Bianca Ribeiro Lima Marques

Evasão no ensino superior da Paraíba em 2019: uma análise com modelos estatísticos

João Pessoa, Brasil

2021

Bianca Ribeiro Lima Marques

Evasão no ensino superior da Paraíba em 2019: uma análise com modelos estatísticos

Monografia apresentada ao curso de Bacharelado em Estatística da Universidade Federal da Paraíba, como requisito fundamental para obtenção do grau de Bacharel em Estatística.

Universidade Federal da Paraíba – UFPB

Centro de Ciências Exatas da Natureza

Departamento de Estatística

Profa. Dra. Ana Hermínia Andrade e Silva

João Pessoa, Brasil

2021

Catalogação na publicação
Seção de Catalogação e Classificação

M357e Marques, Bianca Ribeiro Lima.
Evasão no ensino superior da Paraíba em 2019 : uma análise com modelos estatísticos / Bianca Ribeiro Lima Marques. - João Pessoa, 2021.
75 p. : il.

Orientação: Ana Hermínia Andrade e Silva.
TCC (Graduação/Bacharelado em Estatística) - UFPB/CCEN.

1. Modelagem estatística. 2. Aprendizagem de máquina. 3. Evasão no ensino superior. I. Silva, Ana Hermínia Andrade e. II. Título.

UFPB/CCEN CDU 311(043.2)

Elaborado por Josélia Maria Oliveira da Silva - CRB-15/113

Trabalho dedicado à minha mãe, a maior incentivadora dos meus estudos.

AGRADECIMENTOS

O meu primeiro e maior agradecimento é dado ao Senhor, que conduz minha vida em todos os aspectos, inclusive no âmbito acadêmico. Que toda honra, glória, louvor sejam dados a Cristo Jesus, a quem por misericórdia tornou-se meu Salvador, por meio do seu precioso sangue.

Grata também ao meu querido parceiro em que posso chamar, hoje, de esposo. Obrigada, meu bem, por todos os nossos momentos, por sempre me apoiar e me incentivar a continuar, me lembrando sempre que tudo isso deve ser feito para a glória de Cristo. Te amo, enquanto Deus permitir que a gente respire!

À minha mãe e irmã que representam boa parte de quem sou hoje. Também minhas companheiras nas lutas e felicidades. Conquistamos isso juntas!

Ao meu pai, que da sua maneira, demonstra o seu amor.

Aos meus sogros e cunhadas, que tenho a honra de também chamá-los família.

Aos meus amigos da graduação. Me alegro por termos formado amizades verdadeiras.

Aos meus professores, que foram participantes especiais no amor que surgiu pela estatística quando nem mesmo eu sabia o que esperar do curso. Em especial, grata pela minha professora, Ana Hermínia, que me orientou nesse trabalho com tanta paciência e incentivo.

À Cru, movimento acolhedor de cristãos apaixonados por conectar pessoas a Cristo. A experiência com a universidade, sem dúvidas, não teria sido a mesma sem esse movimento e os amigos formados.

"Mas em nada tenho a minha vida por preciosa, contanto que cumpra com alegria a minha carreira e o ministério que recebi do Senhor Jesus, para dar testemunho do evangelho da graça de Deus." (Bíblia Sagrada, Atos 20:24)

RESUMO

A educação sofreu forte expansão no país nas duas últimas décadas, inclusive no estado da Paraíba. Tendo em vista que o ensino superior é de grande importância para o desenvolvimento da população, estudar as causas que levam os discentes a evadirem do ensino superior é indispensável para o contexto educacional. Por isso, este trabalho visa levantar dados relativos a essa temática para o ano de 2019, bem como modelar os dados de evasão para identificar, de modo estatístico, o que pode influenciar a evasão de tais discentes. Os dados são provenientes do censo anual do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) e, como forma a alcançar os objetivos propostos, foram realizadas análises estatísticas e aplicados métodos de modelagem utilizando o *software* R e a linguagem *Python*. Os métodos estatísticos envolveram os Modelos Lineares Generalizados (MLGs), especificamente a Regressão Logística, o modelo KNN, a Árvore de decisão e o modelo *Random Forest*. Após a metodologia ter sido aplicada com o ajuste de cada modelo, foi identificado qual o modelo com maior poder de predição, utilizando técnicas de avaliação de modelos, como a acurácia, curva ROC e a medida AUC. Assim, foi possível constatar que o modelo da Árvore de Decisão apresentou maior vantagem em prever a evasão de discentes do ensino superior. Mesmo diante disso, outros modelos tiveram suas importâncias devidas, trazendo conclusões importantes ao trabalho de modo a contribuir com o âmbito educacional da Paraíba.

Palavras-chave: modelagem estatística, aprendizagem de máquina, evasão, ensino superior.

ABSTRACT

Education has undergone a strong expansion in the country in the last two decades, also reaching the state of Paraíba. Considering that higher education has great importance for the development of the population, studying the causes that lead students to evade higher education is indispensable for the educational context. Therefore, this work aims to raise data related to this topic for the year 2019, as well as model the data of evasion to identify, statistically, what can influence the evasion of such students. The data come from the annual census of the National Institute of Studies and Educational Research Anísio Teixeira (Inep) and all the objectives are put into practice through the application of modeling methods added to the use of the R software and the Python language. Statistical methods involved Generalized Linear Models (MLGs), specifically the Logistic Regression, the KNN model, the Decision Tree and the Random Forest model. After the methodology has been applied with the fitting of each model, it was identified which model had the greatest predictive power, through model evaluation techniques, such as accuracy, ROC curve and AUC measure. Thus, it was possible to verify that the Decision Tree model presented a greater advantage in predicting the evasion of students from higher education. Even with this result, other models had their due importance, bringing important conclusions to the project in order to contribute to the scope of education of Paraíba.

Keywords: statistical model, machine learning, evasion, higher education.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Inep	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação
IES	Instituições de Educação Superior
AM	Aprendizagem de Máquina
MLGs	Modelos Lineares Generalizados
KNN	k vizinhos mais próximos
AUC	Área abaixo da curva
PNE	Plano Nacional de Educação
PEE	Plano Estadual de Educação
IDE	Ambiente de Desenvolvimento Integrado
RFE	Eliminação de Recurso Recursivo
i.i.d.	Independentes e identicamente distribuídos
OR	Odds Ratio
FAVAP	Faculdades Vale do Piancó
NA	Não aplicável

SUMÁRIO

1 INTRODUÇÃO

Motivação: É notório o papel que o ensino superior tem para o estímulo do desenvolvimento de competências cognitivas, funcionais e sociais, por isso este é um tema considerado prioritário e estratégico para o futuro das nações (LOPES; MESQUITA, 2017). Além disso, essa mesma discussão é de extrema importância ao observar o processo de expansão no Brasil e, particularmente, no estado da Paraíba, nas duas últimas décadas.

Em 9 de janeiro de 2001, o presidente da República aprovou a Lei n.º 10.172/ 2001, referente ao Plano Nacional de Educação (PNE). Este plano é referente às políticas educacionais e funciona como um norteador do planejamento educacional a longo prazo. O atual PNE, editado por meio da Lei Federal 13.005, de 2014, e na Paraíba, o PEE (Plano Estadual de Educação) editado por meio da Lei n.º 10.488, de 2015, ambos com vigência de 10 anos, possuem, entre todas as metas contidas no plano, seis e três metas, respectivamente, destinadas à educação superior. Observe tais metas na Tabela 1.

Dada a devida importância, é estimulado cada vez mais ao jovem a entrada no ensino superior. No entanto, esse cenário ainda apresenta muitos contrapontos. O processo de permanência nas universidades muitas vezes entra em confronto com a necessidade dos alunos de conciliar trabalho e estudo, a adaptação a um novo sistema de ensino, aos conhecimentos anteriores de maior complexidade e aos aprendizados nem sempre vivenciados pelos alunos de camadas mais populares. Além disso, um outro fator que deve ser considerado para permanência do aluno é a situação econômica, que para muitos é desfavorável, implicando em dificuldades financeiras para se manter, comprar materiais e deslocar-se por meio de transporte. (DIAS; COSTA, 2016).

Desafios: Poucos trabalhos foram encontrados acerca da evasão no ensino superior. Outro desafio tratou-se em referenciar pesquisas que utilizam do mesmo conceito de evasão abordado nesse trabalho, devido às inúmeras divergências nessa temática. Por fim, como o censo de discentes do Inep carrega uma abundância de dados, foi um desafio trabalhar com uma grande abrangência, escolhendo, assim, abordar a Paraíba, além de fazer divisões do estado em gerências regionais.

Objetivo e metodologia: Nesse contexto, foram traçados como objetivos para o estudo em questão: levantar dados relativos aos discentes do ensino superior no ano de 2019, delimitando as gerências regionais do estado da Paraíba; e, modelar como atividade preditiva os dados de evasão com o fim de compreender quais variáveis influenciam na evasão de um aluno do ensino superior. Para cumprir tal objetivo, neste trabalho é proposta a metodologia da modelagem estatística, mais especificamente, a Regressão Logística, modelo KNN, Árvore de Decisão e *Random Forest*.

Tabela 1 – Metas do Plano de Educação em vigência

PNE	PEE	Metas
—	Meta 11	Ampliar a oferta, garantir a permanência e melhorar a qualidade da educação do campo.
—	Meta 12	Ampliar a oferta de cursos de educação a distância nas diversas etapas e modalidades de ensino no Estado da Paraíba, triplicando até o final de vigência deste PEE.
—	Meta 15	Ampliar a oferta, garantir a permanência e melhorar a qualidade da educação escolar indígena.
Meta 12	Meta 20	Elevar a taxa bruta de matrícula na educação superior para 50% e a taxa líquida para 33% da população de 18 a 24 anos, assegurada a qualidade da oferta e expansão para, pelo menos, 40% das novas matrículas, no segmento público.
Meta 13	Meta 21	Elevar a qualidade da educação superior e ampliar a proporção de mestres doutores do corpo docente em efetivo exercício no conjunto do sistema de educação superior para 75%, sendo, do total, no mínimo, 35% doutores.
Meta 14	Meta 22	Elevar gradualmente o número de matrículas na pós-graduação stricto sensu, de modo a atingir a titulação anual de 60.000 mestres e 25.000 doutores.
Meta 15	Meta 23	Garantir, em regime de colaboração entre a União, os Estados, o Distrito Federal e os Municípios, no prazo de 1 ano de vigência deste PNE, política nacional de formação dos profissionais da educação de que tratam os incisos I, II e III do caput do art. 61 da Lei nº 9.394, de 20 de dezembro de 1996, assegurado que todos os professorese as professoras da educação básica possuam formação específica de nível superior, obtida em curso de licenciatura na área de conhecimento em que atuam.
Meta 16	Meta 24	formar, em nível de pós-graduação, 50% dos professores da educação básica, até o último ano de vigência deste PNE, e garantir a todos (as) os (as) profissionais da educação básica formação continuada em sua área de atuação, considerando as necessidades, demandas e contextualizações dos sistemas de ensino.
Meta 18	Meta 26	Assegurar, no prazo de 2 anos, a existência de planos de Carreira para os (as) profissionais da educação básica e superior pública de todos os sistemas de ensino e, para o plano de Carreira dos (as) profissionais da educação básica pública, tomar como referência o piso salarial nacional profissional, definido em lei federal, nos termos do inciso VIII do art. 206 da Constituição Federal.

Organização: O restante deste texto está organizado da seguinte forma. No Capítulo 2 é apresentada toda a metodologia do trabalho, a primeira seção, a descrição dos materiais utilizados, e a segunda seção, a descrição dos métodos aplicados. Após isso, no Capítulo 3 estão expostos todos os resultados da metologia aplicada aos dados da pesquisa, bem como a discussão acerca do que foi levantado sobre a evasão. Sendo assim, o Capítulo 4 trata-se da conclusão do trabalho, trazendo as principais informações e resultados observados ao longo do estudo somado às sugestões de temas para futuros trabalhos.

2 METODOLOGIA

Neste Capítulo, é descrita a fonte dos microdados utilizados, assim como os materiais utilizados para aplicação dos métodos. Somado a isso, foram expostos os assuntos acerca da modelagem dos dados com o intuito de obter um maior entendimento nos resultados.

2.1 Materiais

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) vinculado ao Ministério da Educação (MEC) tem como missão contribuir no desenvolvimento econômico e social do Brasil. Essa missão é realizada por meio do auxílio na elaboração de políticas educacionais e para isso atua em diversas áreas, como nas estatísticas educacionais originadas dos censos.

Os dados do trabalho em questão tratam-se do censo do ensino superior do Inep. Primeiramente, é de extrema importância entender a origem e o conceito da palavra "censo" para maior clareza do estudo. Censo origina-se do latim *census*, e quer dizer "conjunto dos dados característicos dos habitantes de uma localidade ou país, para fins estatísticos" (HOUAISS, 2014). Eles trazem sempre informações estatísticas confiáveis que guiam as tomadas de decisões para uma certa melhoria.

No censo da educação superior é utilizado o sistema e-Mec, sistema em que estão registradas as Instituições de Educação Superior (IES), seus cursos e locais de oferta, que juntos a elas, somam-se os dados sobre docentes e discentes. Desse modo, para cumprir os objetivos, o estudo foi realizado por meio do censo da educação superior do ano de 2019, último censo disponível na realização de todo o trabalho, destacando-se apenas os discentes da Paraíba. Com o intuito de compreender melhor a abrangência desse censo, a Tabela com as principais variáveis trabalhadas no estudo com as suas respectivas descrições segundo o Inep encontra-se no Anexo A.

Com base nisso, uma parte dos tratamentos dos dados foi realizado por meio do *software* R, "uma linguagem orientada a objetos, criada em 1996 por Ross Ihaka e Robert Gentleman que aliada a um ambiente integrado permite a manipulação de dados, realização de cálculos e geração de gráficos" (SOUZA; PETERNELLI; MELLO, 2014). A linguagem R, utilizada na versão 4.0.2 e o Ambiente de Desenvolvimento Integrado (*Integrated Development Environment* - IDE) *RStudio*, estão disponíveis para *download* em *https://www.r-project.org/* e *https://www.rstudio.com/products/rstudio/*, respectivamente. Além disso, para a modelagem dos dados também se utilizou a linguagem de programação Python que tem a mesma finalidade, mas que pode ser mais recomendada para grandes

bases de dados, como se enquadra a pesquisa em questão. Para esta foi utilizada a versão 3.8.3 por meio do Jupyter Notebook, uma *interface* gráfica disponível para *download* em *https://jupyter.org/*, com o fim de programar em diversas linguagens como Julia, Python e R.

Em R, utilizou-se das bibliotecas data.table, dplyr, caTools, em que realizou-se a filtragem apenas dos dados da Paraíba, a criação das colunas de gerências regionais e evasão e a modelagem inicial, especificamente os modelos univariados de regressão logística. Todas as demais implementações computacionais foram realizadas em Python, utilizando-se das seguintes bibliotecas: numpy, pandas, matplotlib, pingouin, seaborn, imblearn, statsmodels e sklearn.

2.2 Métodos

2.2.1 Pré-processamento de dados

O pré-processamento de dados trata-se da união de técnicas que incluem a preparação, organização e estruturação dos dados. Esse conjunto de atividades tem como finalidade aperfeiçoar a precisão, a qualidade e a eficiência da modelagem efetuada em subsequente (HAN; JIAN; MICHELIN, 2006). Assim, dentre todas as técnicas, foram utilizadas nessa pesquisa, as seguintes:

(*i*) **Categorização de atributos ou criação de novas informações com base em outra** — na seção 3.1 foi destacada a criação da coluna evasão, com base no atributo referente à situação do aluno, a criação da coluna das gerências regionais a partir das informações dos municípios e a transformação das variáveis com mais de duas categorias em *dummys*, técnica de formação de variáveis binárias que "surge como uma forma de permitir a introdução de fatores explicativos de natureza qualitativa em modelos de regressão linear" (VALLE et al., 2002).

(*ii*) **Limpeza dos dados com o tratamento das informações faltantes**, podendo ser excluídas ou substituídas — em variáveis categóricas, criando-se uma nova classe, em variáveis numéricas, substituindo-se pela média, mediana, moda ou com informações previstas conforme as existentes.

