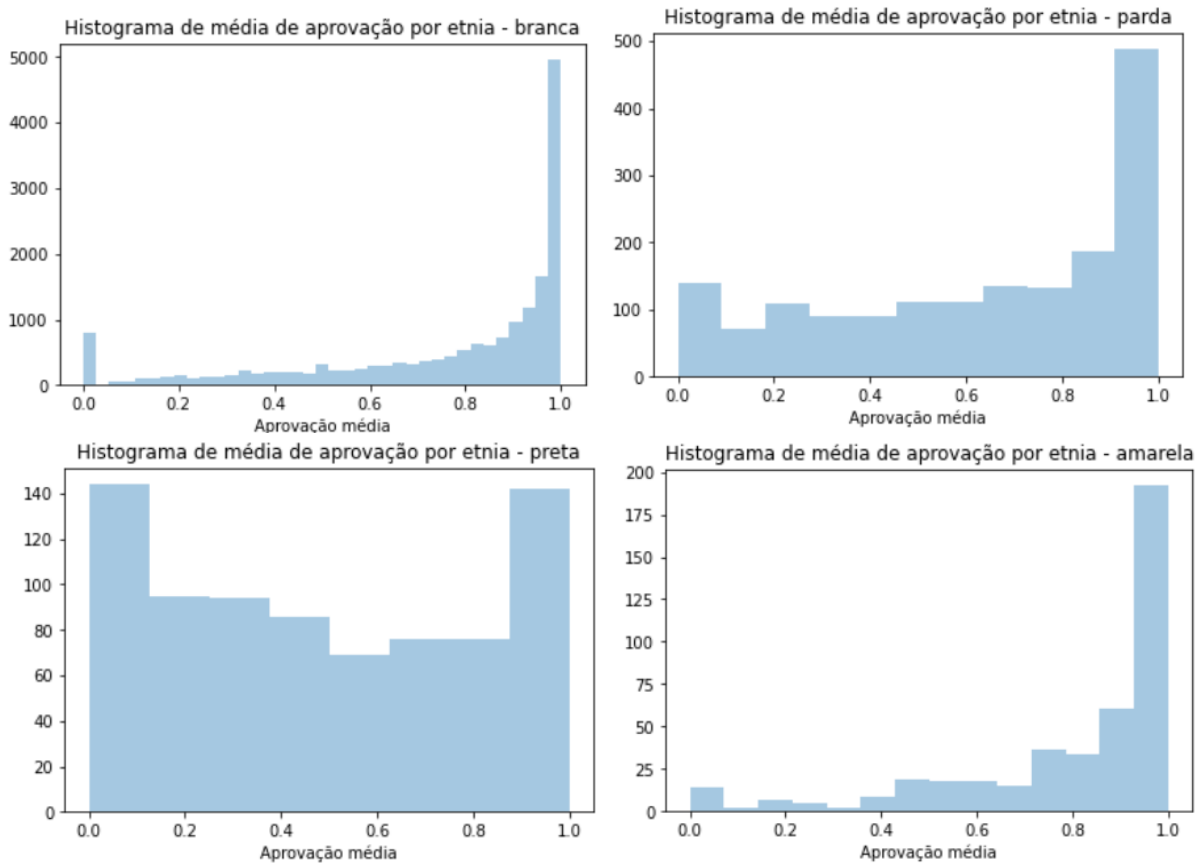
Figura 18 - Boxplot de IAA por etnia



Fonte: elaborado pelo autor

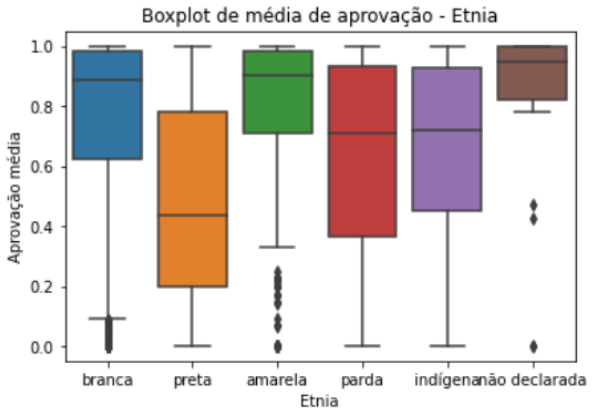
3.1.3.4.3 Análise aprovação

Figura 19 - Histogramas de aprovação média por etnia



Fonte: elaborado pelo autor

Figura 20 - Boxplot de aprovação média por etnia



Fonte: elaborado pelo autor

Assim como observado sobre o gênero dos discentes, a grande maioria dos alunos se identifica como branco, ultrapassando 70%. O número enxuto de alunos que se identificam como indígena ou não declarados é muito pouco expressivo, tornando difícil a sua avaliação

em histograma. Os grupos étnicos de pardos e pretos possui uma expressividade não negligenciável apesar de pequena de 6,7% e 3,2%, respectivamente.

Pela representatividade dos brancos na base, era esperado e se confirma que os valores apontados na tabela 8 são muito próximos dos valores gerais. Podemos simplificadamente alocar 3 clusters por nível de desempenho: em primeiro as etnias amarela, branca e não declarada, em segundo parda e indígena e, por fim, preta. Esses grupos apresentam resultados bastante similares tanto em IAA quanto em aprovação, exceto pelo fato de os indígenas terem uma taxa de formatura bastante elevada. Uma hipótese para explicar esse fato pode ser a de maior demora para formatura e menor taxa de desistência, mesmo dado esse fato, porém isso deve ser checado com os dados.

3.1.3.5 Categoria de ingresso X etnia

Assim como na análise de categoria de ingresso, os dados anteriores a 2008, quando não havia o sistema de cotas, foi excluída a fim de permitir uma comparação de dados mais confiável.

3.1.3.5.1 Dados gerais

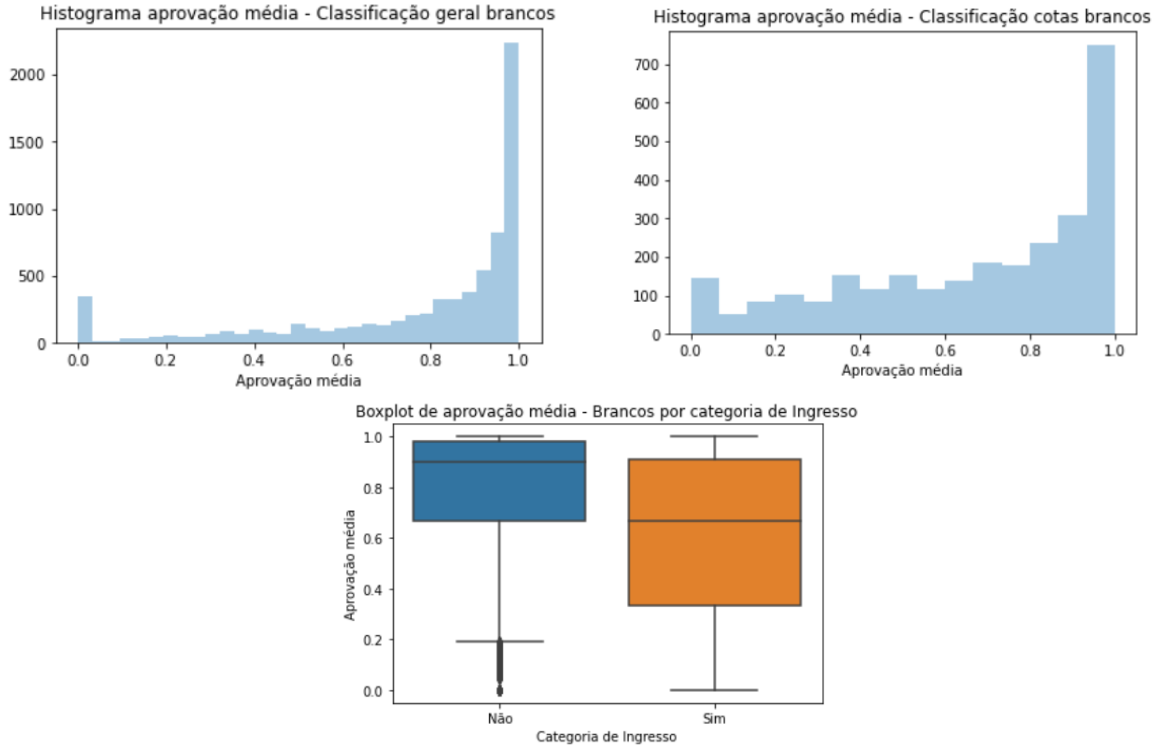
Tabela 9 - Dados gerais dos alunos por etnia, controlado por categoria de ingresso

CATEGORIA DE INGRESSO	MATRÍCULAS	TAXA DE APROVAÇÃO	MÉDIA IAA	FORMADO	REGULAR	DESISTENTES
BRANCA GERAL	74%	78%	6,40	34%	38%	28%
BRANCA COTA	26%	67%	5,55	20%	51%	29%
PARDA GERAL	76%	72%	6,09	32%	36%	32%
PARDA COTA	34%	52%	4,38	9%	56%	35%
PRETA GERAL	10%	69%	5,72	24%	41%	34%
PRETA COTA	90%	41%	3,50	10%	35%	55%

Fonte: Elaborado pelo autor

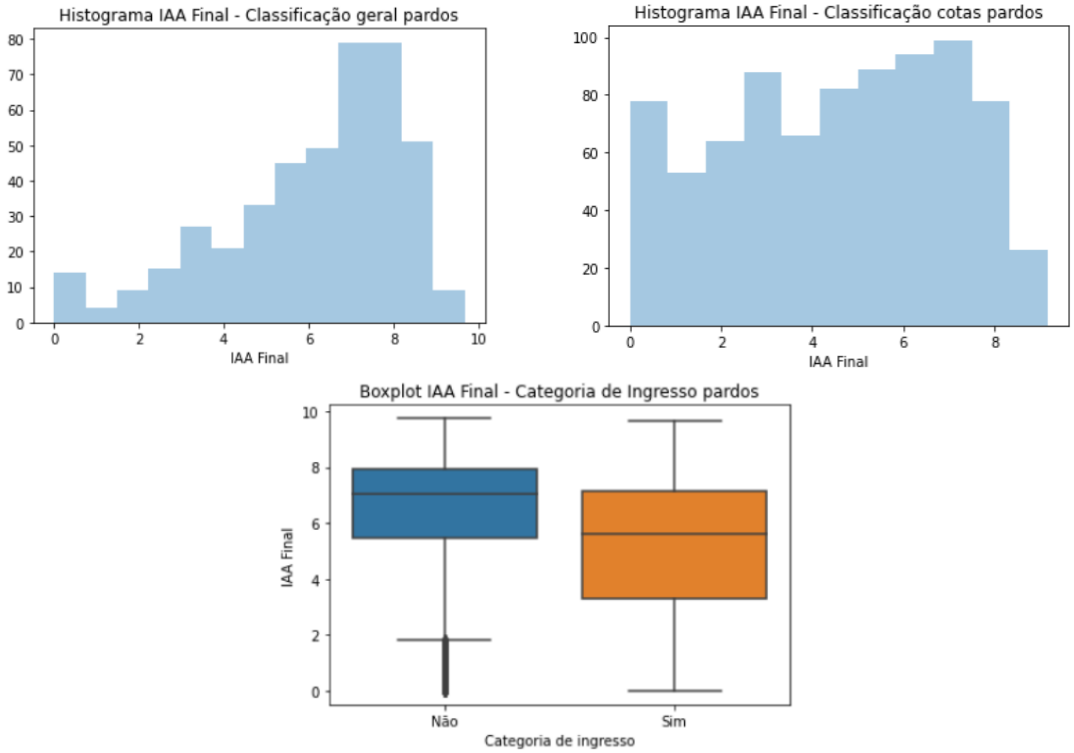
3.1.3.5.2 Análise IAA

Figura 21 - Histogramas e boxplot de IAA por classificação - etnia branca



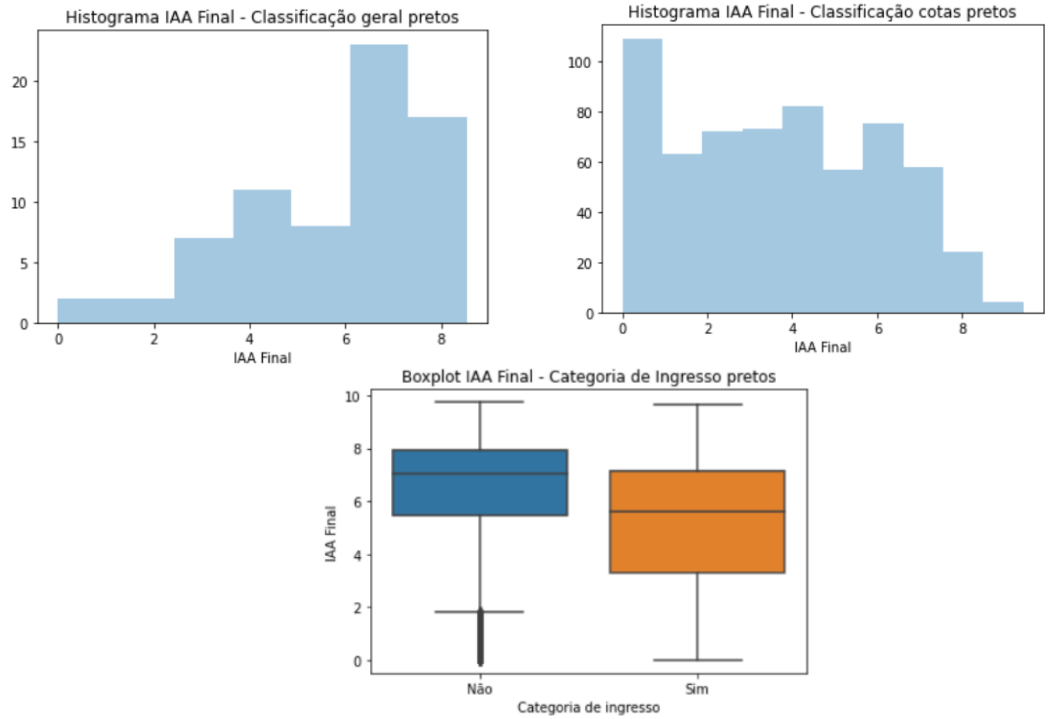
Fonte: elaborado pelo autor

Figura 22 - Histogramas e boxplot de IAA por categoria de ingresso e etnia parda



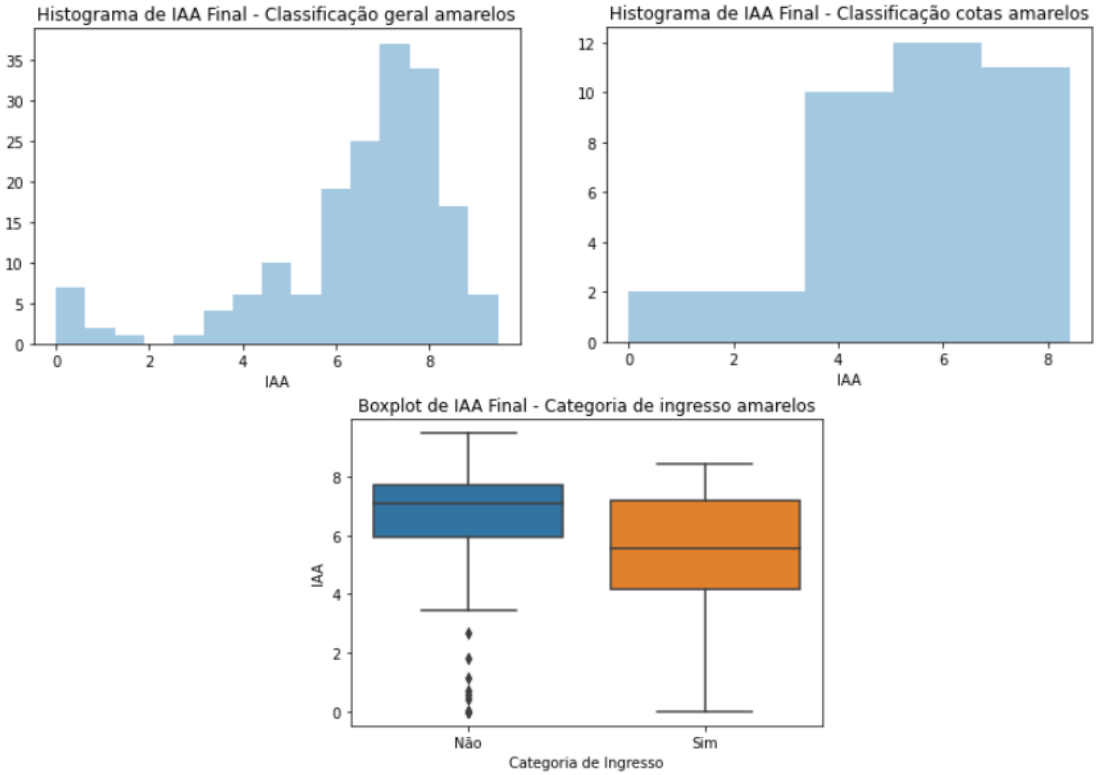
Fonte: elaborado pelo autor

Figura 23 - Histogramas e boxplot de IAA por categoria de ingresso e etnia preta



Fonte: elaborado pelo autor

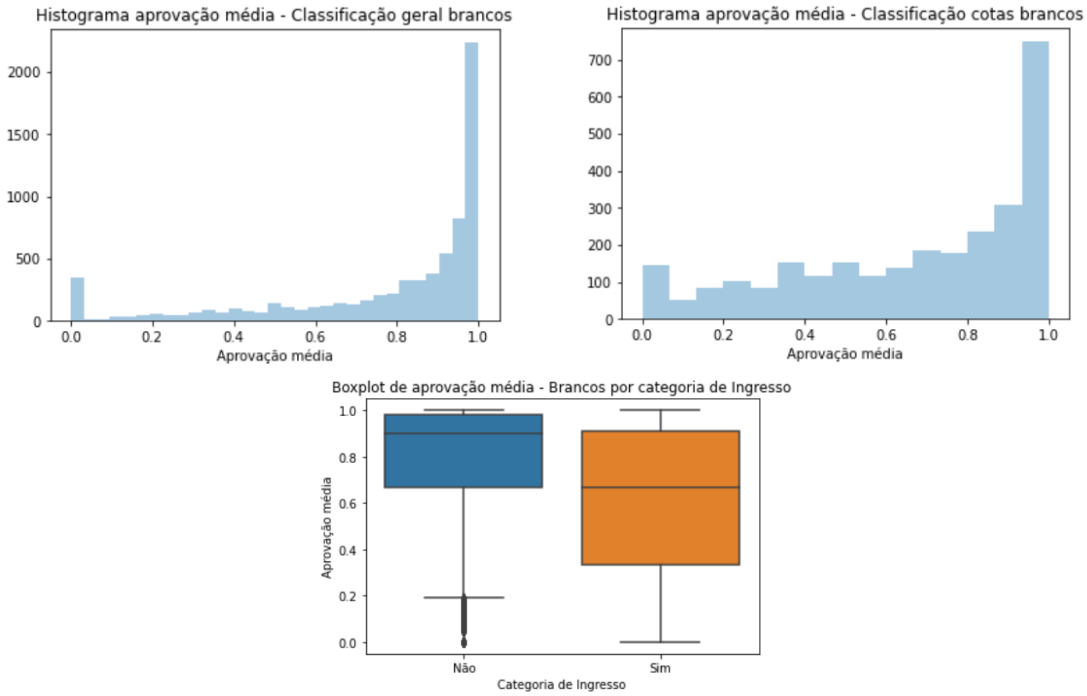
Figura 24 - Histogramas e boxplot de IAA por categoria de ingresso e etnia amarela



Fonte: elaborado pelo autor

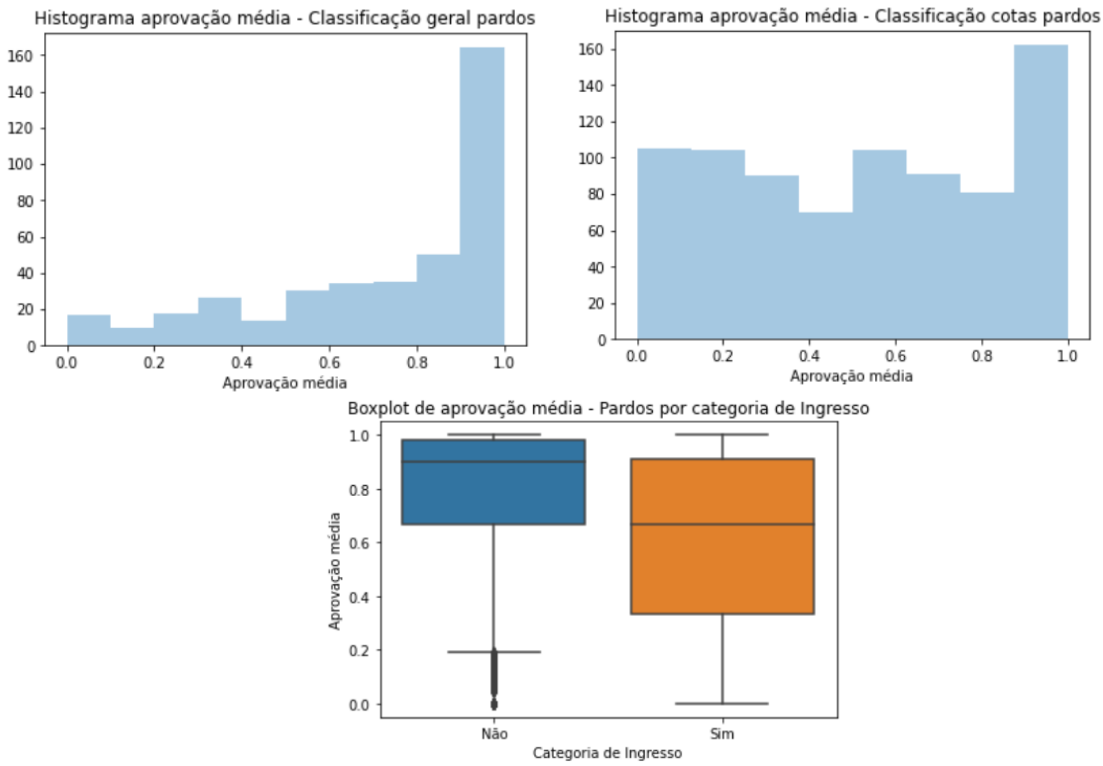
3.1.3.5.3 Análise aprovação média

Figura 25 - Histogramas e boxplot de aprovação média por categoria de ingresso e etnia branca



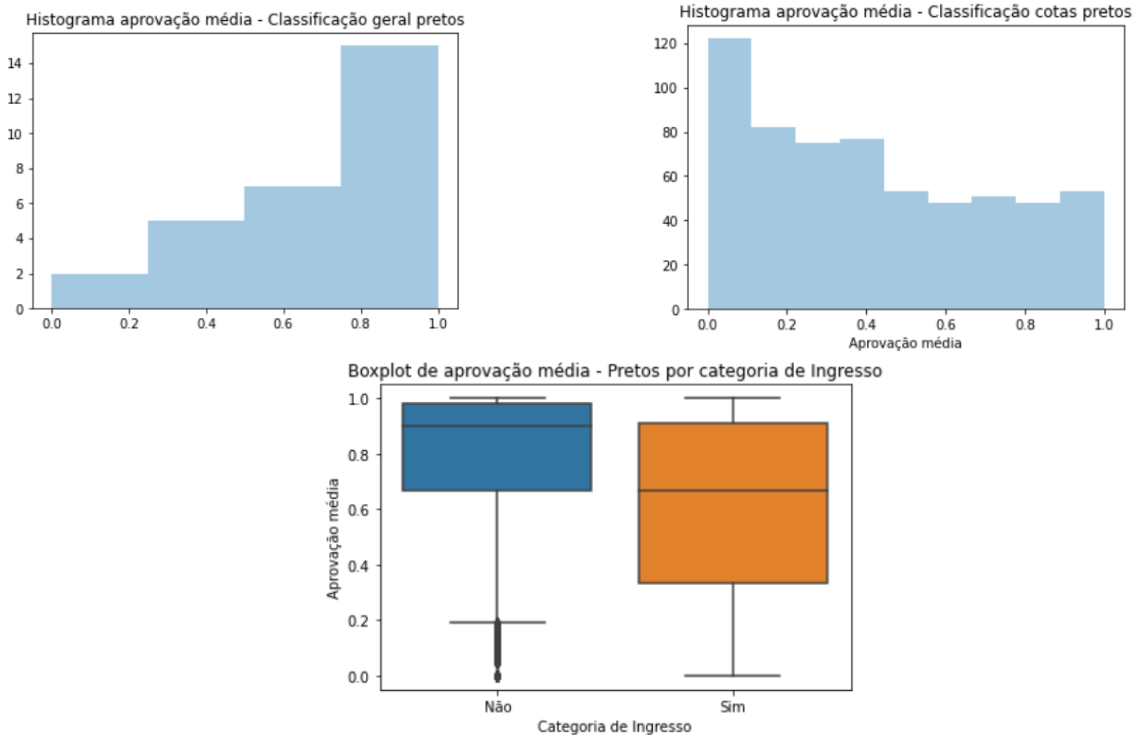
Fonte: elaborado pelo autor

Figura 26 - Histogramas e boxplot de aprovação média por categoria de ingresso e etnia parda