(*iii*) **Escolha de quais atributos serão previamente utilizados na pesquisa**, por meio do conhecimento no assunto que está sendo abordado bem como a matriz de correlação — o coeficiente de Pearson determina o grau de relação entre as variáveis variando de -1 (inversamente proporcionais) e 1 (diretamente proporcionais) (JOHNSON; WICHERN et al., 2014).

(*iv*) **Balanceamento dos dados**, em que modelos de classificação como os que são utilizados nessa pesquisa, podem tratar os dados desbalanceados de forma enviesada,

prevendo de uma melhor forma a classe majoritária sem obter o mesmo comportamento para a classe minoritária. Por isso, diversos algoritmos, que balanceiam os dados, podem ser utilizados. Nesse trabalho utilizou-se do algoritmo SMOTE, desenvolvido por Chawla (CHAWLA et al., 2002) com a intenção de diminuir o efeito *overfitting*, ou seja, modelos muito específicos com menor poder de generalização para a classe de interesse. Essa técnica procura gerar dados sintéticos — sem serem replicados os casos já existentes — com base nos vizinhos.

2.2.2 Aprendizagem de Máquina

A aprendizagem de máquina (AM) trata-se de um campo da inteligência artificial e como o próprio nome já sugere trata-se do aprendizado contínuo dos sistemas computacionais. Isso significa que são utilizados algoritmos que identificam e extraem regras e padrões dos dados de modo a aprender e modificar os comportamentos como resposta aos estímulos externos ou como acúmulo de experiência durante as operações (ALPAYDIN, 2020).

De forma específica, a aprendizagem de máquina pode ser considerada supervisionada ou não supervisionada. Na primeira, o modelo extrai informações dos exemplos passados a ponto de fazer conclusões sobre novos exemplos ainda não vistos. Na segunda, não existem exemplos já com as definições. Nesse caso, a máquina busca identificar algum padrão contido nos dados (SCHMITT, 2013).

Com isso, a AM foi utilizada nesse trabalho com o propósito de abordar diversos modelos, bem como utilizar técnicas que identificam qual o modelo mais adequado para os dados em questão. Tais modelos são referentes à aprendizagem de máquina supervisionada de modo que a base é dividida em dados de treinamento e de teste. Aqueles que estão no banco de treinamento servirão de coleta de informações e estudo da máquina, assim como os que estão no banco de teste, que ainda não foram vistos pela máquina, permitirão concluir a eficácia do modelo ao constatar informações sobre eles.

2.2.3 Seleção de variáveis

Além da seleção de variáveis efetivada no pré-processamento, existem outras técnicas de seleção que reduzem as chances de *overfitting* e multicolinearidade que podem ocorrer em um modelo com muitos atributos. Para essa pesquisa, foram cumpridas as técnicas de Eliminação de Recurso Recursivo (do inglês Recursive Feature Elimination — RFE) e o teste Qui-Quadrado. A primeira trata-se da medição do grau de importância de cada atributo para o modelo, analisando os atributos que contribuem de melhor forma na previsão (GUYON et al., 2002). Assim, são retiradas, recursivamente, as variáveis de menor importância. No entanto, essa técnica não trouxe melhoria na previsão dos modelos.

Por isso, utilizou-se a segunda técnica.

O teste Qui-Quadrado (PEARSON, 1900), um teste não-paramétrico utilizado apenas para variáveis categóricas, avalia a hipótese nula de que não há dependência entre as variáveis. Isso implica dizer que o teste mede a relação de dependência que existe entre duas variáveis categóricas de modo a identificar o quanto os valores esperados se desviam dos valores observados na previsão do modelo. Esse teste trouxe melhores resultados para os modelos, sendo assim, foram apresentados nessa pesquisa.

2.2.4 Modelagem estatística

Nesta seção serão abordados alguns dos modelos estatísticos, que envolvem uma variável de interesse Y. Assim, $y_1, ..., y_N$ são os valores das variáveis de interesse Y na população finita, que são considerados observações de $Y_1, ..., Y_N$, variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com distribuição $f(y;\theta)$, onde $\theta \, \epsilon \, \Theta$ é o parâmetro indexador da distribuição f , e θ é o espaço paramétrico.

É de maior costume na modelagem estatística utilizar-se da aleatoriedade de uma amostra a utilizar a população. Isso se dá devido à dificuldade de conseguir dados populacionais. Como afirma Thompson, a amostragem "consiste em selecionar parte de uma população para observar, de modo que seja possível estimar algo sobre toda a população" (THOMPSON, 1992). No entanto, na maioria das aplicações práticas, embora a população seja finita, nunca será observada por inteiro. Por meio do censo do ensino superior trabalhado na pesquisa em questão, encontra-se uma certa quantidade de informações faltantes dos discentes que exemplificam tal fato exposto. Portanto, mesmo que as observações do presente estudo sejam de uma população, o argumento citado justifica a aleatoriedade da população de interesse (PESSOA; SILVA, 1998).

2.2.4.1 Modelos Lineares Generalizados

O modelo de regressão linear é a maneira mais simples de representar uma equação matemática que modela o relacionamento entre as variáveis. Esse modelo é considerado simples, interpretativo e com um custo computacional baixo, sendo muito útil para conjuntos de dados pequenos ou ruidosos (CHEIN, 2019).

Na regressão podemos ter uma ou mais variáveis independentes (X), também conhecidas como variáveis explicativas ou preditoras, que influenciam no valor de Y, a variável resposta. O termo foi iniciado por Francis Galton em um experimento realizado com o diâmetro de sementes e foi replicado para a altura de adultos. Embora o conceito imposto por Galton não seja mais aplicado, o termo permaneceu sendo utilizado até hoje (ALVES, 2016).

Quando uma função linear é aplicada sobre as variáveis independentes, trata-se de

um modelo de regressão linear. Este é o modo mais simples de representar uma equação matemática que modela o relacionamento entre as variáveis:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon_i,$$

em que Y é o vetor de dimensão N x 1, da variável resposta, $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ é o vetor de parâmetros do modelo, sendo β_0 a constante do modelo e os demais coeficientes angulares, X a matriz de variáveis independentes, de dimensões N x p e o subscrito i representa cada uma das observações em análise podendo assumir valores entre 1 e N (população). Além disso, existe a necessidade de o erro aleatório, ϵ_i, ser adicionado ao modelo, tratando-se da diferença entre o valor observado de y e o obtido a partir da reta de regressão.

A forma mais comum de estimar os coeficientes $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_k)^T$ é utilizando método dos mínimos quadrados, que é um procedimento simples e fácil de se aplicar, que visa escolher os valores para os parâmetros de forma a minimizar os erros envolvidos no modelo. Observe:

$$S(\beta) = \sum_{i=1}^{N} \epsilon_i^2,$$

$$S(\beta) = \sum_{i=1}^{N} (Y_i - f(X_i))^2,$$

$$S(\beta) = \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} X_{ij} \beta_j \right). \tag{2.1}$$

Tratando o S como a soma, observa-se o Y_i como o valor original, o $f(X_i)$ igual ao valor resultado da função e o i representando cada instância. Desse modo, a Equação (2.1) concerne da aplicação do método ao modelo de regressão. Mais desse método pode ser encontrado em (HELENE, 2006).

No entanto, é necessário destacar que o modelo de regressão linear exige suposições que quando violadas podem não trazer conclusões fidedignas, como a não eficiência na estimação dos parâmetros por meio do método dos mínimos quadrados. São elas: o relacionamento entre as variáveis independentes com a variável resposta necessita ser linear, não deve haver correlação entre as variáveis independentes, os erros aleatórios devem ser distribuídos normalmente, devendo apresentar variância constante (homoscedasticidade) e não serem auto-correlacionados. Essas suposições, no entanto, nem sempre são atendidas nas aplicações práticas e, diante disso, surge a necessidade de utilizar modelos que se adequem melhor aos dados, como os modelos lineares generalizados.

Muitos dos modelos "apresentam uma estrutura de regressão normal linear e têm em comum, o fato da variável resposta seguir uma distribuição dentro de uma família de distribuições com propriedades muito específicas: a família exponencial" (TURKMAN; SILVA, 2000). Os modelos lineares generalizados são então definidos por modelos que seguem a estrutura da regressão linear de modo a serem a extensão dela e possuem a variável resposta com uma distribuição pertencente à família exponencial de distribuições do seguinte modo:

$$f(y|\theta,\phi) = \left\{ exp\frac{y\theta - b(\theta)}{a(\theta)} + c(y,\phi) \right\},$$

em que θ e ϕ são parâmetros escalares e a(.), b(.) e c(.) trata-se de funções reais conhecidas.

Dessa forma, semelhantemente ao que já foi apresentado, o MLG é composto por três componentes: a componente aleatória (resíduo), a componente sistemática que se trata da função linear entre as variáveis explicativas ($X\beta$) e por fim, a função de ligação, função que lineariza a relação entre a componente sistemática e o valor esperado da componente aleatória (Xβ e E(x)). Observe:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k,$$

apresentando o η referente a função de ligação canônica, em que para o modelo de regressão normal linear, possui a mais simples função de ligação entre os componentes, a função de identidade. Observe a Tabela seguinte apresentada por Belfiore (FÁVERO; BELFIORE, 2017) em que são destacados outros exemplos de MLGs com suas respectivas funções de ligações.

Tabela 2 – Modelos lineares generalizados

Modelo de regressão	Característica da variável dependente Y	Distribuição	Função de Ligação canônica (η)
Linear	Quantitativa	Normal	$\hat{Y}$
Logística binária	Qualitativa com duas categorias	Bernoulli	$\ln\left(\frac{p}{1-p}\right)$
Logística multinomial	Qualitativa com mais de duas categorias	Binomial	$\ln\left(\frac{p_m}{1-p_m}\right)$
Poisson	Quantitativa com dados de contagem	Poisson	$\ln(\lambda)$
Binomial negativa	Quantitativa com dados de contagem	Poisson Gama	$\ln(\lambda)$

2.2.4.2 Regressão Logística

Dentre todos os modelos lineares generalizados, esse estudo irá especificar-se em um caso particular conhecido por modelo de regressão logística. Tal modelo é bastante utilizado na área da saúde para destacar os fatores que importam ou não para uma variável resposta. A exemplo do estudo desenvolvido por Muniz (MUNIZ et al., 2012) referente aos fatores de risco comportamentais acumulados para doenças cardiovasculares no sul do Brasil. No entanto, não só na área da saúde, mas em diversas áreas como a educação já foram elaborados estudos como a predição de desempenho de escolas privadas como, por exemplo, o estudo realizado por Adeodato e Rodrigues (ADEODATO; FILHO; RODRIGUES, 2014).

A natureza da variável resposta é um dos pontos de partida para a escolha do modelo que melhor se adequa ao problema. A comprovar isso, têm-se que a regressão logística é específica para quando a variável resposta é categórica. Quando há uma ordem nas classes de uma variável, essa é chamada ordinal, por isso, o modelo utilizado para quando a variável resposta segue esse tipo é o específico da regressão logística ordinal. Não existindo ordem na variável resposta, é caracterizada como nominal e o modelo realizado é o de regressão logística nominal. Por outro lado, referindo-se às variáveis preditoras, constantemente as aplicações de análise de regressão são de variáveis contínuas, porém para tratar os atributos categóricos, utiliza-se *dummys*, categorias particionadas em variáveis binárias distintas (GUEDES; IVANQUI; MARTINS, 2001).

Somado a isso, um modelo de regressão logístico também pode ser binário ou multinomial. Mais uma vez a definição está relacionada ao tipo da variável resposta, pois se ela for dicotômica, possuindo apenas duas categorias, é utilizado o modelo binário. Do mesmo modo, se ela possuir mais de duas classes, será considerado o modelo multinomial. Dessa forma, destaca-se que o estudo em questão trata do modelo de regressão logística binário e nominal. Sua função de distribuição e a derivação da mesma ou função de densidade são expressas, respectivamente, como:

$$F(x) = \frac{exp(-x)}{1 + exp(-x)},$$

$$f(x) = \frac{exp(-x)}{(1 + exp(-x))^2}.$$

O modelo é definido pela Equação (2.2), que utiliza a função logística (ou função sigmoide, dado o formato em "S") retornando valores no intervalo [0,1] acrescentada a equação:

$$logit\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i = \eta_i + \epsilon_i, \tag{2.2}$$

sendo π_i a estimativa para a esperança condicional de probabilidade de sucesso representada por:

$$\pi_i = \frac{exp^(\hat{\eta}_i)}{1 + exp^(\hat{\eta}_i)}.$$

A estimação dos parâmetros é realizada por meio do método da máxima-verossimilhança. Com a utilização desse método, é possível encontrar o vetor que maximiza a função densidade de probabilidade que está em função do vetor dos parâmetros desconhecidos, também conhecida como função de verossimilhança. Para isso, Portugal afirma que "basta igualar a zero as derivadas parciais da função de verossimilhança e achar o vetor de estimadores que resolve este conjunto de equações" (PORTUGAL, 1995).

Além disso, uma grande vantagem do modelo de regressão linear é a possibilidade de diagnosticar a importância das variáveis independentes, permitindo assim, uma seleção de quais variáveis devem ser mantidas ou retiradas do modelo. Normalmente, a seleção de variáveis inicia-se com a relação da variável resposta para cada variável seguindo com a análise de múltiplas variáveis selecionadas.

A seleção ocorre por meio do teste de Wald cuja estatística de teste segue uma distribuição qui-quadrado com g graus de liberdade, em que g é o número de restrições testadas. Esse teste, na regressão logística, é responsável por identificar o grau de significância de cada coeficiente, verificando se o parâmetro estimado é significantemente diferente de zero (hipótese nula) (WALD, 2004). Pode-se chegar à interpretação do teste por meio do p-valor (SHI; TAO, 2008), valor que indica a probabilidade de obter os resultados observados, assumindo a hipótese nula como verdadeira. Os p-valores que forem menores que o nível de significância adotado, rejeitam a hipótese nula de que a variável não é significativa para o modelo, por isso, serão candidatas ao modelo múltiplo. Quanto menor for o p-valor, maior será a significância da variável para o modelo. Do contrário, as variáveis são retiradas. No entanto, é possível ainda forçar a entrada de alguma variável caso seja de extrema importância para o estudo.

Embora seja possível aplicar o teste de Wald para dois ou mais parâmetros, o uso mais frequente para esse teste de hipóteses é um parâmetro por vez. Por isso, a análise múltipla é feita normalmente por meio da técnica *stepwise* (NETER et al., 1996). O objetivo desta técnica é identificar os atributos que maximizam a previsão da resposta com o menor número de variáveis explicativas empregadas. Dessa forma, por meio da seleção são identificadas quais variáveis são estatisticamente significantes para o modelo final.

Por fim, referindo-se a grande vantagem do modelo de regressão logística em ter os coeficientes interpretados é destacado pela *Odds Ratio* (OR). Essa medida, apresentada na Equação 2.3. trata-se da razão entre duas *odds*, ou seja, chances ou possibilidades, podendo ser resultada na exponencial dos coeficientes. A *odds* também refere-se a uma razão, sendo

a divisão entre a probabilidade de sucesso e a probabilidade de fracasso (complemento). Vale destacar, então, que chance não é o mesmo que probabilidade, tendo em vista que probabilidade seria a divisão entre o sucesso e o total.

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}, \tag{2.3}$$

sendo π_0 e π_1 as estimativas para a esperança condicional de probabilidade de sucesso do primeiro e segundo grupo já destacados anteriormente.

2.2.4.3 Modelo KNN

O algoritmo k *nearest neighbor* ou k vizinhos mais próximos, também conhecido como KNN, foi proposto por Cover e Hart (COVER; HART, 1967) e é um dos algoritmos mais utilizados na aprendizagem de máquina. Nesse método, utiliza-se os k dados de treinamento mais próximos ao vetor x para decidir sua resposta $\hat{y}$. Pequenos valores de k tendem a resultar em predições mais precisas, porém mais sensíveis a ruídos e irregularidades. Contudo, valores altos de k podem ocasionar em informações perdidas.