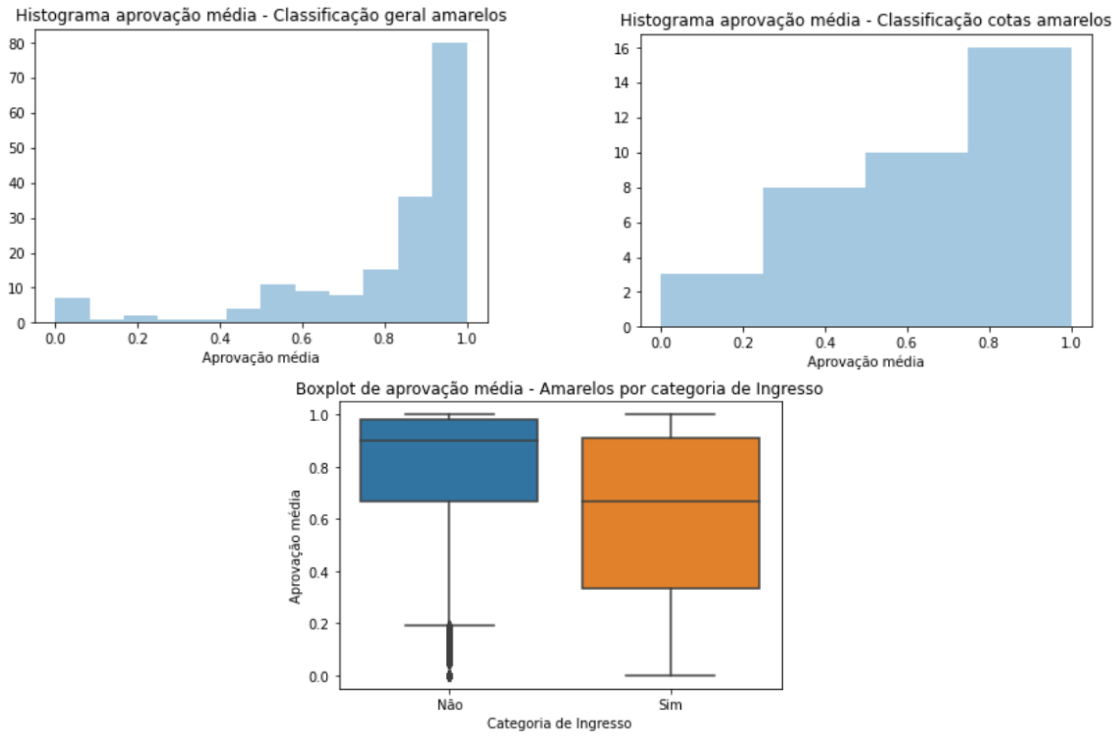
Fonte: elaborado pelo autor

Figura 28 - Histogramas e boxplot de aprovação média por categoria de ingresso e etnia preta



Fonte: elaborado pelo autor

Figura 27 - Histogramas e boxplot de aprovação média por categoria de ingresso e etnia amarela



Fonte: elaborado pelo autor

Da tabela 9, observa-se que enquanto os grupos de alunos brancos e pardos ingressaram majoritariamente pelo sistema de cotas, o grupo de alunos pretos ingressou quase

em sua totalidade por esse sistema. Existe também uma enorme disparidade de resultados entre alunos de classificação geral e cotas em cada grupo étnico, aparente em todos os indicadores de aprovação, IAA, formatura e desistência em todas as etnias. Essa questão fica muito bem ilustrada quando comparados os gráficos de histograma de classificação geral e cotas para cada etnia. O *boxplot* evidencia que no caso de todas as etnias, a mediana da classificação de cotas está aproximadamente no nível do primeiro quartil da classificação geral, ou seja, existe um *gap* razoável entre um público e outro.

Apenas as etnias branca e parda de classificação geral tem média acima do nível exigido na universidade e tem indicadores muito similares como um todo. Existe uma tendência de maior número de alunos regulares nos grupos de cotas, exceto pela etnia preta, que possui maior nível de evasão. Talvez essa situação seja explicada, conforme apontado anteriormente, que esse grupo de alunos demora mais tempo para se formar, já que sua taxa de aprovação também é substancialmente menor.

Pode-se concluir a partir do apresentado que, se controlado pela categoria de ingresso, a diferença entre as etnias é significativamente reduzida. Grande parte do diferencial é explicado, apesar de não poder explicar toda a diferenciação. A população preta ingressa na universidade majoritariamente por meio do sistema de cotas, estando excluída dos cursos sem esse artifício, mas uma vez dentro da universidade parece ter muita dificuldade de se formar, com uma desistência evasão atual maior que 50% e uma taxa de formatura muito reduzida.

3.1.3.6 Idade

3.1.3.6.1 Dados gerais

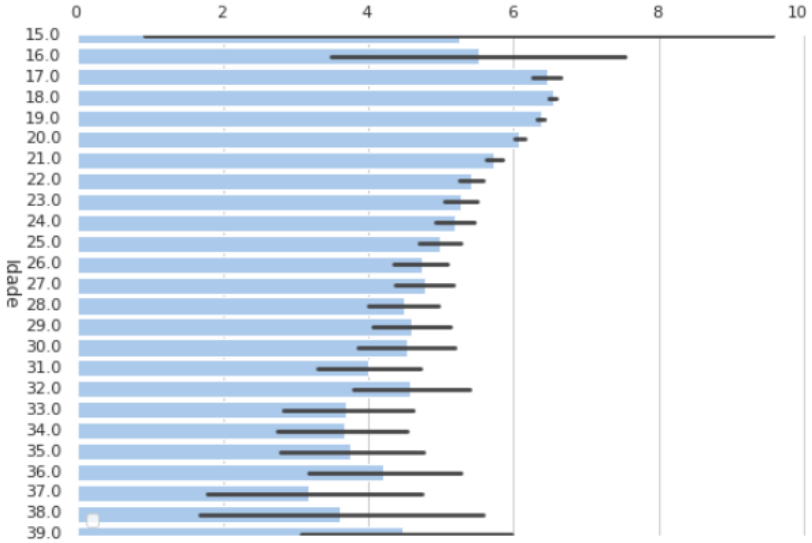
Tabela 10 - Dados gerais dos alunos por faixa de idade

CATEGORIA DE INGRESSO	MATRÍCULAS	TAXA DE APROVAÇÃO	MÉDIA IAA	FORMADO	REGULAR	DESISTENTES
< 22	86%	72%	6,02	51%	21%	28%
>= 22 < 28	11%	61%	5,23	35%	28%	37%
>= 28	3%	42%	3,98	20%	32%	48%

Fonte: Elaborado pelo autor

3.1.3.6.2 Análise IAA

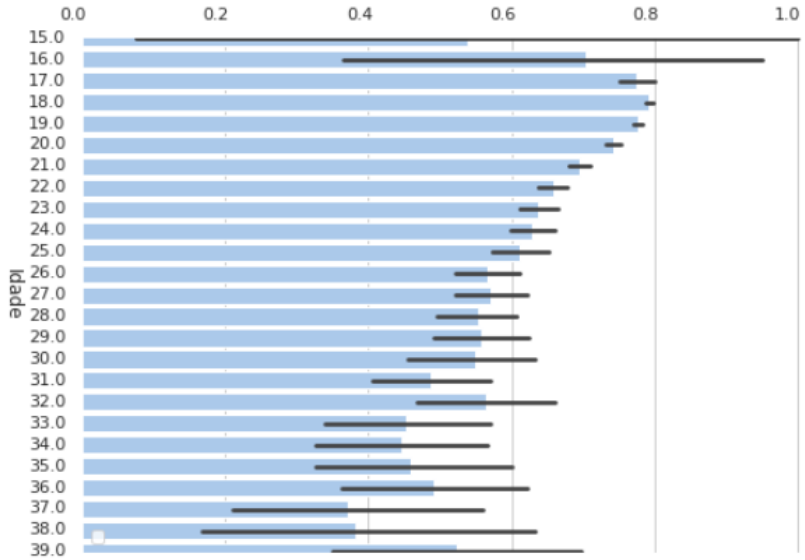
Figura 29 - Média e dispersão de IAA por idade



Fonte: elaborado pelo autor

3.1.3.7 Análise aprovação

Figura 30 - Média e dispersão de média de aprovação por idade



Fonte: elaborado pelo autor

O grupo de alunos que fizeram o vestibular com 15 a 21 anos compreende 87% dos alunos. Esse também é o grupo com melhor desempenho apresentado em nota média assim como média de aprovações. Existe uma clara tendência indicando que maiores idades desempenham pior. Essa tendência não se mostra nas idades de 15 e 16 anos, mas esses grupos são pouco representativos e são os que indicam maior dispersão, ou seja, apesar da média baixa, existe um nível de discentes com resultados bastante elevados e baixos.

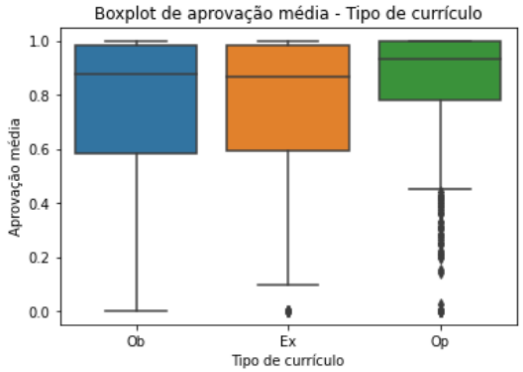
3.1.3.8 Tipo currículo

Tabela 11 - Dados gerais do desempenho dos alunos por classificação da matéria no currículo

CATEGORIA DE INGRESSO	DISCIPLINAS	TAXA DE APROVAÇÃO	MÉDIA NOTA
OBRIGATÓRIA	88,9%	82%	6,65
OPTATIVA	8,2%	88%	7,43
EXTRA	2,9%	78%	6,53

Fonte: Elaborado pelo autor

Figura 31 - Boxplot de aprovação média por categoria de matéria no currículo



Fonte: elaborado pelo autor

A grande maioria de disciplinas cursadas é obrigatória, indicando que há pouca carga horária opcional nos cursos analisados. Disciplinas optativas possuem uma aprovação e média acima das obrigatórias e extras, ambas as quais possuem um desempenho muito similar.

3.1.3.9 Repetição de matéria

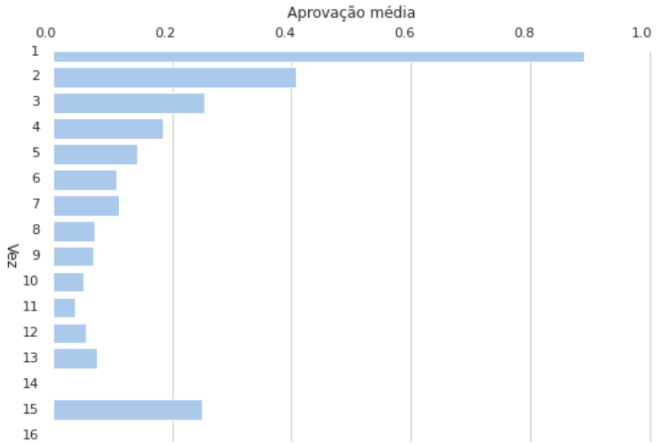
Foi elaborado um rastreamento para contabilizar alunos que foram reprovados e se matricularam novamente na mesma matéria. Esse é o indicador que será analisado no presente tópico.

Tabela 12 - Dados gerais por vez matriculado na matéria

VEZ	DISCIPLINAS	TAXA DE APROVAÇÃO	NOTA
PRIMEIRA	88%	89%	7,14
SEGUNDA	8%	41%	4,09
TERCEIRA	2%	26%	2,95
QUARTA	1%	19%	2,31
> 5	1%	12%	1,65

Fonte: elaborado pelo autor

Figura 32 - Aprovação média por vez em matriculado na mesma matéria



Fonte: Elaborado pelo autor

Existe uma queda visível de desempenho dos alunos a cada vez que o aluno cursa a mesma matéria. Conforme apontado anteriormente, alunos que se formaram estão quase todos com uma aprovação de acima de 70%, indicando que esse grupo de alunos em grande medida cursa as suas disciplinas apenas uma vez. Já alunos que não se formam, tem taxas de aprovação muito menores, sendo que aproximadamente 50% dos alunos tem taxas de aprovação igual ou menores do que 50%, o que indica que esses alunos devem cursar muitas disciplinas mais de uma vez. Enquanto para algumas disciplinas a maioria deve passar diretamente, outras disciplinas apresentam uma dificuldade muito maior, sendo assim, para chegar à taxa de 50% ou abaixo de aprovação média, esses alunos devem reprovar essas disciplinas mais de uma vez.

3.1.3.10 Contagem de disciplinas

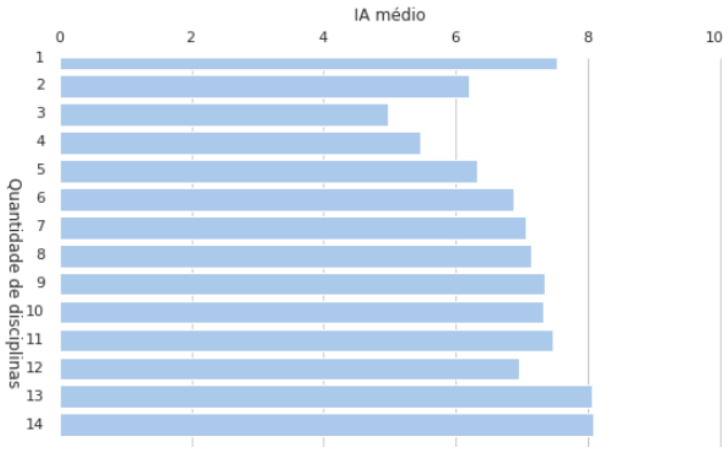
Esse tópico é referente à quantidade de disciplinas cursadas pelo aluno num semestre específico.

Tabela 13 - Dados gerais por quantidade de disciplinas cursada em um semestre

CONTAGEM DE DISCIPLINAS	MATRÍCULAS	TAXA DE APROVAÇÃO	IA
6	31%	85%	6,87
7	22%	88%	7,01
5	18%	78%	6,32
8	11%	89%	7,14
4	6%	78%	5,46
< 4	5%	71%	6,22
9	4%	92%	7,33
> 9	3%	97%	7,58

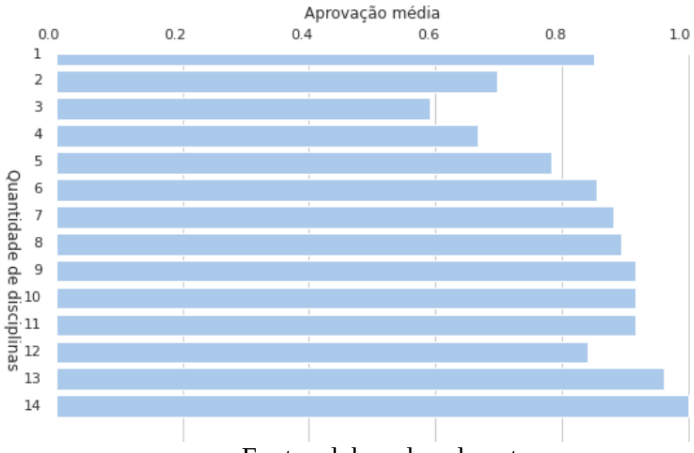
Fonte: Elaborado pelo autor

Figura 33 - Gráfico de barras de IA médio de alunos por quantidade de disciplinas



Fonte: elaborado pelo autor

Figura 34 - Gráfico de barras de aprovação média de alunos por quantidade de disciplinas



Fonte: elaborado pelo autor

Conforme indicado na tabela 13, 82% dos casos se encontra dentro da faixa de 5 a 8 disciplinas. Dentro dessa faixa, quanto menos disciplinas sendo cursadas, pior o desempenho. Esse fato é diferente do que se imaginaria, pois, uma possibilidade seria que um maior número de tarefas, maior a dificuldade de lidar com a carga de trabalho. Nos modelos será

testado o efeito de disciplinas contraposto com a carga horária que essas disciplinas representam.

3.1.3.11 Disciplinas

O aspecto desse tópico representa a taxa de aprovação média de uma disciplina específica.

Tabela 14 - Dados gerais referentes a disciplinas

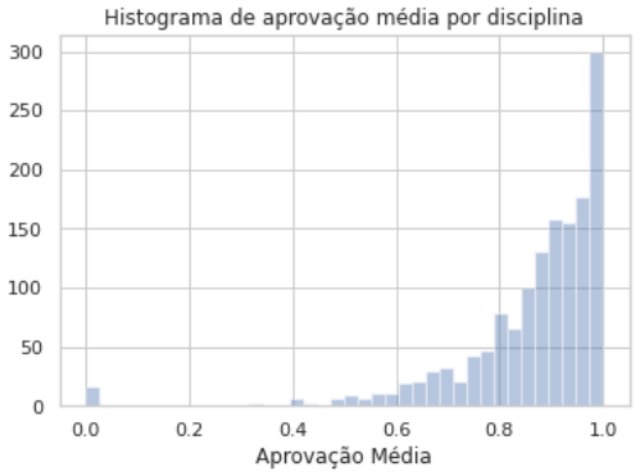
	< 4	**4 < X < 6**	**6 < X < 8**	**> 8**
APROVAÇÃO MÉDIA	7%	52%	72%	93%
NOTA MÉDIA	1,31	5,37	7,10	8,68

Fonte: Elaborado pelo autor

Figura 35 - Histograma de aprovação média de disciplinas

Fonte: elaborado pelo autor

Figura 36 - Histograma de aprovação média de disciplinas



Fonte: elaborado pelo autor

Existe uma diferença razoável de taxa de aprovação das disciplinas, mas a maioria possui taxas de aprovação mais elevadas. Em geral as notas médias estão no intervalo de 6 a 8,5.

3.1.3.12 Professores

O aspecto desse tópico representa a taxa de aprovação média dos professores. No total são 2115 matrículas distintas.

Tabela 15 - Dados gerais referentes a desempenho dos alunos para um respectivo professor

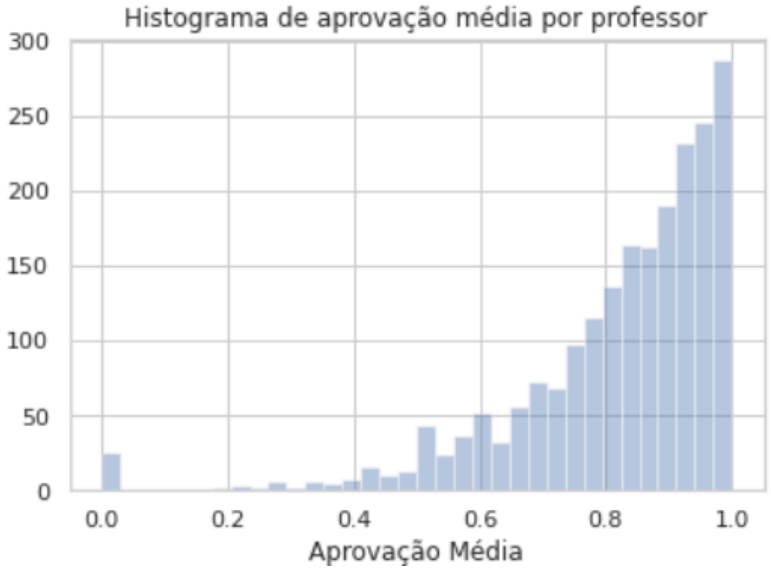
	X < 4	**4 < X < 6**	**6 < X < 8**	**> 8**
APROVAÇÃO MÉDIA	3%	8%	24%	65%
NOTA MÉDIA	4%	21%	57%	18%

Fonte: elaborado pelo autor

Figura 37 - Histograma de nota média de professores

Fonte: Elaborado pelo autor

Figura 38 - Histograma de aprovação média de professores

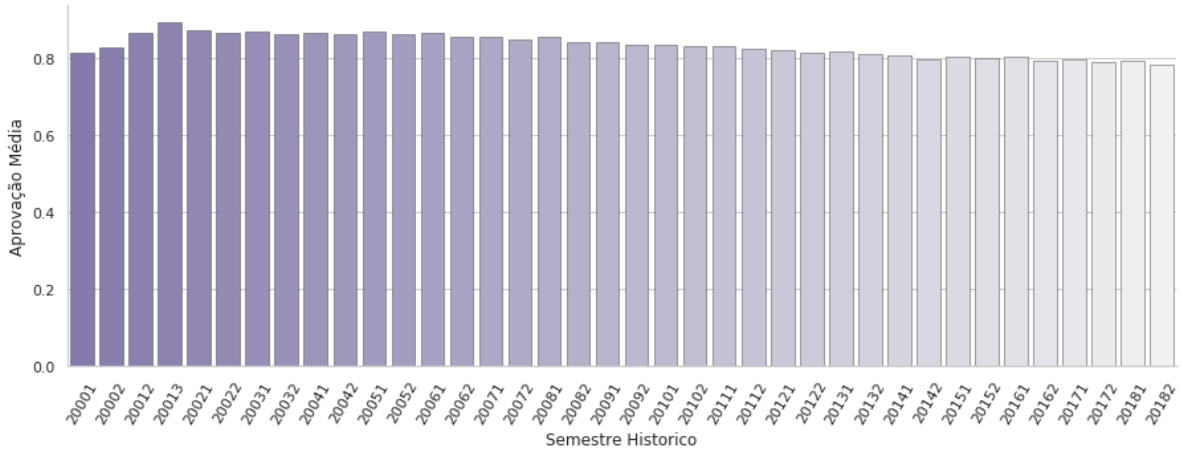


Fonte: elaborado pelo autor

A análise das aprovações e notas médias dos professores é muito similar com o desempenho dos alunos nas disciplinas. É relevante observar que ambos os gráficos são um pouco mais assimétricos, deslocando os resultados mais à esquerda, do que os de disciplinas, ou seja, possuem um comportamento similar, mas pior do que as disciplinas. Essa situação pode ser explicada para uma diferenciação entre professores dentro de uma única disciplina.

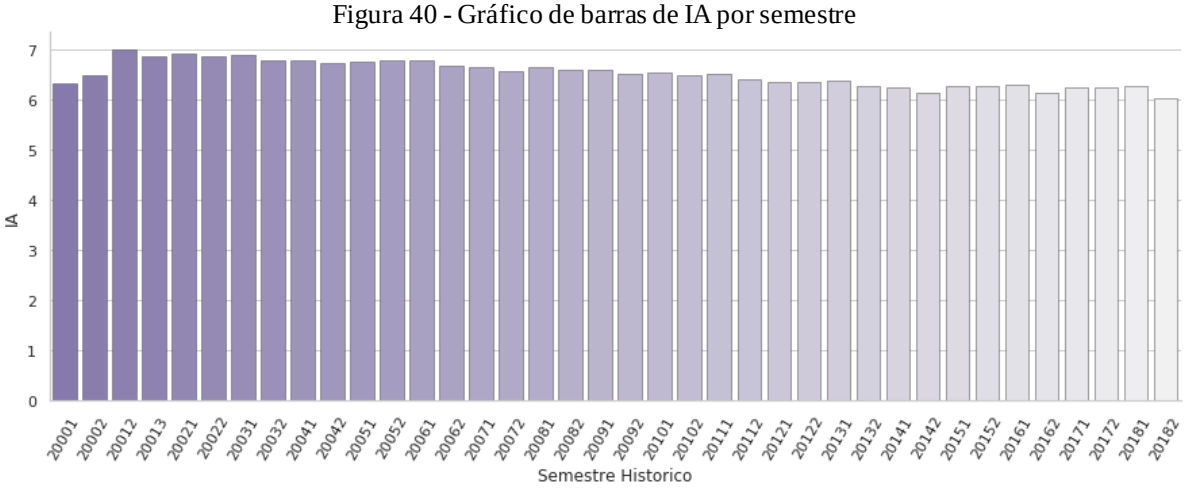
3.1.3.13 Histórico

Figura 39 – Gráfico de barras de aprovação média por semestre



Fonte: elaborado pelo autor

Figura 40 - Gráfico de barras de IA por semestre



Fonte: elaborado pelo autor

Os gráficos acima indicam as médias de aprovação e IA agregadas por semestre. Dentro do processo histórico, existe uma clara tendência negativa de resultado em ambos os casos. A variação da média de aprovação foi por volta de 5% enquanto a variação de IA foi por volta de 1 ponto.

3.1.4 Feature engineering

Conforme descrito no item 2.1.7. foram executados os processos de elaboração, processamento e elaboração das *features* para se obter o agrupamento final de *features* imputadas no modelo preditivo. Cabe ressaltar que, por conta da agilidade computacional

apresentada pelo modelo de florestas aleatórias, a etapa de *feature selection* foi feita a partir do mesmo, fazendo a filtragem por meio da importância das variáveis nesse modelo específico.