Conforme apresentado em (FIX; HODGES, 1989), Fix e Hodges introduziram, em 1951, o método não-paramétrico, por um relatório não publicado da Escola de Medicina de Aviação da Força Aérea dos Estados Unidos, como sendo apenas para $k = 1$. O modelo de KNN tem como vantagem a simplicidade, podendo ser usado mesmo quando não se sabe a distribuição dos dados. No entanto, é totalmente dependente dos dados observados.

2.2.4.4 Árvore de Decisão

Uma árvore de decisão é composta por nós internos e externos ligados por ramos. No método é utilizado uma hierarquia com base nas informações que influenciam a classificação de forma que há uma informação adquirida a cada variável analisada, ou seja, uma progressão na análise. Como afirma Schmitt (SCHMITT, 2013), o modelo "recebe como entrada um objeto ou uma situação descrita por um conjunto de propriedades ou atributos, e retorna como saída uma decisão", uma conclusão de classificação.

2.2.4.5 *Random Forest* (Floresta Aleatória)

Random Forest ou também conhecida como Floresta Aleatória introduzido por Breiman (BREIMAN, 2001), trata-se de um dos modelos que gera uma floresta de decisão ou conjuntos de árvores de decisões. O algoritmo possui dois parâmetros principais que são a *ntree* correspondente ao número de árvores aleatórias contidas, assim como o *mtry* relacionado ao número de características em cada nó interno da árvore (RIQUETI; RIBEIRO; ZÁRATE, 2018).

O algoritmo é considerado mais eficiente por não haver a necessidade das podas de árvores de decisões, pois é acrescentado a variabilidade controlada. Somado a isso, a classificação realizada pela Floresta Aleatória se dá pela escolha de qual instância foi votada por uma quantidade maior de árvores de decisões. Dessa forma, é observado na maioria das vezes, uma melhor acurácia do *Random Forest* quando comparado a outros modelos.

2.2.5 Avaliação de modelos

2.2.5.1 Acurácia

Diante do término da aplicação de diversos métodos à base de dados, é necessário avaliar se tais modelos foram bons, o quanto os modelos generalizam as informações aprendidas por meio do conjunto de treinamento ao conjunto de teste, e qual pode ser considerado melhor. Na avaliação de classificadores binários, por só apresentar duas classes, pode-se de dizer que uma classe trata-se da positiva e a outra, da negativa. Vale salientar que a classe positiva é destacada apenas por ser o evento de interesse, não necessitando ser um evento bom. Esse cenário é encontrado na pesquisa em questão por a classe positiva ser a evasão de alunos do ensino superior da Paraíba.

Dessa forma, por meio do conjunto de teste, a categoria real da variável resposta de uma determinada observação é comparada com a categoria prevista pelo modelo. Essa comparação permite quatro possibilidades de previsão diante de um classificador binário, conforme é apresentado na Tabela 3.

Tabela 3 – Possíveis resultados de um classificador binário

Resultado	Descrição	Classificação
Verdadeiro positivo (VP)	Positivo identificado como positivo	Classificação correta
Falso positivo (FP)	Negativo identificado como positivo	Erro tipo I
Verdadeiro negativo (VN)	Negativo identificado como negativo	Classificação correta
Falso negativo (FN)	Positivo identificado como negativo	Erro tipo II

As quantidades para cada um dos resultados, apresentados na Tabela 3, formam a matriz de confusão, uma matriz de probabilidades em que são indicadas as probabilidades de acerto, dada pela diagonal principal, e as de erro. Assim, a matriz é construída de modo a obter linhas que fazem referência às classes reais e colunas que fazem referência às classes preditas (DAVIS; GOADRICH, 2006).

Por meio da matriz de confusão, podem ser calculadas diversas medidas, sendo a acurácia a mais comum entre elas. Tal medida é calculada por meio da divisão entre os

resultados classificados corretamente e a soma de todos os elementos da matriz. Observe a Equação 2.4:

$$Acc = \frac{VP + VN}{VP + FN + VN + FP}. \tag{2.4}$$

Para análise dessa medida, tem-se a probabilidade de acertos para cada modelo. Sendo assim, quanto maior for a probabilidade, mais o modelo acerta na previsão. No entanto, muitas vezes pode ser preferível escolher um modelo com acurácia menor ao analisar a matriz de confusão, pois pode acontecer da classe maior ser bem predita pelo modelo, ou seja, com muitos acertos na previsão, mas o da classe menor, podendo ser a principal ou de interesse, apresentar maiores quantidades do erro tipo I ou tipo II. Sendo assim, é necessário cautela na análise.

2.2.5.2 Curva ROC e AUC

A Curva ROC é um gráfico que possibilita o estudo do desempenho do modelo por meio da sensibilidade, taxa de verdadeiro positivo apresentada no eixo y da curva, e especificidade, taxa de falso positivo apresentada no eixo x. Esse desempenho está relacionado a análise do poder preditivo do modelo. É esperado, então, que o modelo detecte o máximo possível de verdadeiros positivos, enquanto minimiza os falsos positivos.

Devido à dificuldade de comparar visualmente diferentes curvas ROC, reconhece-se a necessidade de uma medida como o AUC (área abaixo da curva) que possua esse objetivo. O valor do AUC varia de 0 a 1 (em porcentagem, de 0 a 100) de forma que quanto mais próximo de 1, indica uma maior área sob a curva, ou seja, um melhor desempenho do modelo, pois acerta mais do que erra na previsão (MEURER; TOLLES, 2017).

2.2.6 Fundamentando a evasão

Segundo Schmitt, historicamente a evasão é um termo utilizado para tratar das perdas estudantis (SCHMITT, 2014). No entanto, o Brasil não possui um consenso sobre o conceito de forma que diversos autores expõem suas diferentes perspectivas a respeito desse tema. Essas exposições trazem a discussão normalmente sobre quais dos discentes devem ser considerados evadidos ou não.

Existe a contribuição do estudo por Pereira (PEREIRA, 1996) que constata que na evasão "existem categorias que podem ser observadas, como abandono, cancelamento a pedido, cancelamento pela universidade e transferência para outra instituição". Esse mesmo autor, em acordo com Ristoff (RISTOFF, 1999), considera os discentes que migram de um curso para o outro não como evasão e sim, mobilidade. Em contrapartida, o MEC afirma que a evasão se difere em três modalidades: a evasão do curso, como o abandono, a desistência, exclusão por norma institucional, transferência ou reopção do curso; a evasão

da instituição, quando o discente desliga-se da instituição; por fim, a evasão do sistema, quando o ensino superior é abandonado de forma temporária ou definitiva pelo discente (UNIVERSIDADES; ESPECIAL; BORDAS, 1996).

É de grande importância definir o conceito de evasão que será utilizado na pesquisa dado que os critérios irão refletir nos resultados. Portanto, dentre os estudos citados no que diz respeito a evasão, esse trabalho irá abordar o conceito exposto pelo MEC quanto à evasão de sistema. Tendo em vista, então, que o censo do INEP divide a situação do aluno nas categorias "cursando", "matrícula trancada", "desvinculado do curso", "transferido para outro curso da mesma IES", "formado" e "falecido", os discentes com matrícula trancada ou que foram desvinculados do curso assumiu-se como evadidos. Vale salientar que foram desconsiderados todos os discentes falecidos. Observe:

Tabela 4 – Variável resposta dicotômica: evasão

Y	Evasão	Situação
$Y = 0$	Não	Discente cursando, transferido para outro curso da mesma IES ou formado.
$Y = 1$	Sim	Discente com matrícula trancada ou desvinculado do curso.

3 RESULTADOS

3.1 Pré-processamento dos dados

Por meio da importação dos microdados de discentes do censo do 2019 do Inep, realizou-se primeiramente o filtro dos dados da Paraíba. Para isso, necessitou-se realizar a mesclagem dos dados dos discentes com os dados das instituições, assim seria possível pegar as localidades de cada IES e consequentemente, dos alunos. Para melhor descrição sobre a situação dos discentes no estado, realizou-se a divisão das gerências regionais. No entanto, observou-se a falta de dados para as seguintes regiões: 4 (Cuité), 7 (Itaporanga), 11 (Princesa Isabel) e por fim, 14 (Mamanguape). Isso é justificado por as instituições presentes nos municípios citados não serem cadastradas no site do e-Mec, com exceção do município de Itaporanga que possui o cadastro da instituição "Faculdades Vale do Piancó — FAVAP", mas que em consulta direta ao código da instituição no banco de dados, não foi obtido nenhum registro de discente no ano em questão.

Em seguida, houve uma avaliação diante todas as variáveis apresentadas no censo. No entanto, observou-se que muitas delas não entram no contexto do estudo da evasão. Por isso, optou-se em primeiro momento, iniciar o estudo apenas com as variáveis listadas na Tabela do Anexo A somadas à variável de classificação das gerências regionais do estado.

Para a limpeza dos dados, foi avaliado o número de dados faltantes (NA — não aplicável) em todas as variáveis selecionadas contidas no banco. Dessa forma, dos 176.949 registros, 94.753 são registros faltantes distribuídos nas variáveis apresentadas na Tabela 5. Vale salientar, que essas informações faltantes, tratam-se de campos não preenchidos ou registrados no censo, por isso, os NA's apresentados na Seção 3.2, foram registros já classificados pelo Inep como dados faltantes ou não aplicáveis. Prosseguindo, nota-se que existe um excesso de NA's nas variáveis referentes ao turno de estudo do discente, ao estado e município de nascimento, ao discente possuir financiamento estudantil ou não e, finalmente, ao aluno estar regularmente matriculado em um curso de graduação, que se vincula temporariamente a outra instituição, sendo ela nacional ou internacional.

Escolheu-se dessa forma, tratar as variáveis referentes ao turno e grau acadêmico do discente, pois o número de informações faltantes não foi tão alto comparado às demais, que se optou por serem excluídas do banco. Para isso, o tratamento constou-se da criação de uma nova categoria que assuma apenas os NA's de cada variável, tendo em vista a possibilidade do turno e o grau acadêmico do discente não ser aplicável por determinado motivo. Ficando, assim, a distribuição das categorias conforme apresentado na Tabela 6.

Somado a isso, esse processo envolveu a transformação da variável "situação"

(TP_SITUAÇÃO) — antes, classificada como “cursando”, “matrícula trancada”, “desvinculado do curso”, “transferido para outro curso da mesma IES”, “formado” e “falecido” — para a variável binária “evasão”. Isso envolveu a retirada dos discentes falecidos, pois não se enquadravam no foco da pesquisa, conforme apresentado na Seção 2.2.6.

Tabela 5 – Variáveis com informações faltantes

Variáveis	Quantidade de NA’s
TP_TURNO	8177
TP_GRAU_ACADEMICO	40
CO_UF_NASCIMENTO	63126
CO_MUNICIPIO_NASCIMENTO	63126
IN_FINANCIAMENTO_ESTUDANTIL	93758
IN_MOBILIDADE_ACADEMICA	13132
Demais variáveis	0

Tabela 6 – Distribuição de categorias das variáveis tratadas

Categorias	TP_TURNO	Categorias	TP_GRAU_ACADEMICO
Matutino	38210 (21.59%)	Bacharelado	125739 (71.06%)
Vespertino	6675 (3.77%)	Licenciatura	35372 (19,98%)
Noturno	76511 (43.24%)	Tecnológico	15798 (8.93%)
Integral	47376 (26.77%)	Bacharelado e Licenciatura	-
Não aplicável (NA)	8177 (4.62%)	Não aplicável (NA)	40 (0.02%)

Tendo sido realizadas todas as alterações necessárias nas variáveis, preparou-se o banco para a aplicação dos modelos separando as variáveis em *dummys* e utilizando das técnicas de aprendizagem de máquina. Assim, 70% do banco foi selecionado aleatoriamente para ser o banco de treinamento enquanto os 30% restantes foram usados como teste para verificação da eficácia.

Por último, destacou-se um desbalanceamento na variável binária evasão, 77,80% dos dados de discentes não evadidos conforme apresentado na Figura e Tabela abaixo. É fato que o desbalanceamento influencia negativamente na modelagem, por isso, foi aplicado o algoritmo SMOTE para balanceamento dos dados de treinamento de modo que cada classe do atributo evasão passará a ter 96.350 registros.

Tabela 7 – Distribuição de categorias da evasão do conjunto de treinamento

Categorias	Proporção
0 (Não evadidos)	96.350 (77,80%)
1 (Evadidos)	27.514 (22,21%)

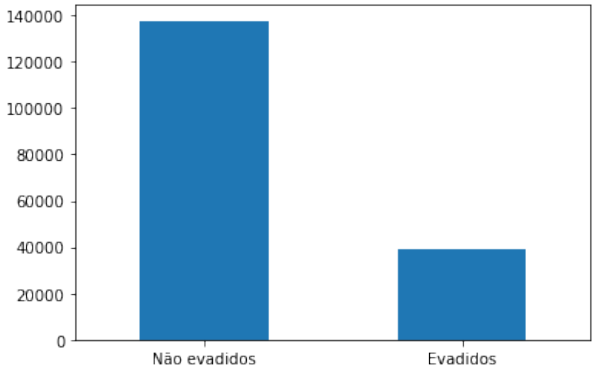

Figura 1 – Distribuição de categorias da evasão do conjunto de treinamento

3.2 Análise descritiva

Para descrição e obtenção de mais informações no tocante a alguns atributos presentes no banco de dados, realizou-se a análise descritiva. Vale salientar que essa etapa aconteceu antes mesmo da divisão entre dados de teste e treinamento.

Em relação às variáveis quantitativas, observou-se que os discentes apresentaram uma idade média de 26 anos com desvio padrão de 7 anos, variando entre discentes com 15 e 96 anos. Quanto à carga horária dos componentes curriculares que o discente tenha aproveitado, constatou-se em torno de 1.682 horas, em média, com desvio padrão de 1.411 horas com carga horária máxima de 14.740 horas. Observe nas Figuras 2 e 3 a distribuição dessas duas variáveis.

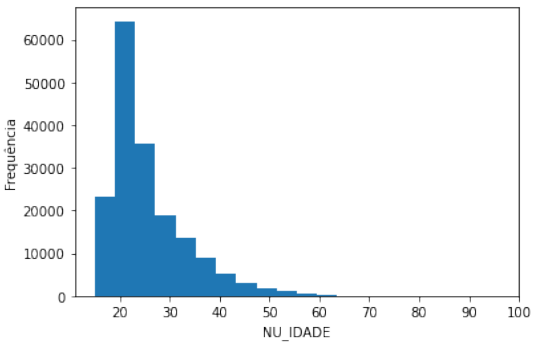

Figura 2 – Idade

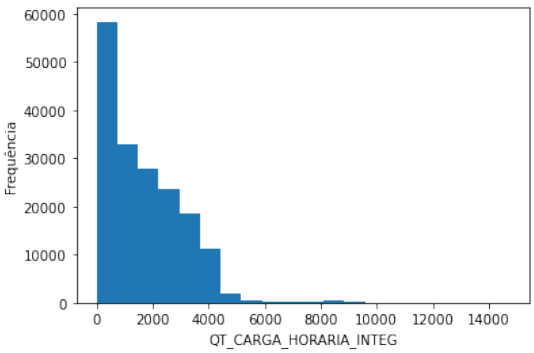

Figura 3 – Carga horária

Em seguida, observou-se por meio das variáveis categóricas, que se predominam discentes que estudam em instituições privadas com fins lucrativos (Figura 4), representando 41,14% de todos os discentes, e logo em seguida, discentes de instituições de categoria pública federal (39,85%). Ademais, conforme observado na Figura 5, se sobressai a quantidade de discentes que estudam no tipo de organização acadêmica universidade (45,08%), com o curso noturno (43,24%), como é possível observar na Figura 6, e finalmente, com grau bacharelado (71,06%) apresentado na Figura 7.

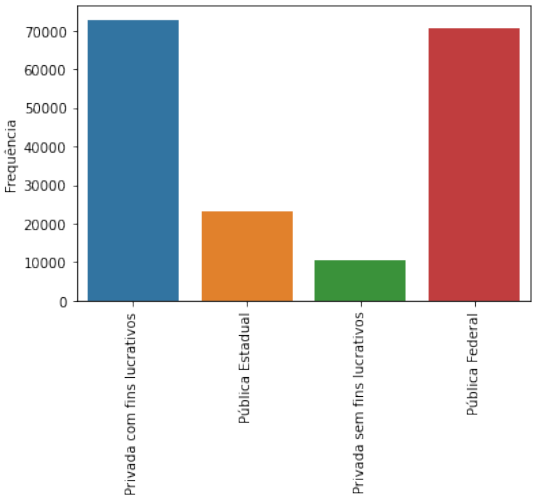

Figura 4 – Categoria administrativa

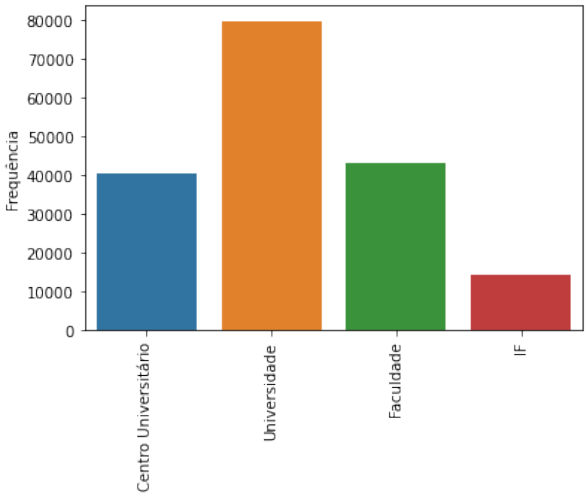

Figura 5 – Organização acadêmica

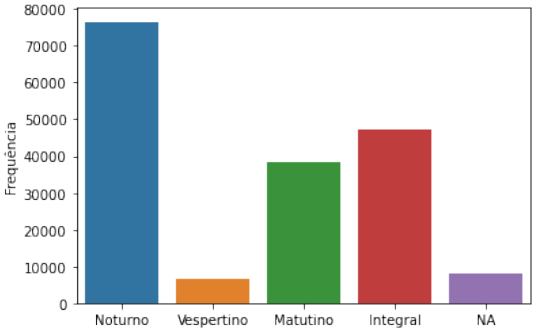

Figura 6 – Turno de estudo

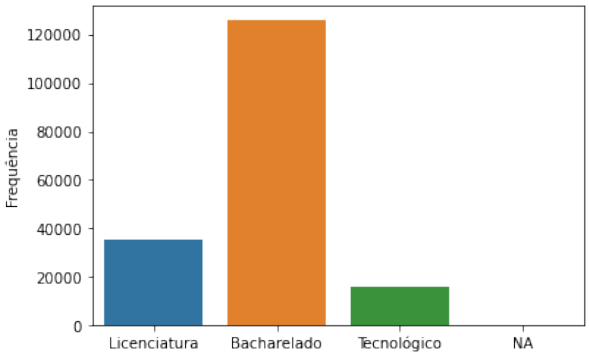

Figura 7 – Grau acadêmico

Quanto a relação entre os discentes que evadiram e outros atributos, notou-se que foi predominante, discentes que concluíram o ensino médio em escolas públicas (Figura 8), não participaram do programa de reserva de vagas (Figura 9), não receberam apoio social (Figura 10), não participaram de atividades (Figura 11) e não são ingressantes no curso (Figura 12), estando em qualquer outra situação diferente de ingresso. Destaca-se que o discente pode apresentar uma ou mais dessas características citadas.

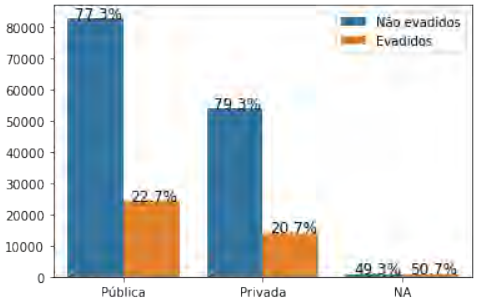

Figura 8 – Tipo de escola de conclusão do ensino médio X Evasão

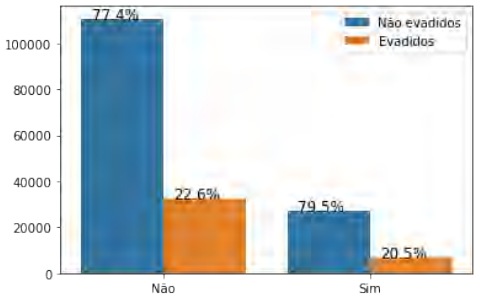

Figura 9 – Discente participante ou não do programa de reserva de vagas X Evasão

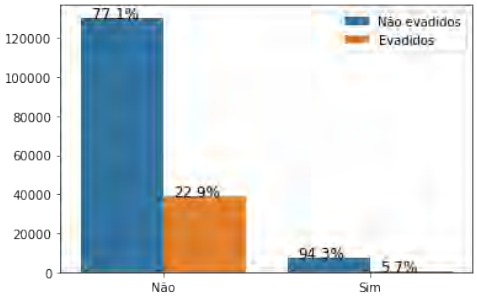

Figura 10 – Discente que recebe ou não apoio social X Evasão

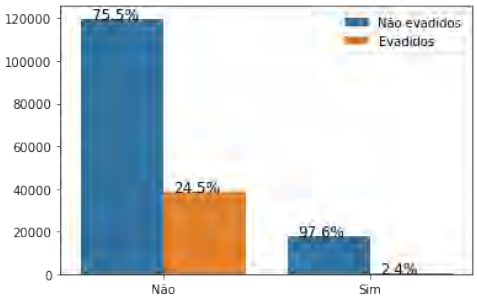

Figura 11 – Discente participante ou não de atividades extracurriculares X Evasão

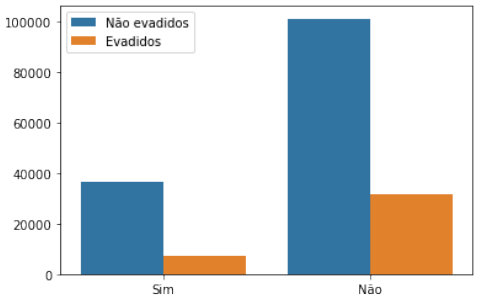

Figura 12 – Discente ingressante no curso ou não X Evasão

3.3 Seleção de variáveis

Quanto a seleção de variáveis, foram realizadas duas técnicas: a primeira, RFE, que aplicada como teste ao modelo de Regressão Logística, não obteve um bom ajuste, e a segunda, por meio do teste Qui-Quadrado. As técnicas foram utilizadas em conjunto com o algoritmo SMOTE, porém nesse trabalho serão apresentados apenas os resultados do teste Qui-Quadrado.

Dessa forma, por meio do teste Qui-quadrado, optou-se por selecionar as 10 variáveis que mais influenciam na evasão. Foram elas: a gerência geográfica, idade, a categoria administrativa, turno de estudo, grau acadêmico, se o aluno participa de programa de reserva de vagas, se o aluno recebe algum apoio social, se participa de alguma atividade extracurricular, se é ingressante no curso e por fim, a carga horária, aproveitada pelo discente, dos componentes curriculares que fazem parte da matriz do curso.

Essa seleção de variáveis foi utilizada para todos os modelos. No entanto, na Seção 3.4.1, foi apresentado uma etapa anterior a essa, como aplicação ao método de seleção por meio do teste de Wald apresentado na Seção 2.2.4.2, que possui maior precisão.

3.4 Aplicação dos modelos

3.4.1 Regressão Logística

Visando chegar ao modelo final de Regressão Logística, o processo de seleção das variáveis para o modelo iniciou-se de forma simples, com apenas uma variável explicativa. Portanto, para cada variável existente dentre as selecionadas na Seção 3.1, foi visualizado, uma a uma, o p-valor do teste de Wald com a variável de interesse ou resposta. Dentre elas, não foi viável fazer o modelo univariado com a variável referente ao curso do discente (CO_CURSO) e o município de nascimento do mesmo (CO_MUNICIPIO_NASCIMENTO), devido às muitas categorias contidas em cada variável. No mais, os resultados podem ser observados na Tabela 8 a seguir.

Conforme apresentado, por meio do p-valor foi possível verificar quais variáveis são ditas significativas para a evasão, ou seja, com o p-valor menor que o nível de significância de 5% adotado, implicando na hipótese nula rejeitada, ou seja, variável significativa para o modelo univariado. Já as variáveis que não apresentaram p-valores significativos para todas as categorias, são interpretadas por algumas das categorias não influenciarem na evasão dos discentes, mesmo assim, como foram minoria, optou-se por manter tais variáveis no modelo. No entanto, como esse primeiro processo de modelagem teve o intuito de identificar as variáveis que explicam, individualmente, a evasão, foram retiradas do modelo as variáveis não significativas, ou seja, que não rejeitaram a hipótese nula por apresentarem p-valor maior que o nível de significância adotado.

Tabela 8 – Modelo univariado com a variável resposta evasão

Classificação do p-valor	Variáveis testadas
Significativa	TP_ORGANIZACAO_ACADEMICA, TP_GRAU_ACADEMICO, TP_MODALIDADE_ENSINO, TP_SEXO, IN_DEFICIENCIA, IN_RESERVA_VAGAS, IN_FINANCIAMENTO_ESTUDANTIL, IN_APOIO_SOCIAL, IN_ATIVIDADE_EXTRACURRICULAR, TP_ESCOLA_CONCLUSAO_ENS_MEDIO, TP_SEMESTRE_REFERENCIA, IN_INGRESSO_TOTAL, NU_IDADE, QT_CARGA_HORARIA_TOTAL, QT_CARGA_HORARIA_INTEG
Muitas categorias significativas	CO_IES, TP_CATEGORIA_ADMINISTRATIVA, CO_CURSO, TP_COR_RACA, GERÊNCIAS REGIONAIS
Não significativa	TP_NIVEL_ACADEMICO, TP_NACIONALIDADE, CO_UF_NASCIMENTO, IN_MOBILIDADE_ACADEMICA, IN_MATRICULA, IN_CONCLUINTE

Prosseguiu-se, então, para o modelo de regressão logística múltipla, sendo as variáveis dependentes do modelo, todas as variáveis ditas significativas ou com muitas de suas categorias significativas no modelo univariado. Dessa forma, das 20 variáveis que estavam no modelo, seis foram retiradas, pois em conjunto, não foram significativas ao modelo, apresentando p-valor maior que o nível de 5% de significância. Para concluir, como as 14 variáveis restantes no modelo são significativas, aplicou-se a seleção de variáveis, por meio do teste Qui-quadrado, comum a todos os outros modelos apresentados na Seção 3.3, com o intuito de diminuir a quantidade de atributos deixando apenas os dez mais importantes.

Observando a eficácia do modelo de Regressão Logística, foram calculadas a acurácia e a matriz de confusão, apresentadas na Figura 13. Verificou-se uma acurácia de 76%, ou seja, essa é a probabilidade do modelo acertar na previsão do discente ter sido evadido ou não, com base nas informações das variáveis dependentes. Além disso, constatou-se, por meio da matriz de confusão que o modelo acertou 76% das vezes que o discente não foi evadido e 72% das vezes que o discente foi evadido. Em seguida, ao analisar a Figura 14, constatou-se que a curva se encontra relativamente próxima ao ponto (0,1), resultando em uma área sob a curva (AUC) igual a 82%, tendo um ponto de equilíbrio um pouco acima de 0,7 da taxa de verdadeiros positivos e um pouco mais que 0,2 da taxa de falsos positivos, tratando-se de um modelo moderado.

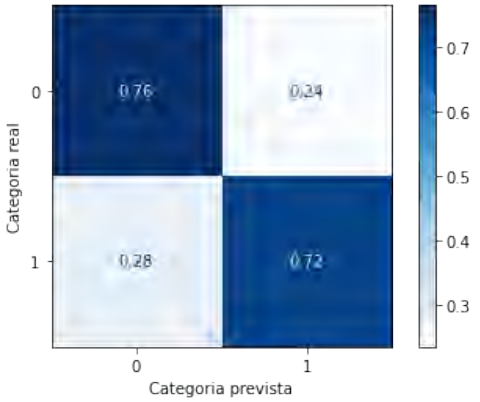

Figura 13 – Matriz de confusão
Regressão Logística

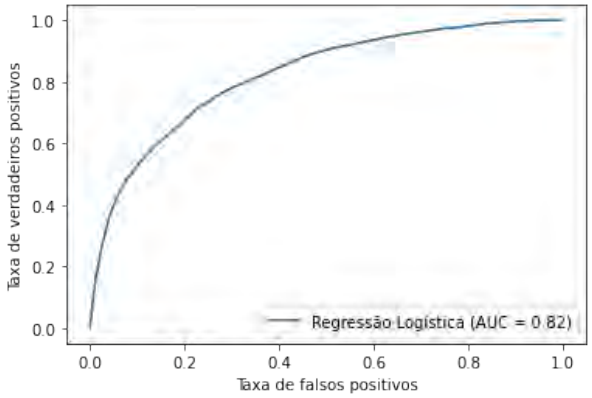

Figura 14 – Curva ROC
Regressão Logística

Como dito anteriormente, o grande diferencial do modelo de regressão logística para os demais apresentados nessa pesquisa é a possibilidade de medir a importância de cada variável, observando o efeito das variáveis independentes na variável resposta. Para isso, na Tabela 9 são apresentadas as porcentagens do inverso das exponenciais para cada variável do modelo ajustado para melhor interpretação dos dados.

Tabela 9 – Odds Ratio - Regressão Logística

Variável	Porcentagem
TP_CATEGORIA_ADMINISTRATIVA_4	41,76%
TP_CATEGORIA_ADMINISTRATIVA_5	85,07%
TP_TURNO_2	38,90%
TP_TURNO_3	43,18%
TP_TURNO_4	77,18%
TP_GRAU_ACADEMICO_2	7,50%
TP_GRAU_ACADEMICO_3	-18,36%
IN_RESERVA_VAGAS	-42,51%
IN_APOIO_SOCIAL	-85,57%
IN_ATIVIDADE_EXTRACURRICULAR	-94,45%
GERENCIA_GEOGRAFICA_2	-92,98%
GERENCIA_GEOGRAFICA_3	16,54%
GERENCIA_GEOGRAFICA_5	-17,20%
GERENCIA_GEOGRAFICA_6	-55,23%
GERENCIA_GEOGRAFICA_9	-52,07%
GERENCIA_GEOGRAFICA_10	-78,31%
NU_IDADE	6,65%
QT_CARGA_HORARIA_INTEG	-0,12%

Portanto, é possível chegar as seguintes afirmações:

a) Discentes estudando em instituições privadas com ou sem fins lucrativos têm maiores chances de 41,76% e 85,07%, respectivamente, de evadirem quando comparados a discentes das instituições públicas federais;

b) Discentes estudando em turno vespertino, noturno ou integral têm maiores chances de 38,90%, 43,18% e 77,18%, respectivamente, de evadirem do que discentes que

estudam no turno matutino;

c) Discentes de licenciatura possuem uma chance maior de 7,50% de evadirem do ensino superior do que discentes do bacharelado, porém, alunos de grau acadêmico tecnológico possuem menor chance, 18,36%, de evadirem do que alunos do bacharelado;

d) Se o discente participa de reserva de vagas, ele possui 42,51% menos chance de evadir em relação aos que não participam;

e) Se o discente recebe algum apoio social como moradia, transporte, alimentação, material didático e bolsas (trabalho/permanência), ele possui 85,57% menos chance de evadir em relação aos que não recebem nenhum apoio social;

f) Aqueles que participam de alguma atividade extracurricular (estágio não obrigatório, extensão, monitoria e pesquisa) possuem 94,45% menos chance de evadir do ensino superior do que os discentes que não participam de nenhuma atividade extracurricular;

g) Os discentes que estudam nas IES de Guarabira, Monteiro, Patos, Cajazeiras e Sousa possuem menor chance de evadirem do que os alunos que estudam nas instituições de João Pessoa, principalmente os estudantes de Guarabira com 92,98% e em menor proporção os estudantes de Patos com 17,20%, no entanto, os estudantes em Campina Grande possuem 16,54% maior chance de evadirem quando comparados aos alunos em João Pessoa;

h) Para cada ano a mais de idade do discente, as chances aumentam em 6,65% de evadirem do ensino superior da Paraíba;

i) A carga horária dos componentes curriculares que o discente já aproveitou e faz parte da matriz do curso pouco influencia na evasão, mostrando apenas que a cada hora a mais diminui em apenas 0,12% a chance do discente evadir.