Segue abaixo a tabela de variáveis resultante e que será utilizada para a previsão:

Tabela 16 - Descrição das *features* inseridas nos modelos

Variável	Descrição	Cálculo
Gênero	Gênero cadastrado no perfil do aluno	Binário: 0 para masculino 1 para feminino
Etnia_branca	Etnia cadastrada como branca no perfil do aluno	Binário: 0 para não branca 1 para branca
Etnia_indígena	Etnia cadastrada como indígena no perfil do aluno	Binário: 0 para não indígena 1 para indígena
Etnia_parda	Etnia cadastrada como parda no perfil do aluno	Binário: 0 para não parda 1 para parda
Etnia_preta	Etnia cadastrada como preta no perfil do aluno	Binário: 0 para não preta 1 para preta
Idade no vestibular	Idade cadastrada no perfil do discente no momento do vestibular	-
Categoria de Ingresso	Categoria de ingresso do perfil do aluno	Binário: 0 para classificação geral 1 para classificação por cotas
Horas Aula	Cadastro de número de horas-aula da disciplina	Somatório de horas-aula semestrais
Tipo currículo	Categoria da disciplina perante o seu currículo. Pode ser obrigatória, optativa ou extracurricular	-
IA	Índice de aproveitamento	Pontos obtidos no somatório dos produtos das notas pelas cargas horárias matriculadas de um determinado semestre
IAA	Índice de aproveitamento acumulado	Pontos obtidos no somatório dos produtos dos IAs pelas cargas horárias somadas dos semestres
IAA -1	Diferença de aproveitamento entre os 2 semestres anteriores	Subtração do IA do semestre pelo IA do semestre anterior
Taxa de aprovação do aluno	Percentual de aprovação histórica do aluno	Número de aprovações do aluno / número total de disciplinas cursadas

Taxa de aprovação do aluno - 3	Percentual de aprovação dos 3 semestres anteriores cursados pelo aluno	Número de aprovações do aluno / número total de disciplinas cursadas
Taxa de aprovação do aluno - 1	Percentual de aprovação do semestre anteriores cursados pelo aluno	Número de aprovações do aluno / número total de disciplinas cursadas
Taxa de aprovação do professor	Percentual de aprovação histórica do professor	Número de aprovações dos alunos que cursaram uma disciplina do professor / número total de alunos que cursaram suas disciplinas
Taxa de aprovação da disciplina	Percentual de aprovação histórica da disciplina	Número de aprovações de alunos na disciplina / número total de alunos que cursaram a disciplina
Pontos Vestibular	Pontuação obtida pelo aluno na prova do vestibular	-
Pontos SISU	Pontuação obtida pelo aluno na prova do ENEM	-
FI	Quantidade de frequências insuficientes	Número de frequências insuficientes / Total de disciplinas cursadas
Contagem de disciplinas	Número de disciplinas que o aluno está matriculado no respectivo semestre	Somatório de disciplinas da matrícula no mesmo semestre
Carga horária	Carga horária total cursada no respectivo semestre	Somatório das horas aula referentes a todas as disciplinas no mesmo semestre
Repetência de matérias	Número de vezes que cursou a disciplina em questão	Contagem das vezes que cursou a disciplina

Fonte: Elaborado pelo autor

3.1.5 Validação cruzada

Conforme apresentado no tópico 2.1.4. a validação cruzada do presente trabalho tem a especificidade de ser dependente temporalmente. O que isso significa é que não se pode utilizar uma divisão simples do conjunto total de dados da base, por exemplo, de 30% e 70%. A maneira adotada para lidar com a questão foi fazer essa divisão temporal onde se use como base de treino todos os semestres anteriores ao semestre a ser testado. Sendo assim, foi utilizado como primeira base de teste o semestre de 2015.1. Isso significa que o modelo foi treinado a partir de todos os dados de 2000.1 a 2014.2, e as *features* do semestre 2015.1 foram calculadas com base na mesma amostra.

3.1.6 Tunagem de hiperparâmetros

Para a tunagem foram escolhidos alguns hiperparâmetros do modelo de florestas aleatórias. Esses parâmetros, conforme descrito na seção 2, não são facilmente definidos teoricamente, sendo necessária uma abordagem empírica ao iterar entre vários valores e avaliando o resultado obtido em cada situação para a escolha da melhor configuração.

Como instruído por Koehrsen (2018) foram testados os seguintes hiperparâmetros, com os valores das respectivas listas:

Hiperparâmetros do modelo de árvores aleatórias

1) Quantidade de árvores geradas
2) Quantidade de features consideradas para cada divisão dos nós
3) Número máximo de hierarquia
4) Número mínimo de instâncias alocadas para um nó antes da divisão
5) Número mínimo de instâncias permitidas por folha
6) Método de amostragem de instâncias

Hiperparâmetros do modelo de regressão logística.

1) Parâmetro C
2) Número de iterações

A escolha de um dado valor para um parâmetro específico pode alterar significativamente o valor ótimo de outro. Dessa maneira, as possibilidades de solução são muitas. Idealmente, conforme sugerido por Koehrsen, existe uma abordagem interessante na qual se realiza um processamento de combinações aleatórias dentro de listas de valores para cada parâmetro. Essa busca não é exaustiva, mas aponta para qual faixa de valores pode ser mais adequada. Em posse de uma faixa de valores mais estreita, pode-se aplicar uma abordagem de listas menores, de menor dispersão de valores, e aplicada exaustivamente, para a definição dos valores exatos dos hiperparâmetros do modelo. O presente trabalho, porém, é executado com uma base de dados relativamente grande e a aplicação do modelo de florestas aleatórias é relativamente exigente em nível de processamento. Essas características fazem com que a abordagem optada foi de treinar as variáveis do modelo controlando o parâmetro de quantidade de árvores geradas e posteriormente ajustado com os demais parâmetros fixos.

É importante ressaltar que esse teste de parâmetros deve ter uma função objetiva a fim de comparar as alternativas e mensurar o desempenho, escolhendo assim a alternativa de melhor desempenho.

3.1.7 Avaliação dos modelos

Cada modelo tem seu objetivo específico e ele direciona a definição das configurações. No capítulo 2 foi apresentado um ferramental diverso de possibilidades para avaliar os modelos a fim de trazer uma análise holística sobre os resultados. Assim, serão apresentados os objetivos de cada modelo, o ferramental adequado para mensurar e avaliar o modelo e as decisões intermediárias.

4 RESULTADOS

Após a elaboração e refinamento do modelo, esta seção é dedicada a uma análise dos resultados obtidos para avaliação da sua possibilidade de aplicação e confiabilidade. Conforme descrito no capítulo 3 e fazendo uso das ferramentas apresentadas no capítulo 2, os resultados serão avaliados conforme as métricas de avaliação de modelo. Na sequência, serão apresentadas as *feature importances* dos modelos.

4.1FLORESTAS ALEATÓRIAS - PREVISÃO DE REPROVAÇÃO

O primeiro modelo tem o objetivo de prever os casos mais críticos de reprovação. Ele pode ser utilizado de algumas maneiras, mas foi definido que o valor a ser perseguido seja de 500 casos. Optou-se por esse número arbitrariamente por conta de ser um número razoável para a criação de um programa que pudesse apoiar tais alunos de alguma forma. Assim, a função objetiva escolhida para esse modelo configurar esse modelo foi a precisão em K para o valor de K = 500.

Para a definição dos valores dos hiperparâmetros, conforme apontado no capítulo 3, inicialmente foi fixado o valor de quantidade de árvores geradas em 100 e aplicado as seguintes listas de valores para cada parâmetro. Por fim, estabeleceu-se os demais parâmetros como fixos e aumentado o valor de árvores. A primeira etapa consiste em 18 alternativas que utilizando *cross validation* 2 *fold* soma 36 testes. No segundo momento são 3 tentativas, com *cross validation* 2-*fold*, totalizando 42 alternativas testadas. A tabela 17 aponta os valores testados e o valor da melhor configuração:

Tabela 17 - Definição de hiperparâmetros do modelo de florestas aleatórias

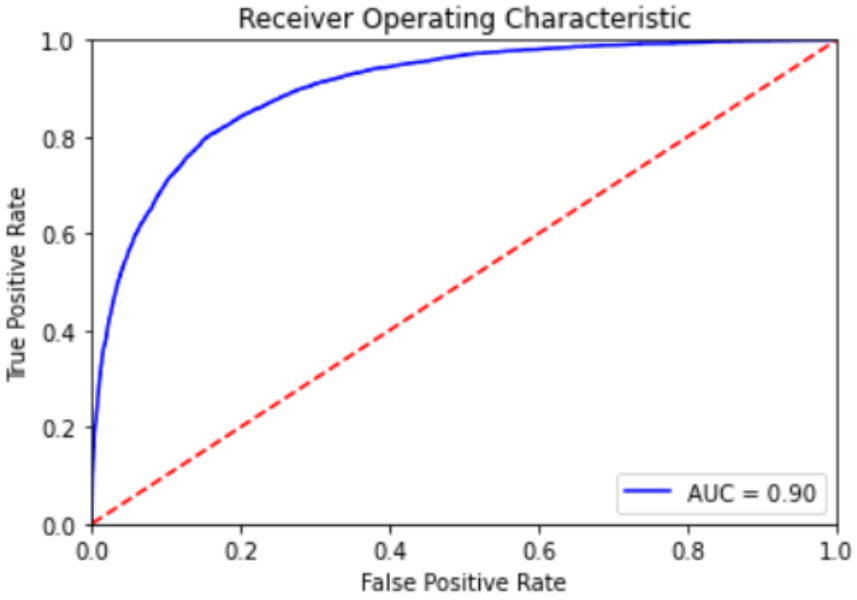
Parâmetro	Valores testados	Valor final
Quantidade de árvores geradas	[100, 200, 500]	500
Quantidade de *features* consideradas para cada divisão nos nós	Automático	Automático
Número máximo de hierarquia	[10, 20, 30]	30

Número mínimo de instâncias alocadas para um nó antes do da divisão	[5, 10]	10
Número mínimo de instâncias permitidos por folha	[3, 6, 10]	3
Método de amostragem de instâncias	*Bootstrap*	*Bootstrap*

Fonte: Elaborado pelo autor

A partir dessa definição, pode-se treinar o modelo com os parâmetros definidos. Desse ponto, é possível a determinação da curva ROC, ilustrada pela figura 40:

Figura 41 - Curva ROC do modelo de florestas aleatórias



Fonte: elaborado pelo autor

Os valores das métricas principais derivadas da matriz de confusão, principalmente as de precisão e sensibilidade que compõem o gráfico, estão apresentadas na tabela 18 para cada valor discreto de limiar de decisão:

Tabela 18 - Métricas de avaliação para cada limiar de decisão

Limiar de decisão	Classe	Precisão	Especificidade	F1 - Score	P@K
0,9	Reprovado	0,41	0,94	0,57	0,23
	Aprovado	0,85	0.61	0,75	
	Média ponderada	0,85	0,68	0,71	
0,8	Reprovado	0,52	0,86	0,65	0,42
	Aprovado	0,95	0,77	0,85	
	Média ponderada	0,85	0,79	0,80	
0,7	Reprovado	0,61	0,79	0,69	0,56
	Aprovado	0,93	0,85	0,89	
	Média ponderada	0,86	0,84	0,84	
0,6	Reprovado	0,67	0,70	0,69	0,68
	Aprovado	0,91	0,90	0,91	
	Média ponderada	0,86	0,86	0,86	
0,5	Reprovado	0,74	0,61	0,67	0,75
	Aprovado	0,89	0,94	0,91	
	Média ponderada	0,86	0,86	0,86	
0,4	Reprovado	0,80	0,50	0,62	0,82
	Aprovado	0,87	0,96	0,91	
	Média ponderada	0,85	0,86	0,85	
0,7	Reprovado	0,86	0,37	0,52	0,87
	Aprovado	0,84	0,98	0,91	
	Média ponderada	0,85	0,84	0,82	
0,2	Reprovado	0,91	0,23	0,37	0,91
	Aprovado	0,81	0,99	0,90	
	Média ponderada	0,82	0,82	0,78	
0,1	Reprovado	0,96	0,12	0,22	0,96
	Aprovado	0,80	1,00	0,89	
	Média ponderada	0,83	0,80	0,73	

Fonte: Elaborado pelo autor

Assim como comentado no capítulo 2, a escolha do limiar de decisão deve levar em conta não apenas uma métrica, mas fazer a tentativa de analisar holisticamente qual o resultado de cada métrica. Como 77% das instâncias, que significam um aluno em uma

disciplina, da base de teste foram aprovadas, ou seja, a base é bastante assimétrica o que faz com que a média ponderada sempre tenda a essa classe, que por sua vez é mais simples de ser prevista. Consequentemente, quase todos limiares de decisão alcançam um patamar razoável de precisão de aprovação e um nível controlado de especificidade e até mesmo o F1. Por conta desses pontos, optou-se por se focar para a escolha do limiar de decisão baseando-se principalmente na precisão, especificidade e precisão@500 de reprovação. Pela dificuldade de se prever uma reprovação, não é de se surpreender que os limiares de decisão que flexibilizem o valor de classificação de reprovação e tendem a aumentar o número de verdadeiros positivos, ao mesmo tempo que aumentando o número de falsos positivos. Essa questão perpassa todos os valores da tabela. A partir do limiar de 0,5, é contata-se um acerto de 75% dentre os piores 500 valores, o que é uma marca interessante, chegando a 87% com o limiar de 0,3. Na mesma comparação, as métricas de precisão e especificidade se modificam de 0,74 a 0,86 e 0,61 a 0,37, respectivamente.

Existem muitas opções de utilização e para cada uma poderia ser escolhido um limiar diferente, mas nesse caso, considera-se um 80% um nível interessante de precisão e 50% um nível controlado de erros positivos. Para uma série de utilizações, conforme será apontado, é mais crítico que se encontre os verdadeiros positivos, mesmo com erros, pois o tratamento será sistêmico e não estritamente individual. Ao mesmo tempo, conforme algumas pessoas serão classificadas como falsos positivos, podem ter alta probabilidade de ser reprovado e seria interessante um apoio adicional. Agora, a especificidade do problema também é reguladora, pois os recursos são finitos e a universidade não disponibiliza dos recursos para tratar a todos os preditos positivos, portanto esse é o *trade-off* a ser considerado e acredita-se que o nível de 0,5 de taxa de falsos positivos seja aceitável.

Tabela 19 - Matriz de confusão para primeira calibragem do modelo de florestas aleatórias

	Negativo Previsto	*Positivo previsto*
Negativo Real	17.269	2.696
Positivo real	1.274	4.697

Fonte: Elaborado pelo autor

A título de comparação a matriz de confusão para o limiar de decisão de 0,7, ou seja, para identificação dos casos mais críticos possui um número de falsos negativos bastante

grande, mas uma taxa de falsos positivos extremamente baixa, sendo especificamente muito interessante para a identificação de casos mais críticos:

Tabela 20 - Matriz de confusão para segunda calibragem do modelo de florestas aleatórias

	Negativo Previsto	*Positivo previsto*
Negativo Real	19.989	355
Positivo real	3.756	2.215

Fonte: Elaborado pelo autor

4.2 REGRESSÃO LOGÍSTICA - PREVISÃO DE REPROVAÇÃO

O modelo tem o objetivo de prever os casos mais críticos de reprovação. A fim de comparabilidade com o primeiro modelo, foi definido que o valor a ser perseguido seja de 500 casos. Assim, a função objetiva escolhida para esse modelo configurar esse modelo foi a precisão em K para o valor de K = 500.

Como apenas dois parâmetros foram calibrados e o modelo de regressão logística é menos exigente computacionalmente, testou-se mais valores para cada parâmetro e foi feita um teste exaustivo das alternativas de combinações, somando 28 combinações e, contando com um cross validation de 2 fold, somando 56 alternativas. A tabela 21 aponta os valores testados e o valor da melhor configuração:

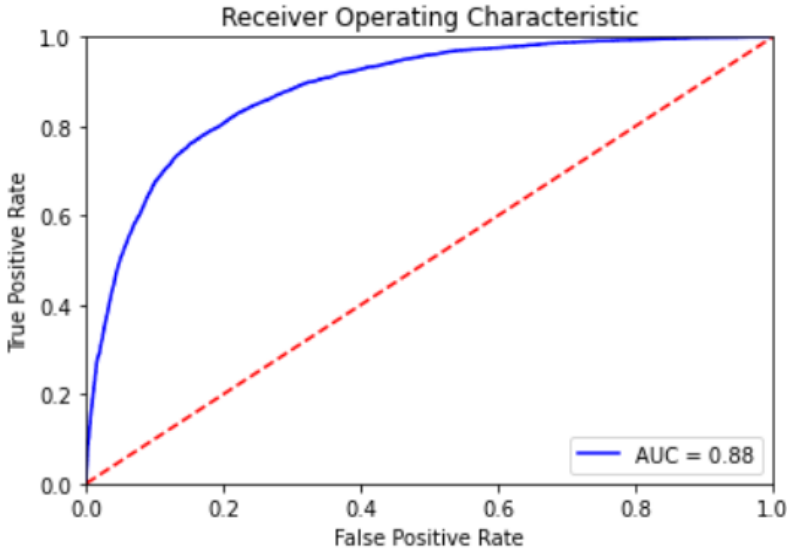
Tabela 21 - Definição de hiperparâmeros para modelo de regressão logística

Parâmetro	**Valores testados**	**Valor final**
C	[0,001 , 0,01 , 0,1 , 1 , 10 , 100, 1000]	0,001
Quantidade de iterações	[1000 , 10000 , 15000, 20000]	1000

Fonte: Elaborado pelo autor

A partir dessa definição, pode-se treinar o modelo com os parâmetros definidos. Desse ponto, é possível a determinação da curva ROC, ilustrada pela figura 42:

Figura 42 - Curva ROC do modelo de regressão logística



Fonte: elaborado pelo autor

Os valores das métricas principais derivadas da matriz de confusão, principalmente as de precisão e sensibilidade que compõem o gráfico, estão apresentadas na tabela 22 para cada valor discreto de limiar de decisão:

Tabela 22 - Métricas de avaliação para cada limiar de decisão do modelo de regressão logística

Limiar de decisão	Classe	Precisão	Especificidade	F1 - Score	P@K
0,1	Reprovado	0,36	0,96	0,53	0,34
	Aprovado	0,98	0.51	0,67	
	Média ponderada	0,84	0,61	0,64	
0,2	Reprovado	0,54	0,81	0,65	0,45
	Aprovado	0,93	0,80	0,86	
	Média ponderada	0,84	0,80	0,81	
0,3	Reprovado	0,65	0,69	0,67	0,52
	Aprovado	0,91	0,89	0,90	
	Média ponderada	0,85	0,85	0,85	
0,4	Reprovado	0,71	0,58	0,64	0,61
	Aprovado	0,88	0,93	0,91	
	Média ponderada	0,84	0,85	0,85	

0,5	Reprovado	0,75	0,49	0,60	0,67
	Aprovado	0,86	0,95	0,91	
	Média ponderada	0,84	0,85	0,84	
0,6	Reprovado	0,78	0,41	0,54	0,73
	Aprovado	0,85	0,97	0,90	
	Média ponderada	0,83	0,84	0,82	
0,7	Reprovado	0,80	0,33	0,47	0,80
	Aprovado	0,83	0,98	0,90	
	Média ponderada	0,83	0,83	0,80	
0,8	Reprovado	0,84	0,26	0,39	0,86
	Aprovado	0,82	0,99	0,89	
	Média ponderada	0,82	0,82	0,78	
0,9	Reprovado	0,87	0,16	0,28	0,95
	Aprovado	0,80	0,99	0,89	
	Média ponderada	0,82	0,80	0,75	

Fonte: Elaborado pelo autor

Observando os números da tabela em geral, as métricas são similares aos valores encontrados no modelo de florestas aleatórias, porém um pouco pior em quase todas as situações. Sendo assim, e sendo o objetivo do modelo muito similar, faz sentido fazer uma decisão a partir dos mesmos princípios. Dessa forma, assumindo grosseiramente o valor de 50% de especificidade como o valor máximo aceitável, opta-se por um limiar de decisão de 0,5. A tabela 23 representa a matriz de confusão correspondente:

Tabela 23 - Matriz de confusão para primeira calibragem do modelo de regressão logística

	Negativo Previsto	*Positivo previsto*
Negativo Real	19.375	969
Positivo real	3.030	2.941

Fonte: Elaborado pelo autor

Assim, metade das reprovações é corretamente classificada e cerca de 25% das previsões está equivocada. Alterando o limiar de decisão para 0,7, obtém-se a seguinte matriz de confusão:

Tabela 24 - Matriz de confusão para segunda calibragem do modelo de regressão logística

	Negativo Previsto	*Positivo previsto*
Negativo Real	19.985	492
Positivo Real	3.999	1.972

Fonte: Elaborado pelo autor

Assim como a segunda matriz de confusão do modelo de florestas aleatórias, o nível de falsos positivos cai muito conforme e fica num nível bastante confiável, fazendo do modelo algo interessante para previsão de piores casos.

4.3 FEATURE IMPORTANCE

No treinamento do modelo existe uma qualificação sobre as *features* referente à sua utilização na classificação dos modelos. A tabela 25 e gráficos da figura 42 indicam os valores de importância aferidos a cada variável independente:

Tabela 25 - *Feature importance* em cada modelo

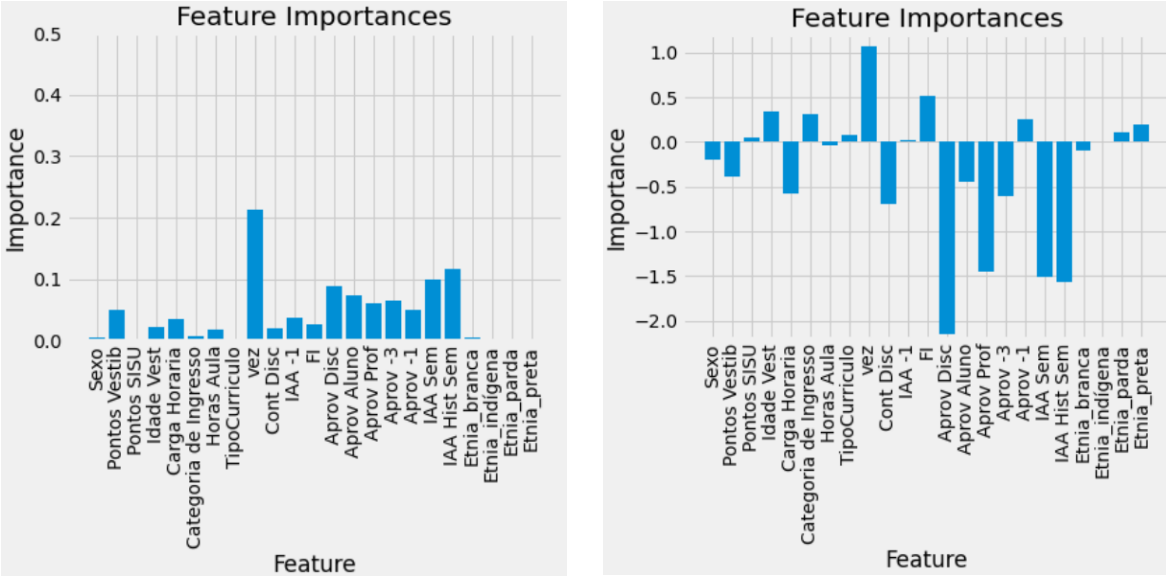
Variável	**Florestas Aleatórias**	**Regressão Logística**
Gênero	0	-0,2
Etnia branca	0,01	-0,1
Etnia indígena	0	0,01
Etnia parda	0	0,25
Etnia preta	0	0,2
Idade no vestibular	0,02	0,34
Categoria de Ingresso	0,01	0,31

Horas Aula	0,02	-0,04
Tipo currículo	0	0,08
IA	0,1	-1,51
IAA	0,12	-1,57
Delta IAA	0,04	0,03
Taxa de aprovação do aluno	0,07	-0,44
Taxa de aprovação do aluno - 3	0,07	-0,6
Taxa de aprovação do aluno - 1	0,05	0,25
Taxa de aprovação do professor	0,06	-1,45
Taxa de aprovação da disciplina	0,09	-0,69
Pontos Vestibular	0,05	-0,38
Pontos SISU	0	0,04
FI	0,03	0,52
Contagem de disciplinas	0,02	-0,69
Repetência de matérias	0,21	1,06

Fonte: Elaborado pelo autor

Os gráficos de barra abaixo demonstram graficamente as *feature importances* de cada variável em ambos os modelos.