3.4.2 Modelo KNN

Para o modelo KNN, utilizou-se a distância euclidiana e buscou-se obter resultados para três diferentes valores de k. Primeiramente, a modelagem foi realizada considerando apenas os dois vizinhos mais próximos ($k = 2$). Neste caso, foi observado uma acurácia de 80%, a maior em comparação aos demais valores de k apresentados mais a frente. Contudo, por meio da matriz de confusão da Figura 15, é possível observar que enquanto o modelo acerta 90% das vezes que os alunos não são evadidos, o modelo acerta apenas 45% das vezes que os alunos são evadidos, sendo este o foco principal da pesquisa. O valor da área abaixo da curva ROC apresentado na Figura 16 confirma a análise ao mostrar um valor de apenas 75%.

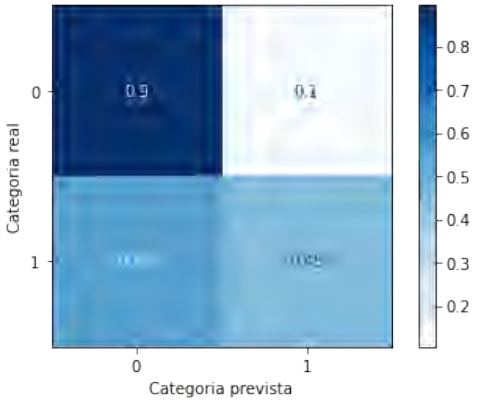

Figura 15 – Matriz de confusão
KNN ($k = 2$)

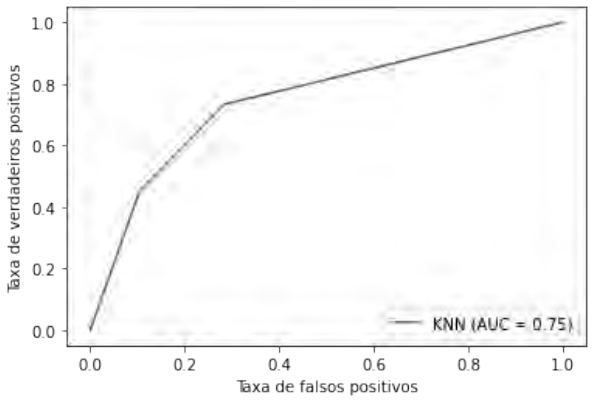

Figura 16 – Curva ROC
KNN ($k = 2$)

Agora, considerando cinco vizinhos mais próximos ($k = 5$), obtiveram-se os resultados apresentados nas Figuras 17 e 18. Apesar de a acurácia ter sofrido uma baixa de 4% (76%), verificou-se uma melhora no modelo para a classificação correta dos alunos evadidos, pois cresceu para 67%, além do AUC ter crescido para 0,79.

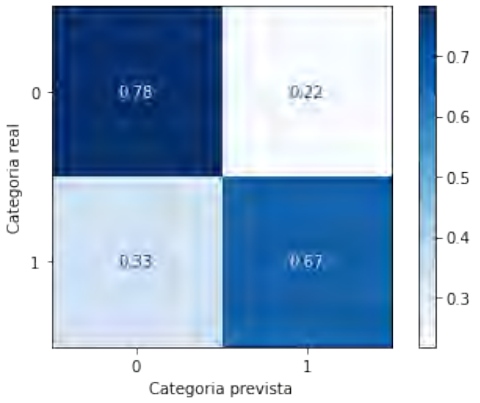

Figura 17 – Matriz de confusão
KNN ($k = 5$)

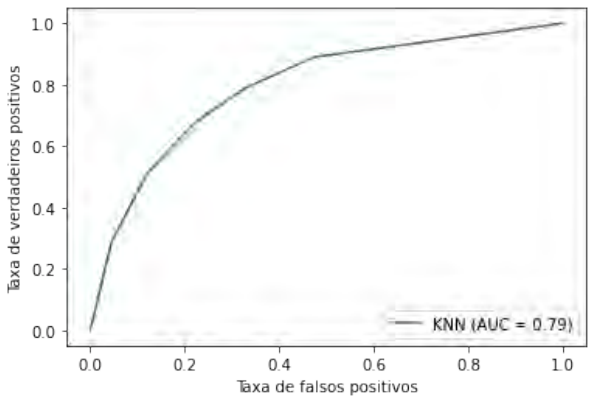

Figura 18 – Curva ROC
KNN ($k = 5$)

Finalmente, ao considerarmos nove vizinhos mais próximos ($k = 9$), constatou-se mais uma vez uma diminuição na acurácia (75%) e o modelo consegue acertar um pouco mais na previsão da classe minoritária, os alunos evadidos. Esse cenário, também influenciou na curva ROC de modo a aumentar o tamanho da área sob a curva, 81%. Observe as Figuras 19 e 20.

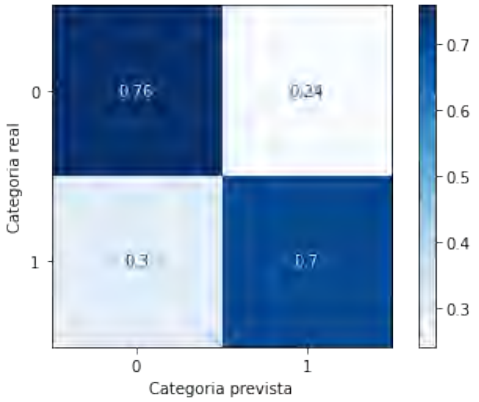

Figura 19 – Matriz de confusão KNN ($k = 9$)

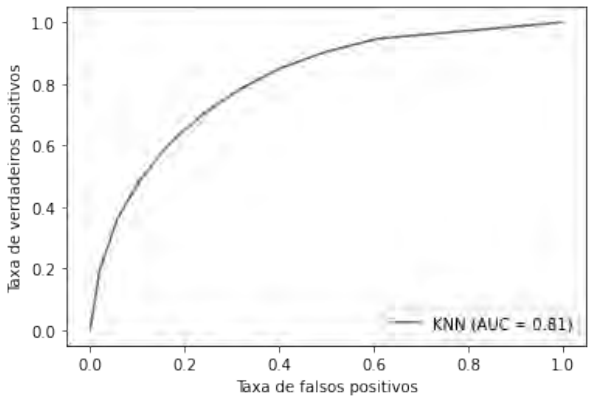

Figura 20 – Curva ROC KNN ($k = 9$)

3.4.3 Árvore de decisão

O ajuste do modelo da árvore de decisão atingiu uma acurácia igual a 78%, apresentando um certo equilíbrio na probabilidade de acertos para a classe majoritária e minoritária do conjunto teste. Para todos os alunos evadidos, o modelo previu corretamente em 76% das vezes e para todos os alunos não evadidos, o modelo previu corretamente em 78% das vezes. Somado a isso, na Figura 22 é possível observar como a curva ROC se aproxima um pouco mais do ponto (0,1) de modo a apresentar um AUC de 85%.

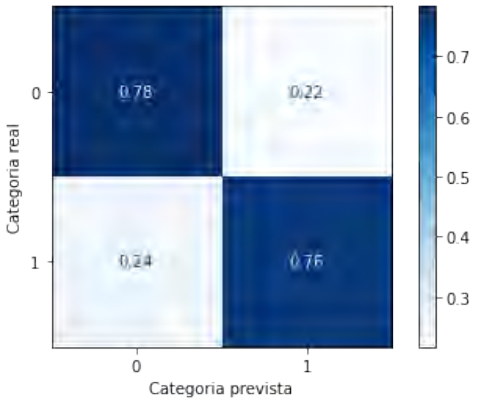

Figura 21 – Matriz de confusão Árvore de decisão

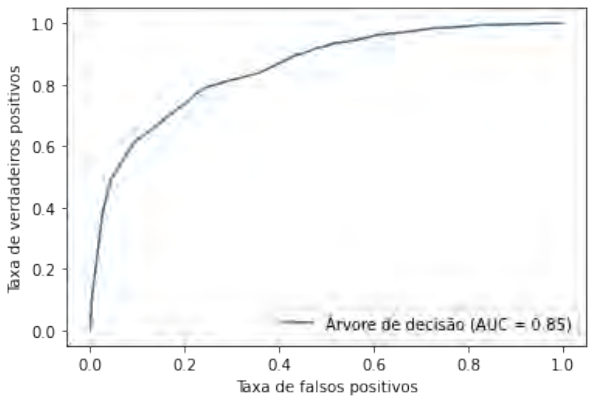

Figura 22 – Curva ROC Árvore de decisão

3.4.4 *Random Forest*

Por fim, o modelo de *Random Forest* parece se ajustar muito bem quando analisada a acurácia cujo valor foi de 82% e o AUC de 86%. No entanto, percebe-se que esse valor foi enganoso ao analisar a matriz de confusão em que apresenta, de forma clara, o modelo influenciado pela classe majoritária. O modelo conseguiu prever muito bem (87%) os

alunos que não evadiram, mas teve mais dificuldade de prever os alunos que evadiram, errando em 37% das vezes.

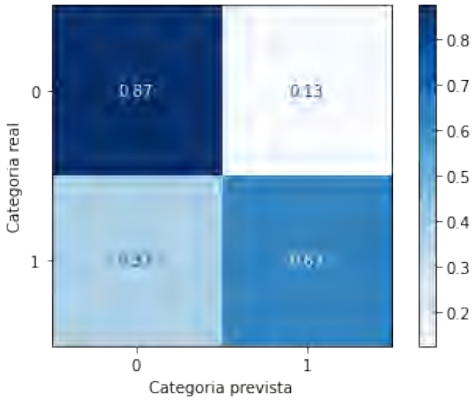

Figura 23 – Matriz de confusão Random Forest

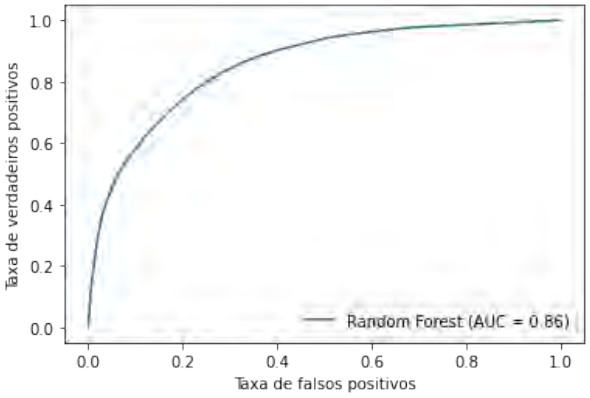

Figura 24 – Curva ROC Random Forest

4 CONCLUSÃO

Tendo em vista a grande expansão do ensino superior da Paraíba nas últimas décadas, considerou-se a necessidade de acrescentar estudos sobre o assunto, principalmente a respeito da evasão, pois possui a temática acompanhada do estudo referente às maiores dificuldades de permanência dos discentes nas instituições. Em particular, foi visto a importância de fazer a modelagem dos dados do censo do Inep do ano de 2019, aplicando uma metodologia de estatística e aprendizagem de máquina.

Uma dificuldade encontrada na pesquisa em questão tratou-se da abundante quantidade de dados. Isso contribuiu para ser considerado apenas o recorte do ano de 2019 e no estado da Paraíba. Ainda assim, vale salientar que os objetivos de levantar dados referentes aos discentes do ensino superior no ano de 2019, delimitando as gerências regionais do estado da Paraíba e modelar os dados de evasão com o fim de compreender quais variáveis influenciam na evasão foram atendidos no presente estudo.

Comparando os resultados obtidos para cada modelo, observou-se que a modelagem de dados reais torna-se de maior dificuldade por na maioria das vezes apresentar um desbalanceamento na variável de interesse. No cenário da pesquisa atual, percebeu-se essa dificuldade devido a quantidade de discentes evadidos ser bem menor que a quantidade de discentes não evadidos. Dessa forma, isso foi considerado ao avaliar cada modelo, que muitas vezes conseguia prever melhor a classe predominante em relação à classe minoritária e de maior interesse.

Na comparação dos modelos, observou-se que o modelo de *Random Forest* foi o modelo que apresentou maior área sob a curva ROC (AUC) e maior acurácia. Entretanto, considerando o problema da classe majoritária antes observada, percebeu-se que este modelo não se tratou do melhor, dado que não previu tão bem a classe minoritária. Então, o modelo que melhor apresentou eficiência na modelagem dos dados de discentes da Paraíba foi o modelo de Árvore de Decisão que apresentou uma acurácia um pouco menor ao modelo de *Random Forest*, mas que conseguiu prever melhor os discentes que evadiram. Em seguida, os melhores modelos foram o de Regressão Logística e o KNN considerando nove vizinhos mais próximos.

No entanto, pode-se concluir que diferentemente do modelo de Regressão Logística, os outros modelos podem se mostrar melhores em suas métricas e até um esforço menor para o desenvolvimento, mas possuem a desvantagem na capacidade de interpretar os seus parâmetros. Nesse ponto, o modelo de Regressão Logística pôde interpretar que os alunos de instituições privadas, em turno diferente ao matutino, em cursos de licenciatura, não participando de reserva de vagas, nem de alguma atividade extracurricular ou sem algum

apoio social possuem maior probabilidade de evadirem do ensino superior da Paraíba. Tais achados são úteis na tomada de decisões como a elaboração das estratégias de gestão das instituições de educação e regulação de políticas e projetos educacionais voltadas aos discentes de educação superior.

Em estudos futuros, sugere-se uma maior abrangência em tempo e localidade, podendo ser acompanhado os demais anos e outras regiões. Além disso, sugere-se um aprimoramento nas técnicas utilizadas, podendo-se aplicar outros modelos com maior complexidade para verificar se há uma melhor eficácia em relação aos modelos já apresentados nesse estudo, bem como diferentes técnicas de validação cruzada que irão aprimorar a prevenção do *overfitting*. Somado a isso, a utilização de outras métricas como a precisão e sensibilidade podem também atender aos objetivos da pesquisa complementando os resultados já apresentados com a curva ROC e acurácia.

REFERÊNCIAS

ADEODATO, P. J.; FILHO, M. M. S.; RODRIGUES, R. L. Predição de desempenho de escolas privadas usando o enem como indicador de qualidade escolar. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2014. v. 25, n. 1, p. 891. Citado na página 21.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020. Citado na página 17.

ALVES, J. M. S. *Dos mínimos quadrados à regressão linear: atividades históricas sobre função afim e estatística usando planilhas eletrônicas*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2016. Citado na página 18.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 23.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado na página 17.

CHEIN, F. Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas. Escola Nacional de Administração Pública (Enap), 2019. Citado na página 18.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado na página 23.

DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd international conference on Machine learning*. [S.l.: s.n.], 2006. p. 233–240. Citado na página 24.

DIAS, S. M. B.; COSTA, S. L. da. A permanência no ensino superior e as estratégias institucionais de enfrentamento da evasão. *Jornal de Políticas Educacionais*, v. 9, n. 17/18, 2016. Citado na página 13.

FÁVERO, L. P.; BELFIORE, P. *Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®*. [S.l.]: Elsevier Brasil, 2017. Citado na página 20.

FIX, E.; HODGES, J. L. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, JSTOR, v. 57, n. 3, p. 238–247, 1989. Citado na página 23.

GUEDES, T. A.; IVANQUI, I. L.; MARTINS, A. B. T. Comparando equações de regressão em dados de saúde. *Acta Scientiarum. Technology*, v. 23, p. 1531–1535, 2001. Citado na página 21.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. *Machine learning*, Springer, v. 46, n. 1, p. 389–422, 2002. Citado na página 17.