Figura 43 - Importância das variáveis nos modelos de florestas aleatórias e regressão logística



Fonte: elaborado pelo autor

Existe algumas diferenciações conceituais importantes para a definição da importância das variáveis em cada modelo. Como cada árvore de decisão do *bundle* possui um set de dados diferente, terá diferentes características dentro de seu grupo. A *feature importance* do modelo é medida pela soma das árvores na qual a respectiva *feature* foi a determinante na divisão do primeiro nó. Essa situação faz com que *features* que podem explicar bastante a diferenciação de algumas instâncias específicas seja ofuscada por alguma com menor poder de explicação sobre a variável dependente. Além disso, *features* que não correspondem a todos os perfis da população, como pontuação do SISU e etnias específicas não sejam relevantes a princípio. Por essas questões é interessante fazer um paralelo com os resultados apresentados pelo modelo de regressão logística, que não possui essas especificidades. A *feature importance* do mesmo é determinada pelos coeficientes de cada uma das variáveis. Similarmente à regressão linear, cada variável possui um coeficiente que determina a relação da variável independente e a variável independente. Essa relação é estabelecida como o log da probabilidade de uma classificação da variável dependente, dada a variável independente. No fim das contas, ele representa o quanto a intensidade daquela variável independente afeta a probabilidade de uma classificação de uma instância enquanto positiva ou negativa, dado que o problema do presente trabalho é de classificação binária.

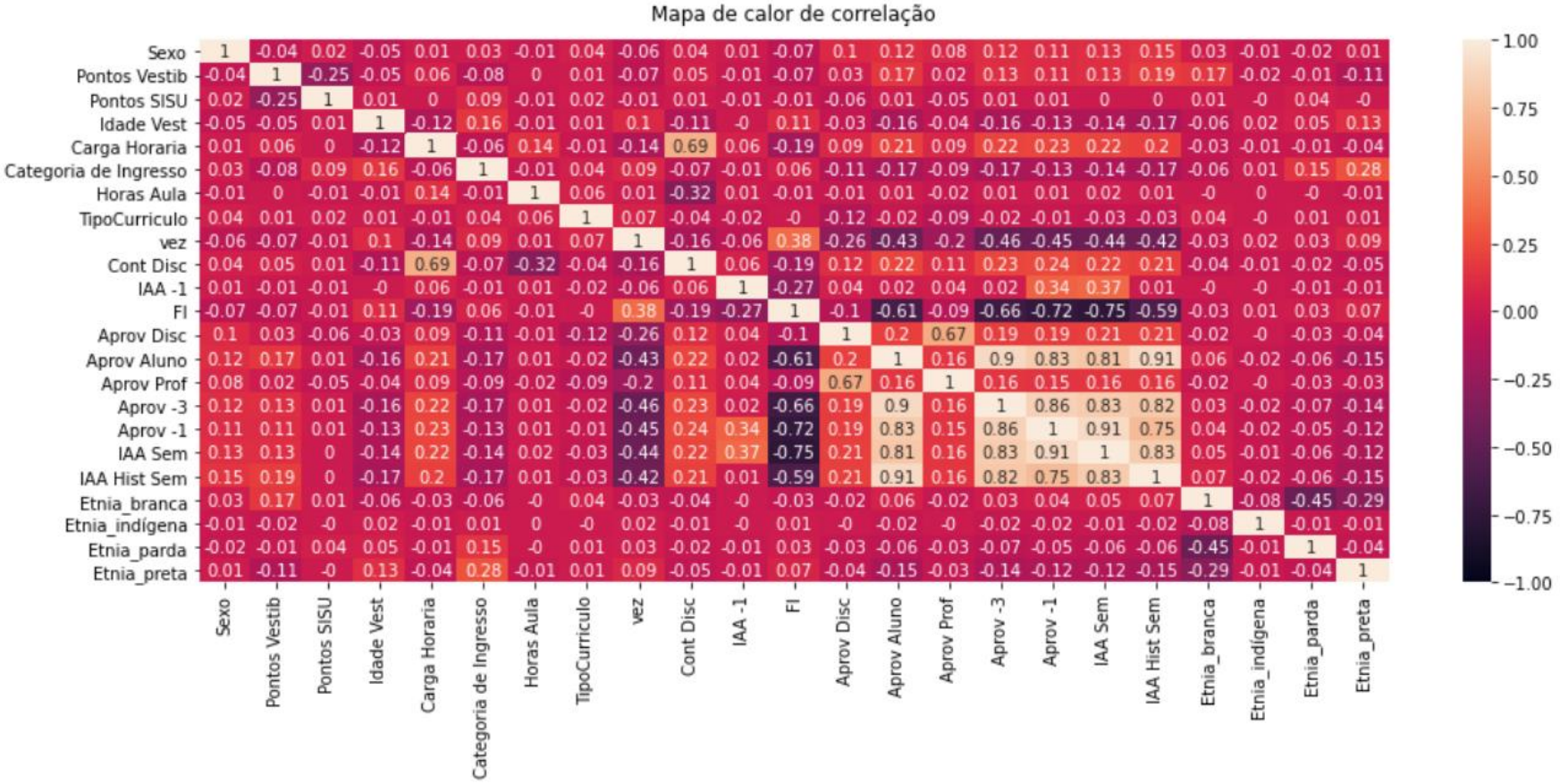
Das explicações acima, é relevante que as *feature importances* do modelo de florestas aleatórias são altamente dependentes dos valores na amostra *bootstrap* associada a cada árvore, assimetrias da distribuição dos perfis dos alunos e de outras variáveis.

Exemplificando este último ponto, se uma variável possui uma correlação muito grande a outra, mas uma capacidade de explicação do resultado um pouco menor, existe uma grande probabilidade de a variável menos importante não constar com grande relevância dentre as *features*. Quanto ao modelo de regressão logística é diferente no sentido de que cada variável independente é avaliada de maneira separada, fazendo com que todas possam ser avaliadas por si mesma, mas potencialmente ofuscando relações de uma variável que possam explicar a diferenciação da outra.

A fim de agregar à discussão, a figura 43 apresenta um mapa de calor para representar a correlação entre as variáveis.

É importante ressaltar que uma alta correlação não implica causalidade ou efeito de uma variável em outra, mas indica uma proximidade de ambas uma com a outra ou de ambas com alguma variável específica.

Figura 44 - Mapa de correlação entre as variáveis independentes



Fonte: elaborado pelo autor

5 DISCUSSÃO

Esse tópico é destinado a um olhar mais aprofundado sobre os resultados observados nos modelos quanto a como se pode utilizá-los para uma revisão de alguns pressupostos e análises anteriores aos modelos, assim como a proposição de formas de inserção deles na gestão do ensino do centro.

5.1 VARIÁVEIS E DIAGNÓSTICO

O valor absoluto do coeficiente das variáveis do modelo de regressão logística representa o quanto o valor daquela variável afeta a probabilidade de classificação de uma classe ou outra. Consequentemente, vemos que as *features* de maior impacto são as de maior valor absoluto, nomeadamente aprovação de disciplina, professor, IA, IAA e a repetição de matéria. A questão de referencial do efeito de uma *feature* na predição é descrita pelo sinal. Ou seja, para a repetição de matéria, valores muito baixos correspondem a um valor alto na predição, considerando que o valor de 1 significa reprovação. Essa era uma característica esperada, pois ficou indicado esse efeito durante a análise descritiva das variáveis. O valor das variáveis de taxa de aprovação de disciplina, professor, IA e IAA tem o efeito inverso: valores muito altos significam baixa probabilidade de uma instância ser classificada como reprovado. Pelo mapa de calor, verifica-se que existe uma alta correlação entre índices de aproveitamento e taxas de aprovação de alunos, mas é interessante apontar que no modelo de regressão os índices de aproveitamento parecem ser melhores preditores do que as taxas de aprovação, ou seja, a diferenciação dentro da linha dos índices de aproveitamento referente à aprovação das matérias aparenta ser maior do que a diferenciação entre médias de aprovação de alunos. Conforme foi visto na análise descritiva, existe uma discricionariedade mais elevada nos valores de índices de aproveitamento do que nas taxas de aprovação, motivo que pode estar alinhado com a constatação proveniente da regressão.

Os valores de taxas de aprovação de semestres anteriores à matéria demonstram ser melhores indicadores do que a taxa de aprovação histórica do aluno. Isso indica que existe uma tendência de curto prazo do aluno que é interessante para a definição de seu desempenho. Da mesma maneira, o delta IA parece apontar para o mesmo sentido, porém de maneira muito menos expressiva. Ao que parece IA de cada semestre parecem possuir um efeito muito próximo ao de IAA, então parece que o efeito de tendência do IA é muito menos

determinante. Se por um lado o IAA tem uma inércia maior do que a taxa de aprovação, por conta de não ser binário, mas permitir intensidades, não é dada a diferença entre os efeitos aprovação e nota. Por fim a correlação entre o delta IAA é relativamente baixo com as taxas de aprovação de semestres anteriores, especificamente para a aprovação dos últimos 3 semestres, que é aproximadamente nula.

A taxa de é, na maioria dos casos, próxima de 0, sendo uma boa métrica para a definição de aprovação de alunos, especificamente identificando alunos com mal desempenho. Essa relação é fortalecida pela alta correlação negativa entre a taxa de FI para com as taxas de aprovação, pois enquanto o FI é esperado ter alto impacto nas notas e índices de aproveitamento por conta do impacto de aumento da discricionariedade da nota, por outro essa relação não é tão óbvia para com as taxas de aprovação.

As taxas de aprovação de professores e disciplinas aparentam ter mais efeito sobre a reprovação de alunos do que as próprias taxas de aproveitamento dos próprios alunos, algo que pode ser explicado pela assimetria das taxas de aprovação de matérias e de professores indicadas na análise descritiva. Indica-se que em geral uma disciplina difícil é mais discricionária do que um professor, porém existe uma alta correlação entre as duas variáveis, assim como existem disciplinas que são lecionadas por apenas um professor, então resta analisar o quanto a dispersão dentro de uma própria disciplina, derivada de diferentes professores, é relevante.

Como era de se esperar, existe uma correlação alta entre a carga horária e a quantidade de disciplinas de um semestre. Ambas as quais parecem ter uma relação inversamente proporcional com a reprovação de disciplinas, indicando que altos valores de carga horária estão associados a altos baixos níveis de reprovação. Enquanto essa relação é razoavelmente evidente, ressalta-se a contrassenso que cargas horárias de disciplinas específicas são proporcionais a níveis altos de reprovação.

Referente ao perfil de alunos, observa-se que eles não são os maiores indicadores para a definição de se alunos passarão ou não em matérias específicas, porém parecem identificar alguns grupos particulares. O fato de existir uma diferenciação a matérias específicas por conta do perfil da pessoa, conforme a importância das variáveis indica, já é por si só algo relevante. Enquanto etnias preta e parda possuem uma relação com o mal desempenho, a branca apresenta uma relação contrária. Esse efeito é, porém, distribuído pelas diferentes variáveis binárias de identificação das etnias. A idade é curiosamente uma das variáveis de perfil mais indicativas de sucesso e fracasso, diretamente proporcional,

apontando para um pior desempenho de idades mais avançadas, igualmente ilustrado na análise descritiva.

A intensidade de valores no SISU parece não ser muito relevante, mas esse valor é provavelmente resultado da opção de preencher os valores faltantes, de classificados pelo vestibular, com o número 0. Para além disso, uma explicação possível pode ser uma baixa amplitude de valores entre os classificados. O valor de pontos no vestibular parece ser mais explicativo das aprovações nas matérias e esse efeito parece ter uma relação inversamente proporcional, enquanto altos valores estão associados a baixas reprovações. Dentro desse ponto existem efeitos como valores muito altos podem estar associados a nenhuma reprovação, como evidenciado pela quantidade de alunos formados em níveis muito altos de aprovação, assim como pode evidenciar que alunos com valor 0, devido a terem sido classificados por SISU, podem ter níveis de reprovação maior.

Dos apontamentos das variáveis de maior importância do modelo de regressão, o comportamento expresso no modelo de florestas aleatórias é muito similar, sendo a repetição de matérias o maior indicador para determinação de reprovações. Se por um lado a análise descritiva parece demonstrar que a reprovação aumenta a cada vez que a matéria é cursada, esses resultados do modelo parecem indicar que grande parte das reprovações estão concentradas alunos que estão cursando outra vez.

As variáveis de perfil têm uma expressividade menor, assim como a categoria de ingresso, do que no modelo de regressão, ao passo que a *feature* de pontos no vestibular ganha maior proeminência. Além do fato já explicado da etnia e pontos no SISU serem prejudicados enquanto importância de explicação por serem referentes a um grupo diminuto de alunos, uma hipótese é de que os pontos no vestibular podem absorver grande parte da diferenciação de perfis de alunos. Isso é principalmente relevante para a questão do sistema de cotas, onde todos os alunos classificados por cotas possuem uma nota abaixo da nota de corte.