HAN, J.; JIAN, P.; MICHELIN, K. Data mining, southeast asia edition. *ProQuest Ebook Central https://ebookcentral. proquest. com*, 2006. Citado na página 16.

HELENE, O. *Metodos dos Minimos Quadrados*. [S.l.]: Editora Livraria da Física, 2006. Citado na página 19.

HOUAISS, A. Grande dicionário houaiss da língua portuguesa on-line. 2012. *Acesso em: jul*, 2014. Citado na página 15.

JOHNSON, R. A.; WICHERN, D. W. et al. *Applied multivariate statistical analysis*. [S.l.]: Pearson London, UK:, 2014. v. 6. Citado na página 16.

LOPES, R. P.; MESQUITA, C. O impacto do ensino superior na construção do pensamento de ordem superior. *Revista de Estudios e Investigación en Psicología y Educación*, Universidade da Coruña, p. 127–131, 2017. Citado na página 13.

MEURER, W. J.; TOLLES, J. Logistic regression diagnostics: understanding how well a model predicts outcomes. *Jama*, American Medical Association, v. 317, n. 10, p. 1068–1069, 2017. Citado na página 25.

MUNIZ, L. C. et al. Fatores de risco comportamentais acumulados para doenças cardiovasculares no sul do brasil. *Revista de Saúde Pública*, SciELO Brasil, v. 46, p. 534–542, 2012. Citado na página 21.

NETER, J. et al. Applied linear statistical models. Irwin Chicago, 1996. Citado na página 22.

PEARSON, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor & Francis, v. 50, n. 302, p. 157–175, 1900. Citado na página 18.

PEREIRA, J. T. V. Uma contribuição para o entendimento da evasão um estudo de caso: Unicamp. *Avaliação: Revista da Avaliação da Educação Superior*, v. 1, n. 2, 1996. Citado na página 25.

PESSOA, D. G. C.; SILVA, P. L. N. Análise de dados amostrais complexos. *São Paulo: Associação Brasileira de Estatística*, v. 1, 1998. Citado na página 18.

PORTUGAL, M. S. Notas introdutórias sobre o princípio de máxima verossimilhança: Estimação e teste de hipóteses. *DECON/UFRGS, Porto Alegre, Abril*, 1995. Citado na página 22.

RIQUETI, G.; RIBEIRO, C.; ZÁRATE, L. Classificando perfis de longevidade de bases de dados longitudinais usando floresta aleatória. 2018. Citado na página 23.

RISTOFF, D. *Universidade em foco: reflexões sobre a educação superior*. Florianópolis: Insular, 1999. Citado na página 25.

SCHMITT, R. E. A evasão na educação superior: uma compreensão ecológica do fenômeno como estratégia para a gestão da permanência estudantil. *X Seminário de Pesquisa em Educação da Região Sul – ANPEDSUL*, 2014. Citado na página 25.

SCHMITT, V. F. Uma análise comparativa de técnicas de aprendizagem de máquina para prever a popularidade de postagens no facebook. 2013. Citado 2 vezes nas páginas 17 e 23.

SHI, N.-Z.; TAO, J. *Statistical hypothesis testing: theory and methods*. [S.l.]: World Scientific Publishing Company, 2008. Citado na página 22.

SOUZA, E. F. M. de; PETERNELLI, L. A.; MELLO, M. P. de. Software livre r: aplicação estatística. 2014. Citado na página 15.

THOMPSON, S. K. *Sampling*. [S.l.]: Wiley Series in Probability and Mathematical Statistics, 1992. Citado na página 18.

TURKMAN, M. A. A.; SILVA, G. L. Modelos lineares generalizados: da teoria à prática. In: *VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa*. [S.l.: s.n.], 2000. Citado na página 20.

UNIVERSIDADES, P. d. A. I. das; ESPECIAL, B. C.; BORDAS, M. C. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas: resumo do relatório apresentado a adifes, abruem e sesu/mec pela comissão especial. *Avaliação: revista da Rede de Avaliação Institucional da Educação Superior. Campinas, SP. Vol. 1, n. 2 (dez. 1996), p. 55-65*, 1996. Citado na página 26.

VALLE, P. O. do et al. Análise de variância e análise de regressão com variáveis dummy: Mais semelhanças do que diferenças. *Revista de Estatística*, v. 1, p. 47–86, 2002. Citado na página 16.

WALD, A. *Sequential analysis*. [S.l.]: Courier Corporation, 2004. Citado na página 22.

Anexos

ANEXO A – CENSO INEP

Tabela 10 – Variáveis do censo de Ensino Superior

Nome da Variável	Descrição da variável
TP_CATEGORIA_ADMINISTRATIVA	Tipo da Categoria Administrativa da IES
TP_ORGANIZACAO_ACADEMICA	Tipo da Organização Acadêmica da IES
TP_TURNO	Tipo do turno do curso ao qual o aluno está vinculado
TP_GRAU_ACADEMICO	Tipo do grau acadêmico conferido ao diplomado pelo curso
TP_MODALIDADE_ENSINO	Tipo da modalidade de ensino do curso
TP_NIVEL_ACADEMICO	Tipo do nível acadêmico do curso
TP_COR_RACA	Tipo da cor/raça do aluno
TP_SEXO	Informa o sexo do aluno
NU_IDADE	Idade que o aluno completa no ano de referência do Censo
TP_NACIONALIDADE	Tipo da nacionalidade do aluno
IN_DEFICIENCIA	Informa se o aluno é uma pessoa com deficiência, transtorno global do desenvolvimento ou altas habilidades/superdotação
TP_SITUACAO	Tipo de situação de vínculo do aluno no curso
QT_CARGA_HORARIA_TOTAL	Somatório do total da carga horária dos componentes curriculares que fazem parte da matriz do curso
QT_CARGA_HORARIA_INTEG	Somatório da carga horária dos componentes curriculares que o aluno tenha aproveitado e que fazem parte da matriz do curso
TP_ESCOLA_CONCLUSAO_ENS_MEDIO	Tipo de escola que o aluno concluiu ensino médio
TP_SEMESTRE_REFERENCIA	Semestre de referência do preenchimento do vínculo do curso
IN_MOBILIDADE_ACADEMICA	Informa se o aluno está regularmente matriculado em curso de graduação, que se vincula temporariamente a outra instituição, sendo ela nacional ou internacional
IN_MATRICULA	Informa se o aluno é matriculado no curso
IN_CONCLUINTE	Informa se o aluno é concluinte
IN_INGRESSO_TOTAL	Informa se o aluno é ingressante no curso, não importando a forma de ingresso utilizada.
IN_FINANCIAMENTO_ESTUDANTIL	Informa se o aluno utiliza financiamento estudantil
IN_RESERVA_VAGAS	Informa se o aluno participa de programa de reserva de vagas
IN_APOIO_SOCIAL	Informa se o aluno recebe algum tipo de apoio social na forma de moradia, transporte, alimentação, material didático e bolsas (trabalho/permanência)
IN_ATIVIDADE_EXTRACURRICULAR	Informa se o aluno participa de algum tipo de atividade extracurricular (estágio não obrigatório, extensão, monitoria e pesquisa)

Tabela 11 – Códigos das variáveis do censo de Ensino Superior

Código da Variável	Descrição da variável
CO_UF	Código do IBGE da Unidade da Federação da IES
CO_IES	Código único de identificação da IES
CO_CURSO	Código único de identificação do curso gerado pelo E-MEC
CO_UF_NASCIMENTO	Código do IBGE da Unidade da Federação de nascimento do aluno
CO_MUNICIPIO_NASCIMENTO	Código do IBGE do município de nascimento do aluno

ANEXO B – GERÊNCIAS REGIONAIS

Tabela 12 – Gerências regionais - Paraíba - Região 1 a 9

Gerência Regional	Municípios
Região 1 - João Pessoa	Mari, Sapé, Riachão do Poço, Sobrado, Cruz do Espírito Santo, Santa Rita, Bayeux, João Pessoa, Conde, Alhandra, Caaporã, Pitimbu, Lucena, Cabedelo
Região 2 - Guarabira	Araruna, Cacimba de Dentro, Casserengue, Solânea, Arara, Serraria, Pilões, Alagoinha, Mulungu, Cuitegi, Guarabira, Araçagi, Borborema, Pilõezinhos, Pirpirituba, Belém, Sertãozinho, Lagoa de Dentro, Serra da Raiz, Logradouro, Caiçara, Dona Inês, Riachão, Tacima, Duas Estradas
Região 3 - Campina Grande	Livramento, Taperoá, Assunção, Tenório, Juazeirinho, Soledade, Olivedos, Boa Vista, Cabaceiras, São Domingos do Cariri, Barra, de São Miguel, Riacho de Santo Antônio, Alcantil, Santa Cecília, Gado Bravo, Umbuzeiro, Natuba, Aroeiras, Itatuba, Fagundes, Campina Grande, Boqueirão, Caturité, Barra de Santana, Queimadas, Pocinhos, Montadas, Areial, Puxinanã, Massaranduba, Lagoa Seca, Matinhas, São Sebastião de Lagoa de Roça, Alagoa Nova, Esperança, Juarez Tavora, Alagoa Grande, Areia, Remígio, Algodão de Jandaíra
Região 4 - Cuité	Frei Martinho, Picuí, Nova Palmeira, Pedra Lavrada, Cubati, São Vicente do Seridó, Sossêgo, Barra de Santa Rosa, Damião, Cuité, Nova Floresta, Baraúna
Região 5 - Monteiro	Santo André, Gurjão, Parari, São José dos Cordeiros, Amparo, Ouro Velho, Prata, Monteiro, Zabelê, São Sebastião do Umbuzeiro, São João do Tigre, Camalaú, Congo, Caraúbas, Coxixola, São João do Cariri, Serra Branca, Sumé
Região 6 - Patos	Emas, Catingueira, Santa Teresinha, Mãe D'água, Maturéia, Teixeira, São José do Bonfim, Cacimba de Areia, Cacimbas, Desterro, Passagem, Areia de Baraúnas, Salgadinho, Junco do Seridó, Quixabá, São Mamede, Santa Luzia, São José do Sabugi, Várzea, Patos, São José de Espinharas, Malta
Região 7 - Itaporanga	Santa Inês, Conceição, Ibiara, Santana de Mangueira, Curral Velho, Boa Ventura, Diamante, Pedra Branca, Nova Olinda, Santana dos Garrotes, Itaporanga, São José de Caiana, Serra Grande, Aguiar, Igaracy, Piancó, Olho D'água, Coremas
Região 8 - Catolé do Rocha	Bom Sucesso, Lagoa, Jericó, Mato Grosso, Riacho dos Cavalos, Brejo dos Santos, São Bento, Catolé do Rocha, Belém do Brejo do Cruz, São José do Brejo do Cruz, Brejo do Cruz
Região 9 - Cajazeiras	Poço Dantas, Bernardino Batista, Joca Claudino, Uiraúna, Triunfo, Poço de José de Moura, São João do Rio do Peixe, Santa Helena, Bom Jesus, Cajazeiras, Cachoeira dos Índios, São José de Piranhas, Carrapateira, Monte Horebe, Bonito de Santa Fé

Tabela 13 – Gerências regionais - Paraíba - Região 10 a 14

Gerência Regional	Municípios
Região 10 - Sousa	Vieirópolis, Lastro, Santa Cruz, São Francisco, Sousa, Marizópolis, São José da Lagoa Tapada, Nazarezinho
Região 11 - Princesa Isabel	Manaíra, São José de Princesa, Princesa Isabel, Tavares, Juru, Água Branca, Imaculada
Região 12 - Itabaiana	Serra Redonda, Riachão do Bacamarte, Ingá, Mogeiro, Salgado de São Félix, Itabaiana, Gurinhém, Caldas Brandão, São José dos Ramos, Pilar, São Miguel de Taipú, Juripiranga, Pedras de Fogo
Região 13 - Pombal	Aparecida, São Domingos, Pombal, Paulista, Vista Serrana, São Bentinho, Condado, Cajazeirinhas
Região 14 - Mamanguape	Jacaraú, Curral de Cima, Itapororoca, Cuité de Mamanguape, Capim, Mamanguape, Rio Tinto, Marcação, Baía de Traição, Mataraca, Pedro Régis

ANEXO C – IMPLEMENTAÇÃO COMPUTACIONAL

C.1 Linguagem R

```
## Bibliotecas

library(data.table) # importação
library(dplyr) # tratamento dos dados
library(caTools) # divisão de treinamento e teste

## Importação de base de dados muito grande

local <- file.choose()
banco_alunos <- fread(file = local)
head(banco_alunos)

## Filtrando apenas a Paraíba

banco <- banco_alunos

local <- file.choose()
banco_ies <- read.csv(local, header = T, sep = "|", dec = ".")
attach(banco_ies)
banco_ies <- banco_ies[CO_UF == 25,]

banco_ies <- banco_ies %>%
  select(CO_IES, CO_MUNICIPIO, CO_UF)

banco_pb <- inner_join(banco_ies, banco, by = c("CO_IES"))

## Criação da coluna de gerências regionais

attach(banco_pb)

for(linha in 1:length(banco_pb$CO_MUNICIPIO))
```

```
{
  if (CO_MUNICIPIO[linha] ==  2509107 | CO_MUNICIPIO[linha] ==  2515302 |
      CO_MUNICIPIO[linha] ==  2512762 | CO_MUNICIPIO[linha] ==  2515971|
      CO_MUNICIPIO[linha] ==  2504900 | CO_MUNICIPIO[linha] ==  2513703 |
      CO_MUNICIPIO[linha] ==  2501807 | CO_MUNICIPIO[linha] ==  2507507 |
      CO_MUNICIPIO[linha] ==  2504603 | CO_MUNICIPIO[linha] ==  2500601|
      CO_MUNICIPIO[linha] ==  2503001 | CO_MUNICIPIO[linha] ==  2511905 |
      CO_MUNICIPIO[linha] ==  2508604 | CO_MUNICIPIO[linha] ==  2503209)
    banco_pb$gerencia_geografica[linha] <-  1
  if (CO_MUNICIPIO[linha] == 2501005 | CO_MUNICIPIO[linha] == 2503506 |
      CO_MUNICIPIO[linha] == 2504157 | CO_MUNICIPIO[linha] == 2516003 |
      CO_MUNICIPIO[linha] == 2500908 | CO_MUNICIPIO[linha] == 2515906 |
      CO_MUNICIPIO[linha] == 2511608 | CO_MUNICIPIO[linha] == 2500502 |
      CO_MUNICIPIO[linha] == 2509800 | CO_MUNICIPIO[linha] == 2505204 |
      CO_MUNICIPIO[linha] == 2506301 | CO_MUNICIPIO[linha] == 2500809 |
      CO_MUNICIPIO[linha] == 2502706 | CO_MUNICIPIO[linha] == 2501500 |
      CO_MUNICIPIO[linha] == 2511707 | CO_MUNICIPIO[linha] == 2511806 |
      CO_MUNICIPIO[linha] == 2501906 | CO_MUNICIPIO[linha] == 2515930 |
      CO_MUNICIPIO[linha] == 2508208 | CO_MUNICIPIO[linha] == 2515609 |
      CO_MUNICIPIO[linha] == 2508554 | CO_MUNICIPIO[linha] == 2503605 |
      CO_MUNICIPIO[linha] == 2505709 | CO_MUNICIPIO[linha] == 2512747 |
      CO_MUNICIPIO[linha] == 2516409 | CO_MUNICIPIO[linha] == 2505808)
    banco_pb$gerencia_geografica[linha] <-  2
  if (CO_MUNICIPIO[linha] == 2508505 | CO_MUNICIPIO[linha] == 2516508 |
      CO_MUNICIPIO[linha] == 2501351 | CO_MUNICIPIO[linha] == 2516755 |
      CO_MUNICIPIO[linha] == 2507705 | CO_MUNICIPIO[linha] == 2516102 |
      CO_MUNICIPIO[linha] == 2510501 | CO_MUNICIPIO[linha] == 2502151 |
      CO_MUNICIPIO[linha] == 2503100 | CO_MUNICIPIO[linha] == 2513943 |
      CO_MUNICIPIO[linha] == 2501708 | CO_MUNICIPIO[linha] == 2512788 |
      CO_MUNICIPIO[linha] == 2500536 | CO_MUNICIPIO[linha] == 2513158 |
      CO_MUNICIPIO[linha] == 2506251 | CO_MUNICIPIO[linha] == 2517001 |
      CO_MUNICIPIO[linha] == 2509909 | CO_MUNICIPIO[linha] == 2501302 |
      CO_MUNICIPIO[linha] == 2507200 | CO_MUNICIPIO[linha] == 2506103 |
      CO_MUNICIPIO[linha] == 2504009 | CO_MUNICIPIO[linha] == 2502508 |
      CO_MUNICIPIO[linha] == 2504355 | CO_MUNICIPIO[linha] == 2501575 |
      CO_MUNICIPIO[linha] == 2512507 | CO_MUNICIPIO[linha] == 2512002 |
      CO_MUNICIPIO[linha] == 2509503 | CO_MUNICIPIO[linha] == 2501203 |
      CO_MUNICIPIO[linha] == 2512408 | CO_MUNICIPIO[linha] == 2509206 |
      CO_MUNICIPIO[linha] == 2508307 | CO_MUNICIPIO[linha] == 2509339 |
```

```
    CO_MUNICIPIO[linha] == 2515104 | CO_MUNICIPIO[linha] == 2500403 |
    CO_MUNICIPIO[linha] == 2506004 | CO_MUNICIPIO[linha] == 2507606 |
    CO_MUNICIPIO[linha] == 2500304 | CO_MUNICIPIO[linha] == 2501104 |
    CO_MUNICIPIO[linha] == 2512705 | CO_MUNICIPIO[linha] == 2500577)
  banco_pb$gerencia_geografica[linha] <-  3
if (CO_MUNICIPIO[linha] == 2506202 | CO_MUNICIPIO[linha] == 2511400 |
    CO_MUNICIPIO[linha] == 2510303 | CO_MUNICIPIO[linha] == 2511103 |
    CO_MUNICIPIO[linha] == 2505006 | CO_MUNICIPIO[linha] == 2515401 |
    CO_MUNICIPIO[linha] == 2516151 | CO_MUNICIPIO[linha] == 2501609 |
    CO_MUNICIPIO[linha] == 2505352 | CO_MUNICIPIO[linha] == 2505105 |
    CO_MUNICIPIO[linha] == 2510105 | CO_MUNICIPIO[linha] == 2501534)
  banco_pb$gerencia_geografica[linha] <- 4
if (CO_MUNICIPIO[linha] == 2513851 | CO_MUNICIPIO[linha] == 2506509 |
    CO_MUNICIPIO[linha] == 2510659 | CO_MUNICIPIO[linha] == 2514800 |
    CO_MUNICIPIO[linha] == 2500734 | CO_MUNICIPIO[linha] == 2510600 |
    CO_MUNICIPIO[linha] == 2512200 | CO_MUNICIPIO[linha] == 2509701 |
    CO_MUNICIPIO[linha] == 2517407 | CO_MUNICIPIO[linha] == 2515203 |
    CO_MUNICIPIO[linha] == 2514107 | CO_MUNICIPIO[linha] == 2503902 |
    CO_MUNICIPIO[linha] == 2504702 | CO_MUNICIPIO[linha] == 2504074 |
    CO_MUNICIPIO[linha] == 2504850 | CO_MUNICIPIO[linha] == 2514008 |
    CO_MUNICIPIO[linha] == 2515500 | CO_MUNICIPIO[linha] == 2516300)
  banco_pb$gerencia_geografica[linha] <-  5
if (CO_MUNICIPIO[linha] == 2505907 | CO_MUNICIPIO[linha] == 2504207 |
    CO_MUNICIPIO[linha] == 2513802 | CO_MUNICIPIO[linha] == 2508703 |
    CO_MUNICIPIO[linha] == 2509396 | CO_MUNICIPIO[linha] == 2516706 |
    CO_MUNICIPIO[linha] == 2514602 | CO_MUNICIPIO[linha] == 2503407 |
    CO_MUNICIPIO[linha] == 2503555 | CO_MUNICIPIO[linha] == 2505402 |
    CO_MUNICIPIO[linha] == 2510709 | CO_MUNICIPIO[linha] == 2501153 |
    CO_MUNICIPIO[linha] == 2513000 | CO_MUNICIPIO[linha] == 2507804 |
    CO_MUNICIPIO[linha] == 2512606 | CO_MUNICIPIO[linha] == 2514909 |
    CO_MUNICIPIO[linha] == 2513406 | CO_MUNICIPIO[linha] == 2514701 |
    CO_MUNICIPIO[linha] == 2517100 | CO_MUNICIPIO[linha] == 2510808 |
    CO_MUNICIPIO[linha] == 2514404 | CO_MUNICIPIO[linha] == 2508802)
  banco_pb$gerencia_geografica[linha] <-  6
if (CO_MUNICIPIO[linha] == 2513356 | CO_MUNICIPIO[linha] == 2504405 |
    CO_MUNICIPIO[linha] == 2506608 | CO_MUNICIPIO[linha] == 2513505 |
    CO_MUNICIPIO[linha] == 2505303 | CO_MUNICIPIO[linha] == 2502102 |
    CO_MUNICIPIO[linha] == 2505600 | CO_MUNICIPIO[linha] == 2511004 |
    CO_MUNICIPIO[linha] == 2510204 | CO_MUNICIPIO[linha] == 2513604 |
```

```
    CO_MUNICIPIO[linha] == 2507002 | CO_MUNICIPIO[linha] == 2514305 |
    CO_MUNICIPIO[linha] == 2515708 | CO_MUNICIPIO[linha] == 2500205 |
    CO_MUNICIPIO[linha] == 2502607 | CO_MUNICIPIO[linha] == 2511301 |
    CO_MUNICIPIO[linha] == 2510402 | CO_MUNICIPIO[linha] == 2504801)
  banco_pb$gerencia_geografica[linha] <- 7
if (CO_MUNICIPIO[linha] == 2502300 | CO_MUNICIPIO[linha] == 2508109 |
    CO_MUNICIPIO[linha] == 2507408 | CO_MUNICIPIO[linha] == 2509370 |
    CO_MUNICIPIO[linha] == 2512804 | CO_MUNICIPIO[linha] == 2502904 |
    CO_MUNICIPIO[linha] == 2513901 | CO_MUNICIPIO[linha] == 2504306 |
    CO_MUNICIPIO[linha] == 2502003 | CO_MUNICIPIO[linha] == 2514651 |
    CO_MUNICIPIO[linha] == 2502805)
  banco_pb$gerencia_geografica[linha] <- 8
if (CO_MUNICIPIO[linha] == 2512036 | CO_MUNICIPIO[linha] == 2502052 |
    CO_MUNICIPIO[linha] == 2513653 | CO_MUNICIPIO[linha] == 2516904 |
    CO_MUNICIPIO[linha] == 2516805 | CO_MUNICIPIO[linha] == 2512077 |
    CO_MUNICIPIO[linha] == 2500700 | CO_MUNICIPIO[linha] == 2513307 |
    CO_MUNICIPIO[linha] == 2502201 | CO_MUNICIPIO[linha] == 2503704 |
    CO_MUNICIPIO[linha] == 2503308 | CO_MUNICIPIO[linha] == 2514503 |
    CO_MUNICIPIO[linha] == 2504108 | CO_MUNICIPIO[linha] == 2509602 |
    CO_MUNICIPIO[linha] == 2502409)
  banco_pb$gerencia_geografica[linha] <-  9
if (CO_MUNICIPIO[linha] == 2517209 | CO_MUNICIPIO[linha] == 2508406 |
    CO_MUNICIPIO[linha] == 2513208 | CO_MUNICIPIO[linha] == 2513984 |
    CO_MUNICIPIO[linha] == 2516201 | CO_MUNICIPIO[linha] == 2509156 |
    CO_MUNICIPIO[linha] == 2514206 | CO_MUNICIPIO[linha] == 2510006)
  banco_pb$gerencia_geografica[linha] <- 10
if (CO_MUNICIPIO[linha] == 2509008 | CO_MUNICIPIO[linha] == 2514552 |
    CO_MUNICIPIO[linha] == 2512309 | CO_MUNICIPIO[linha] == 2516607 |
    CO_MUNICIPIO[linha] == 2508000 | CO_MUNICIPIO[linha] == 2500106 |
    CO_MUNICIPIO[linha] == 2506707)
  banco_pb$gerencia_geografica[linha] <- 11
if (CO_MUNICIPIO[linha] == 2515807 | CO_MUNICIPIO[linha] == 2512754 |
    CO_MUNICIPIO[linha] == 2506806 | CO_MUNICIPIO[linha] == 2509404 |
    CO_MUNICIPIO[linha] == 2513109 | CO_MUNICIPIO[linha] == 2506905 |
    CO_MUNICIPIO[linha] == 2506400 | CO_MUNICIPIO[linha] == 2503803 |
    CO_MUNICIPIO[linha] == 2514453 | CO_MUNICIPIO[linha] == 2511509 |
    CO_MUNICIPIO[linha] == 2515005 | CO_MUNICIPIO[linha] == 2507903 |
    CO_MUNICIPIO[linha] == 2511202)
  banco_pb$gerencia_geografica[linha] <- 12
```

```
  if (CO_MUNICIPIO[linha] == 2500775 | CO_MUNICIPIO[linha] == 2513968 |
      CO_MUNICIPIO[linha] == 2512101 | CO_MUNICIPIO[linha] == 2510907 |
      CO_MUNICIPIO[linha] == 2505501 | CO_MUNICIPIO[linha] == 2513927 |
      CO_MUNICIPIO[linha] == 2504504 | CO_MUNICIPIO[linha] == 2503753)
    banco_pb$gerencia_geografica[linha] <- 13
  if (CO_MUNICIPIO[linha] == 2507309 | CO_MUNICIPIO[linha] == 2505279 |
      CO_MUNICIPIO[linha] == 2507101 | CO_MUNICIPIO[linha] == 2505238 |
      CO_MUNICIPIO[linha] == 2504033 | CO_MUNICIPIO[linha] == 2508901 |
      CO_MUNICIPIO[linha] == 2512903 | CO_MUNICIPIO[linha] == 2509057 |
      CO_MUNICIPIO[linha] == 2501401 | CO_MUNICIPIO[linha] == 2509305 |
      CO_MUNICIPIO[linha] == 2512721)
    banco_pb$gerencia_geografica[linha] <- 14
}

head(banco_pb)
table(banco_pb$gerencia_geografica)

## Separando as variáveis essenciais

banco_pb <- banco_pb %>% select(NU_ANO_CENSO, CO_IES,
                                TP_CATEGORIA_ADMINISTRATIVA,
                                TP_ORGANIZACAO_ACADEMICA,CO_CURSO,
                                TP_TURNO,TP_GRAU_ACADEMICO,
                                TP_MODALIDADE_ENSINO,TP_NIVEL_ACADEMICO,
                                CO_CINE_ROTULO,ID_ALUNO,CO_ALUNO_CURSO,
                                CO_ALUNO_CURSO_ORIGEM,TP_COR_RACA,
                                TP_SEXO,NU_IDADE,TP_NACIONALIDADE,
                                CO_UF_NASCIMENTO,CO_MUNICIPIO_NASCIMENTO,
                                IN_DEFICIENCIA,TP_SITUACAO,
                                QT_CARGA_HORARIA_TOTAL,
                                QT_CARGA_HORARIA_INTEG,
                                DT_INGRESSO_CURSO,IN_RESERVA_VAGAS,
                                IN_FINANCIAMENTO_ESTUDANTIL,
                                IN_APOIO_SOCIAL,
                                IN_ATIVIDADE_EXTRACURRICULAR,
                                TP_ESCOLA_CONCLUSAO_ENS_MEDIO,
                                TP_SEMESTRE_REFERENCIA,
                                IN_MOBILIDADE_ACADEMICA,IN_MATRICULA,
                                IN_CONCLUINTE,IN_INGRESSO_TOTAL,
```

```
                                gerencia_geografica)

## Considerando a coluna TP_SITUAÇÃO, criação e classificação da
## coluna evasão

#TP_SITUACAO
#2. Cursando
#3. Matrícula trancada
#4. Desvinculado do curso
#5. Transferido para outro curso da mesma IES
#6. Formado
#7. Falecido

#0.Cursando, Transferido para outro curso da mesma IES, Formado
#1.Matrícula trancada, Desvinculado do curso

attach(banco_pb)

for(linha in 1:length(banco_pb$TP_SITUACAO)){
  if (TP_SITUACAO[linha] == 2 | TP_SITUACAO[linha] == 5|
      TP_SITUACAO[linha] == 6) banco_pb$evasao[linha] <-  0
  if (TP_SITUACAO[linha] == 3 |
      TP_SITUACAO[linha] == 4) banco_pb$evasao[linha] <-  1
  if (TP_SITUACAO[linha] == 7) banco_pb$evasao[linha] <-  9
}

## Retirando os falecidos

banco_pb <- banco_pb %>% filter(evasao <= 1)
table(banco_pb$evasao)

str(banco_pb)
attach(banco_pb)

######## REGRESSÃO LOGÍSTICA ########

## Separando novamente as variáveis mais essenciais

banco_modelo <- banco_pb %>% select(evasao, CO_IES,
```

```
                                    TP_CATEGORIA_ADMINISTRATIVA,
                                    TP_ORGANIZACAO_ACADEMICA,CO_CURSO,
                                    TP_TURNO,TP_GRAU_ACADEMICO,
                                    TP_MODALIDADE_ENSINO,
                                    TP_NIVEL_ACADEMICO,TP_COR_RACA,
                                    TP_SEXO,TP_NACIONALIDADE,
                                    CO_UF_NASCIMENTO,
                                    CO_MUNICIPIO_NASCIMENTO,
                                    IN_DEFICIENCIA,IN_RESERVA_VAGAS,
                                    IN_FINANCIAMENTO_ESTUDANTIL,
                                    IN_APOIO_SOCIAL,
                                    IN_ATIVIDADE_EXTRACURRICULAR,
                                    TP_ESCOLA_CONCLUSAO_ENS_MEDIO,
                                    TP_SEMESTRE_REFERENCIA,
                                    IN_MOBILIDADE_ACADEMICA,IN_MATRICULA,
                                    IN_CONCLUINTE,IN_INGRESSO_TOTAL,
                                    gerencia_geografica,
                                    NU_IDADE,QT_CARGA_HORARIA_TOTAL,
                                    QT_CARGA_HORARIA_INTEG)

## Transformando as primeiras 26 variáveis em fatores
banco_modelo <- data.frame(as.data.frame(lapply(banco_modelo[1:26],
                                                as.factor)),
                           banco_modelo[27:29])
str(banco_modelo)

## Divisão no grupo de treinamento e teste

smp_size <- floor(0.80 * nrow(banco_modelo))
set.seed(123)
train_ind <- sample(seq_len(nrow(banco_modelo)), size = smp_size)
train <- banco_modelo[train_ind, ]
test <- banco_modelo[-train_ind, ]


#### MODELOS UNIVARIADOS ####

uni1 <- glm(evasao ~ factor(CO_IES), binomial(link="logit"),
            banco_modelo) # muitas classes significativas
```

```
summary(uni1)

uni2 <- glm(evasao ~ factor(TP_CATEGORIA_ADMINISTRATIVA),
            binomial(link="logit"),
            banco_modelo) # muitas classes significativas
summary(uni2)

uni3 <- glm(evasao ~ factor(TP_ORGANIZACAO_ACADEMICA),
            binomial(link="logit"),
            banco_modelo) # todas as classes significativas
summary(uni3)

uni4 <- glm(evasao ~ factor(CO_CURSO), binomial(link="logit"),
            banco_modelo) # inadequação da variável para o modelo
summary(uni4)

uni5 <- glm(evasao ~ factor(TP_TURNO), binomial(link="logit"),
            banco_modelo) # muitas classes significativas
summary(uni5)

uni6 <- glm(evasao ~ factor(TP_GRAU_ACADEMICO), binomial(link="logit"),
            banco_modelo) # todas as classes significativas
summary(uni6)

uni7 <- glm(evasao ~ factor(TP_MODALIDADE_ENSINO),
            binomial(link="logit"),
            banco_modelo) # todas as classes significativas
summary(uni7)

uni8 <- glm(evasao ~ factor(TP_NIVEL_ACADEMICO), binomial(link="logit"),
            banco_modelo)  # não significativo
summary(uni8)

uni9 <- glm(evasao ~ factor(TP_COR_RACA), binomial(link="logit"),
            banco_modelo) # muitas classes significativas
summary(uni9)

uni10 <- glm(evasao ~ factor(TP_SEXO), binomial(link="logit"),
             banco_modelo) # todas as classes significativas
```

```
summary(uni10)

uni11 <- glm(evasao ~ factor(TP_NACIONALIDADE), binomial(link="logit"),
             banco_modelo) # não significativo
summary(uni11)

uni12 <- glm(evasao ~ factor(CO_UF_NASCIMENTO), binomial(link="logit"),
             banco_modelo) # muitas classes não significativas
summary(uni12)

uni13 <- glm(evasao ~ factor(CO_MUNICIPIO_NASCIMENTO),
             binomial(link="logit"),
             banco_modelo) # inadequação da variável para o modelo
summary(uni13)

uni14 <- glm(evasao ~ factor(IN_DEFICIENCIA), binomial(link="logit"),
             banco_modelo) # todas as classes significativas
summary(uni14)

uni15 <- glm(evasao ~ factor(IN_RESERVA_VAGAS), binomial(link="logit"),
             banco_modelo) # todas as classes significativas
summary(uni15)

uni16 <- glm(evasao ~ factor(IN_FINANCIAMENTO_ESTUDANTIL),
             binomial(link="logit"),
             banco_modelo) # todas as classes significativas
summary(uni16)

uni17 <- glm(evasao ~ factor(IN_APOIO_SOCIAL), binomial(link="logit"),
             banco_modelo) # todas as classes significativas
summary(uni17)

uni18 <- glm(evasao ~ factor(IN_ATIVIDADE_EXTRACURRICULAR),
             binomial(link="logit"),
             banco_modelo) # todas as classes significativas
summary(uni18)

uni19 <- glm(evasao ~ factor(TP_ESCOLA_CONCLUSAO_ENS_MEDIO),
             binomial(link="logit"),
```

```
            banco_modelo) # todas as classes significativas
summary(uni19)

uni20 <- glm(evasao ~ factor(TP_SEMESTRE_REFERENCIA),
            binomial(link="logit"),
            banco_modelo) # todas as classes significativas
summary(uni20)

uni21 <- glm(evasao ~ factor(IN_MOBILIDADE_ACADEMICA),
            binomial(link="logit"),
            banco_modelo) # não significativo
summary(uni21)

uni22 <- glm(evasao ~ factor(IN_MATRICULA), binomial(link="logit"),
            banco_modelo) # não significativo
summary(uni22)

uni23 <- glm(evasao ~ factor(IN_CONCLUINTE), binomial(link="logit"),
            banco_modelo) # não significativo
summary(uni23)

uni24 <- glm(evasao ~ factor(IN_INGRESSO_TOTAL), binomial(link="logit"),
            banco_modelo)  # todas as classes significativas
summary(uni24)

uni25 <- glm(evasao ~ factor(gerencia_geografica),
            binomial(link="logit"),
            banco_modelo) # muitas classes significativas
summary(uni25)

uni26 <- glm(evasao ~ NU_IDADE, binomial(link="logit"),
            banco_modelo) # significativo
summary(uni26)

uni27 <- glm(evasao ~ QT_CARGA_HORARIA_TOTAL, binomial(link="logit"),
            banco_modelo) # significativo
summary(uni27)

uni28 <- glm(evasao ~ QT_CARGA_HORARIA_INTEG, binomial(link="logit"),
```

```
            banco_modelo) # significativo
summary(uni28)


## Vamos, então, deixar no banco apenas as variáveis que tiveram muitas
## ou todas as classes significativas

banco_modelo <- banco_modelo %>% select(TP_CATEGORIA_ADMINISTRATIVA,
                                        TP_ORGANIZACAO_ACADEMICA,
                                        TP_TURNO,TP_GRAU_ACADEMICO,
                                        TP_COR_RACA,TP_SEXO,
                                        IN_DEFICIENCIA,IN_RESERVA_VAGAS,
                                        IN_FINANCIAMENTO_ESTUDANTIL,
                                        IN_APOIO_SOCIAL,
                                        IN_ATIVIDADE_EXTRACURRICULAR,
                                        TP_ESCOLA_CONCLUSAO_ENS_MEDIO,
                                        gerencia_geografica,
                                        IN_INGRESSO_TOTAL,NU_IDADE,
                                        QT_CARGA_HORARIA_TOTAL,
                                        QT_CARGA_HORARIA_INTEG,evasao)

## Salvando o banco para importar em Python
write.table(banco_modelo, "banco_modelo.txt")
```