Apesar da alta correlação entre variáveis de aprovação e índices de aproveitamento, existe uma diversidade de árvores geradas com cada uma delas, indicando que não existe uma delas melhor para explicar a reprovação dos alunos, sendo cada uma utilizada para um contexto diferente. Uma possibilidade para isso seja que elas englobem efeitos diversos como dos perfis de alunos, tendências e padrões de comportamento.

5.2 IMPLEMENTAÇÃO DO MODELO

No capítulo 4 evidenciou-se a tomada de decisão nos ajustes considerando pressupostos de utilização. Ao mesmo tempo, com poucos ajustes e redefinições existe uma diversidade de usos possíveis. Dois objetivos principais nortearam a elaboração e ajuste dos modelos: a identificação de casos críticos; o mapeamento de todas as reprovações. Em ambos os casos o modelo de florestas aleatórias teve um desempenho melhor, maior precisão e menor especificidade.

No primeiro objetivo, o modelo de florestas aleatórias indicou 2.570 reprovações, das quais 2.215, foram de fato reprovações, ou seja, precisão de 86%, indicando com um elevado nível de confiabilidade quais os casos mais críticos. Isso significa que ações voltadas para esses alunos devem ser recursos bem direcionados e um alto potencial de resultado, se bem desenhadas políticas de auxílio, suporte ou demais cursos de ação. Conforme existem limitações quanto a recursos, para o caso de uma ação que possa ser aplicada a apenas 1000 alunos, o modelo pode ser refinado para apresentar um grupo mais enxuto de instâncias, e consequentemente uma precisão ainda maior.

No segundo objetivo, o modelo foi calibrado para um apontamento de um valor maior dos casos de reprovação, ou seja, da identificação de um número maior de casos de reprovação, limitado pela especificidade, ou indiretamente pela proporção de falsos positivos. Como resultado, obteve-se uma sensibilidade de 79% dos casos de reprovação a uma taxa de 36% de falsos positivos em relação a todas as previsões de reprovação. Na prática, significa que uma parcela relevante das reprovações está no grupo previsto para reprovar, mas existe um erro relativamente alto para a afirmação de que um caso específico será uma reprovação. Para lidar com essa variabilidade, o modelo pode ser executado por níveis onde o resultado será clusterizado por níveis de risco, níveis derivados da probabilidade de reprovação apontada pelo modelo.

Apesar do desempenho menos confiável, o modelo de regressão logística tem uma série de vantagens, dentre elas a maior facilidade de implementação, a simplicidade, uma maior relação direta das variáveis com os efeitos e eficiência computacional. Por conta desses motivos, pode haver uma decisão de gestão de aplicar esse modelo ao invés do modelo de florestas aleatórias. Se o objetivo for a identificação de casos mais críticos, o modelo apresentou resultados similares. Foram previstos 2.588 casos de reprovação, dos quais 86% estavam corretos, ou seja 86% de precisão. Quanto ao segundo objetivo de aplicação, houve

uma sensibilidade de 49% dos casos, ao custo de uma taxa de 25% de falsos positivos. Por esses números, conclui-se que o modelo é muito menos estável para a aplicação segura na prática.

Do acima apresentado, conclui-se por uma sugerir a aplicação do modelo de florestas aleatórias, pois a sua flexibilidade e estabilidade do *bootstrap* que demonstra assertividade para a manutenção da estabilidade do modelo fazem com que a sua utilização seja mais benéfica. Apesar disso, condições que não estão apontadas no presente trabalho como limitações computacionais podem ser solucionadas pelo uso da regressão logística, que se direcionada ao primeiro objetivo 1, pode levar a resultados promissores. A aplicação de qualquer um dos modelos visando o segundo objetivo pode levar a erros dependendo da política em análise, portanto sugere-se aplicações de identificação de grupos de risco, mais do que casos específicos.

5.3 IMPLICAÇÕES PARA A GESTÃO

Com o desenvolvimento do projeto e as análises contidas nele, pode-se levantar pontos que sejam relevantes no sentido de tomada de decisão e ajuste da gestão do ensino. Por fins didáticos, essa questão será abordada no presente documento por meio de uma segmentação das *features* do modelo.

1) Variáveis de perfil de aluno: são variáveis exógenas ao sistema, portanto não passíveis de controle pela administração, porém se elas são relevantes para a identificação do desempenho do aluno, é interessante acompanhá-las como forma de compreender quem é o discente envolvido, quais os fatores que o levam a esse tipo de resposta dentro do sistema de ensino e consequentemente traçar políticas que sejam efetivas para a melhoria de seu desempenho.

2) Variáveis de cadastro de disciplinas ofertadas e matrícula de alunos: em nível macro, esse tipo de informação pode ser um *feedback* relevante para a construção curricular, seja em divisão, aglutinação de disciplinas, ementas, dentre outros efeitos. Em nível micro, informações provenientes deste tipo de dado podem justificar dois vieses importantes que são alterações da forma de oferta de disciplinas específicas, assim como a orientação discente para que eles possam tomar as decisões mais adequadas

para melhorar seu desempenho e aprendizado, seja com matrícula ou gestão da sua dedicação e rotina.

3) Variáveis de desempenho dos alunos: é de suma importância o acompanhamento do desempenho dos alunos para a identificação de problemas individuais ou do sistema de ensino vigente. Assim como o exemplificado na análise descritiva da seção 3, é necessário utilizar desse tipo de indicador para o acompanhamento das condições atuais, assim como usá-los de parâmetros para avaliar a implementação de novas políticas. Posteriormente o ideal seria que fossem criados dispositivos que disponibilizassem essa informação e análises subsequentes que permitam com que o aluno entenda seu quadro individual, qual sua situação e como ele mesmo pode se organizar para gerir seu próprio aprendizado.

4) Variáveis de desempenho de professores e disciplinas: assim como no tópico de desempenho de alunos, o acompanhamento dos indicadores de desempenho discente referentes às disciplinas e professores são bastante relevantes para o diagnóstico e gestão. No tocante a disciplinas, questões como currículo, organização de conceitos, orientação pedagógica são algumas das questões que deveriam ser avaliadas e reformuladas para uma melhoria do desempenho discente e aprendizado. A avaliação docente, por sua vez, é muito importante para avaliar a capacitação e desempenho enquanto professor. Hoje os professores são geralmente avaliados por outras funções como em pesquisa, mas o gerenciamento de seu desempenho na sala de aula não é avaliado. A maior dispersão de desempenho de professores em relação a disciplinas indica que isso pode ser um fator relevante no desempenho discente.

É importante ressaltar que a perspectiva adotada nesse trabalho foi de mensurar o desempenho discente enquanto resultado de aprovação e nota. Esses são muito relevantes para o trabalho da eficiência de recursos e certos objetivos da gestão do ensino, mas no que se refere a aprendizado, objetivo primário do ensino, eles são indicadores relevantes, mas que não captam o assunto em toda sua essência. Existem distinções importantes a serem feitas que fogem do escopo desse trabalho. Dessa maneira, sugere-se que sejam executadas novas investigações e abordagens para uma melhor gestão desse aspecto específico.

6 CONCLUSÕES

O uso do capital investido na educação superior no decorrer das últimas décadas tem sido elevado, porém essa questão não se refletiu nos níveis de produtividade. Após um olhar no microcosmo do centro de engenharia da UFSC, constata-se que existem muitos problemas quanto à eficiência e desempenho dos alunos. Se por um lado os níveis de reprovação e evasão já eram elevados no início do século, esses valores têm se tornado mais críticos nos anos recentes.

Com o quadro de desempenho discente se agravando, o acompanhamento dos resultados e o ferramental de diagnóstico, proposição e controle do ensino ainda se encontra muito deficiente. Dessa maneira, existe uma necessidade de mapeamento e compreensão mais aprofundada das características relevantes para o desenho de melhores políticas. É nesse contexto que o presente trabalho se propôs a fundamentar e criar uma ferramenta de análise e colocá-la em prática para a investigação de em torno dos resultados deficitários.

Com esse objetivo geral em meio a uma tendência global do direcionamento de políticas e tomada de decisão baseado em dados, delineou-se um caminho de fazer-se uso dos ativos nas bases de dados da instituição como forma de fundamentar a gestão. Dessa maneira, foram cumpridos os objetivos de:

1) Limpar a base de dados e ajustar variáveis existentes para que estejam adaptadas aos modelos
2) Criar variáveis que aumentem a capacidade de previsão do modelo
3) Aplicar modelos de previsão para prever os resultados de aprovação dos alunos do CTC
4) Avaliar, parametrizar e selecionar os modelos de acordo com as métricas adequadas
5) Discutir as características mais relevantes do modelo em torno do que significa em termos de leitura do desempenho acadêmico.
6) Propor uma utilização do modelo desenvolvido para diagnosticar e direcionar esforços voltados a um aumento de eficiência do centro

Os primeiros 4 objetivos são etapas necessárias para a construção e confiabilidade do modelo a ser utilizado no quinto objetivo. Todas essas etapas foram bem sucedidas, principalmente ao avaliar que o modelo apresentou métricas muito positivas de uso e estabilidade. Conforme o modelo é em si uma modelagem para a compreensão da realidade do ensino no centro, espera-se que, se alcançado um modelo com boa aplicabilidade, seus

processos internos e variáveis auxiliam a interpretar o que de fato se passa no processo de graduação. É com esse pressuposto que o diagnóstico realizado no projeto levantou uma série de possibilidades e hipóteses que podem e devem ser exploradas em próximos trabalhos e projetos. Espera-se que o presente projeto sirva como impulso para uma modificação na forma de acompanhamento dos resultados do processo de ensino e para uma atualização constante das características de alunos e seu desempenho, já que com o controle do processo e seus resultados poderá alterar o grave quadro no qual o ensino brasileiro se encontra.

REFERÊNCIAS

ABUT, Fatih et al. PREDICTING STUDENT'S PASS/FAIL STATUS IN AN ACADEMIC COURSE USING DEEP LEARNING: A CASE STUDY, 2017.

ALMEIDA, E; Godoy, E. A evasão nos cursos de engenharia: uma análise a partir do Cobenge. In: CONGRESSO BRASILEIRO DE EDUCAÇÃO EM ENGENHARIA, XLIV, 2016, Natal. **Anais**... São Paulo, SP: Universidade Cruzeiro do Sul; 2016. P. 1-3.

BACCARO, Thais Accioly; SHINYASHIKI, Gilberto Tadeu. Relação entre desempenho no vestibular e rendimento acadêmico no ensino superior. **Revista Brasileira de Orientação Profissional**, v. 15, n. 2, p. 165-176, 2014.

Belachew, E. B., & Gobena, F. A. (2017). Student Performance Prediction Model using Machine Learning Approach: The case of Wolkite University. *International Journal if Advanced Research in Computer Science and Software Engineering*, *7*(2), 46-50.

Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study. *Sustainability*, *11*(10), 2833-2851.

BREIMAN, L. Random forests. *Machine learning*, v. 45, n. 1, p. 5-32, 2001.

CASEY, Kevin; AZCONA, David. Utilizing student activity patterns to predict performance. **International Journal of Educational Technology in Higher Education**, v. 14, n. 1, p-4, 2017.

CLAESEN, Marc; DE MOOR, Bart. Hyperparameter search in machine learning. **arXiv preprint arXiv:1502.02127**, 2015.

DE OLIVEIRA, JOSIMAR ALCANTARA. **Prejuízos da Evasão nos Cursos de Engenharia: Um Estudo de Caso**. 2016. Tese de Doutorado. Universidade Paulista.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861‑874, 2006.

FEHER, Beatriz Zardetto. A relação entre os gastos públicos em educação e a qualidade de ensino. 2018.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001.

GÖNEN, M. Analyzing receiver operating characteristic curves with SAS. *SAS Institute*, 2007.

KAR, Purushottam; NARASIMHAN, Harikrishna; JAIN, Prateek. Surrogate functions for maximizing precision at the top. In: **International Conference on Machine Learning**. 2015. p. 189-198.

OFORI, Francis; MAINA, Elizaphan; GITONGA, Rhoda. Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review. **Journal of Information and Technology**, v. 4, n. 1, p. 36, 2020.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). **Education at a glance 2010: OECD indicators**. Paris: OECD, 2019.

PADILHA, W. Educação popular e atenção à saúde. In: SEMINÁRIO SOBRE A POLÍTICA NACIONAL DE PROMOÇÃO DA SAÚDE, 1., 2006, Brasília. **Anais**... Brasília, DF: Ministério da Saúde, 2009. p. 58-60.

PIERI, Renan. Retratos da educação no Brasil. **São Paulo**, p. 1-38, 2018.

SARKAR, D. Understanding Feature Engineering. *Towards Data Science*, 04 janeiro 2018. Disponível em: <https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b>. Acesso em: 10 Novembro 2020.