C.2 Linguagem Python

```
## Bibliotecas

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import pingouin as pg
import seaborn as sns
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectKBest, chi2
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
```

```
import statsmodels.api as sm
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import plot_roc_curve
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.ensemble import RandomForestClassifier

## Função para importação do banco
def pipeline_Alunos_ES(arquivo):
    # Colunas a serem lidas no arquivo
    features = [
        'CO_UF_NASCIMENTO', 'CO_MUNICIPIO_NASCIMENTO',
        'IN_MOBILIDADE_ACADEMICA','TP_CATEGORIA_ADMINISTRATIVA',
        'TP_ORGANIZACAO_ACADEMICA', 'TP_TURNO', 'TP_GRAU_ACADEMICO',
        'TP_COR_RACA','TP_SEXO','IN_RESERVA_VAGAS',
        'IN_FINANCIAMENTO_ESTUDANTIL','IN_APOIO_SOCIAL',
        'IN_ATIVIDADE_EXTRACURRICULAR','TP_ESCOLA_CONCLUSAO_ENS_MEDIO',
        'gerencia_geografica','IN_INGRESSO_TOTAL','NU_IDADE',
        'QT_CARGA_HORARIA_INTEG','evasao'
    ]
    df = pd.read_table('banco_modelo.txt',sep=" ",usecols = features)

    return df

dados = pipeline_Alunos_ES('banco_modelo.csv')

dados.head()

## Pré-processsamento dos dados

dados.isnull().sum()

# variáveis retiradas: 'CO_UF_NASCIMENTO', 'CO_MUNICIPIO_NASCIMENTO',
'IN_MOBILIDADE_ACADEMICA', 'IN_FINANCIAMENTO_ESTUDANTIL'

# variáveis tratadas: 'TP_TURNO', 'TP_GRAU_ACADEMICO'

dados = dados.drop(columns=['CO_UF_NASCIMENTO'])
dados= dados.drop(columns=['CO_MUNICIPIO_NASCIMENTO'])
dados = dados.drop(columns=[ 'IN_MOBILIDADE_ACADEMICA'])
```

```
dados = dados.drop(columns=['IN_FINANCIAMENTO_ESTUDANTIL'])

dados['TP_TURNO'].fillna(9, inplace = True)
dados['TP_GRAU_ACADEMICO'].fillna(9, inplace = True)
dados.head()

## Análise descritiva

dados.dtypes

dados.isnull().sum()

## Quantidade de observações
len(dados['TP_CATEGORIA_ADMINISTRATIVA'])

dados.TP_CATEGORIA_ADMINISTRATIVA.value_counts()

## Conferindo as classes que ficaram nas variáveis tratadas
dados.TP_TURNO.value_counts()

a = dados['TP_TURNO'].value_counts(normalize=True).map('{:.1%}'.format)
display(a)

dados.TP_GRAU_ACADEMICO.value_counts()

b = 'TP_GRAU_ACADEMICO'
c = dados[b].value_counts(normalize=True).map('{:.1%}'.format)
display(c)

## Análise Descritiva (realizada entre as fases do pré-processamento)

dados.QT_CARGA_HORARIA_INTEG.plot(kind='hist', bins=20)
plt.xlabel('QT_CARGA_HORARIA_INTEG')
plt.ylabel('Frequência')
plt.show()

dados.NU_IDADE.plot(kind='hist', bins=20)
plt.xlabel('NU_IDADE')
plt.ylabel('Frequência')
```

```
plt.show()

dados_copia = dados.copy()
dados_copia.head()

d = 'TP_CATEGORIA_ADMINISTRATIVA'
dados_copia[d] = dados_copia[d].map({1:'Pública Federal',
                                     2:'Pública Estadual',
                                     4:'Privada com fins lucrativos',
                                     5:'Privada sem fins lucrativos'})
e = 'TP_ORGANIZACAO_ACADEMICA'
dados_copia[e] = dados_copia[e].map({1:'Universidade',
                                     2:'Centro Universitário',
                                     3:'Faculdade',
                                     4:'IF'})
dados_copia['TP_TURNO'] = dados_copia['TP_TURNO'].map({1:'Matutino',
                                                       2:'Vespertino',
                                                       3:'Noturno',
                                                       4:'Integral',
                                                       9:'NA'})
f = 'TP_GRAU_ACADEMICO'
dados_copia[f] = dados_copia[f].map({1:'Bacharelado', 2:'Licenciatura',
                                     3:'Tecnológico', 9:'NA'})
g = 'TP_COR_RACA'
dados_copia[g] = dados_copia[g].map({0:'Aluno não quis declarar',
                                     1:'Branca', 2:'Preta',3:'Parda',
                                     4:'Amarela',5:'Indígena',9:'NA'})
dados_copia['TP_SEXO'] = dados_copia['TP_SEXO'].map({1:'Feminino',
                                                     2:'Masculino'})
h = 'IN_RESERVA_VAGAS'
dados_copia[h] = dados_copia[h].map({0:'Não',1:'Sim'})
i = 'IN_APOIO_SOCIAL'
dados_copia[i] = dados_copia[i].map({0:'Não',1:'Sim'})
j = 'IN_ATIVIDADE_EXTRACURRICULAR'
dados_copia[j] = dados_copia[j].map({0:'Não',1:'Sim'})
k = 'TP_ESCOLA_CONCLUSAO_ENS_MEDIO'
dados_copia[k] = dados_copia[k].map({1:'Pública',2:'Privada',9:'NA'})
l = 'IN_INGRESSO_TOTAL'
dados_copia[l] = dados_copia[l].map({0:'Não',1:'Sim'})
```

```
m = 'evasao'
dados_copia[m] = dados_copia[m].map({0:'Não evadidos',1:'Evadidos'})

dados_copia.describe()

## Desbalanceamento dos dados

dados_copia.evasao.value_counts()

dados_copia.evasao.value_counts().plot(kind='bar')
plt.xticks(rotation=360)

## TP_CATEGORIA_ADMINISTRATIVA
dados_copia[d].value_counts()/dados_copia.shape[0]*100

sns.countplot(dados_copia[d])
plt.xlabel('')
plt.ylabel("Frequência")
plt.xticks(rotation=90)
plt.show()

## TP_ORGANIZACAO_ACADEMICA
dados_copia[e].value_counts()/dados_copia.shape[0]*100

sns.countplot(dados_copia[e])
plt.xlabel('')
plt.ylabel("Frequência")
plt.xticks(rotation=90)
plt.show()

## TP_TURNO
dados_copia['TP_TURNO'].value_counts()/dados_copia.shape[0]*100

sns.countplot(dados_copia['TP_TURNO'])
plt.xlabel('')
plt.ylabel("Frequência")
plt.show()

## TP_GRAU_ACADEMICO
```

```
dados_copia[f].value_counts()/dados_copia.shape[0]*100

sns.countplot(dados_copia[f])
plt.xlabel('')
plt.ylabel("Frequência")
plt.show()

## Funções para os gráficos bivariados

def with_hue(plot, feature, Number_of_categories, hue_categories):
    a = [p.get_height() for p in plot.patches]
    patch = [p for p in plot.patches]
    for i in range(Number_of_categories):
        total = feature.value_counts().values[i]
        for j in range(hue_categories):
            percentage = '{:.1f}%'.format(100 *
            a[(j*Number_of_categories +
                                                i)]/total)
            x = patch[(j*Number_of_categories + i)].get_x() +
            patch[(j*Number_of_categories + i)].get_width() / 2 - 0.15
            y = patch[(j*Number_of_categories + i)].get_y() +
            patch[(j*Number_of_categories + i)].get_height()
            ax.annotate(percentage, (x, y), size = 12)
    plt.show()

def without_hue(plot, feature):
    total = len(feature)
    for p in ax.patches:
        percentage = '{:.1f}%'.format(100 * p.get_height()/total)
        x = p.get_x() + p.get_width() / 2 - 0.05
        y = p.get_y() + p.get_height()
        ax.annotate(percentage, (x, y), size = 12)
    plt.show()

## TP_ESCOLA_CONCLUSAO_ENS_MEDIO
ax = sns.countplot(dados_copia[k],hue=dados_copia['evasao'], dodge=True)
plt.legend(loc='best')
plt.xlabel('')
plt.ylabel('')
```

```
with_hue(ax,dados_copia[k],3,2)

## IN_RESERVA_VAGAS
ax = sns.countplot(dados_copia[h],hue=dados_copia['evasao'], dodge=True)
plt.legend(loc='best')
plt.xlabel('')
plt.ylabel('')
with_hue(ax,dados_copia[h],2,2)

## IN_APOIO_SOCIAL
ax = sns.countplot(dados_copia[i],hue=dados_copia['evasao'], dodge=True)
plt.legend(loc='best')
plt.xlabel('')
plt.ylabel('')
with_hue(ax,dados_copia[i],2,2)

## IN_ATIVIDADE_EXTRACURRICULAR
ax = sns.countplot(dados_copia[j],hue=dados_copia['evasao'], dodge=True)
plt.legend(loc='best')
plt.xlabel('')
plt.ylabel('')
with_hue(ax,dados_copia[j],2,2)

## IN_INGRESSO_TOTAL
ax = sns.countplot(dados_copia[l],hue=dados_copia['evasao'], dodge=True)
plt.xlabel('')
plt.ylabel('')
plt.legend(loc='best')
with_hue(ax,dados_copia[l],2,2)

## Up-sample (smote - dados sintéticos sem ser duplicados)
## para balanceamento dos dados e separação do
## dataset em treinamento e teste em 70%/30%

X = dados.loc[:, dados.columns != 'evasao']
y = dados.loc[:, dados.columns == 'evasao']

os = SMOTE(random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y,
```

```
                                                    test_size=0.3,
                                                    random_state=42)
columns = X_train.columns
os_dados_X,os_dados_y=os.fit_resample(X_train, y_train)
os_dados_X = pd.DataFrame(data=os_dados_X,columns=columns )
os_dados_y= pd.DataFrame(data=os_dados_y,columns=['evasao'])
print('O tamanho da base balanceada é ',
      len(os_dados_X))
print('A quantidade de registros sintéticos foi',
      len(os_dados_y[os_dados_y['evasao']==0]))
print('A quantidade de registros originais foi',
      len(os_dados_y[os_dados_y['evasao']==1]))
print('A proporção de registros sintéticos foi ',
      len(os_dados_y[os_dados_y['evasao']==0])/len(os_dados_X))
print('A proporção de registros originais foi ',
      len(os_dados_y[os_dados_y['evasao']==1])/len(os_dados_X))

columns

## ATENÇÃO: Alteração apenas nos dados de treinamento
y_train['evasao'].value_counts()

y_train.evasao.value_counts().plot(kind='bar')

y_train['evasao'].value_counts(normalize=True)

os_dados_y['evasao'].value_counts()

## Seleção de variáveis

test = SelectKBest(chi2, k=10)
fit = test.fit(os_dados_X,os_dados_y.values.ravel())
features = fit.transform(os_dados_X)
print(features)

fit.get_support(indices=True)

dados[['TP_CATEGORIA_ADMINISTRATIVA',
       'TP_TURNO','TP_GRAU_ACADEMICO',
```

```
       'gerencia_geografica']] = dados[['TP_CATEGORIA_ADMINISTRATIVA',
       'TP_TURNO','TP_GRAU_ACADEMICO',
       'gerencia_geografica']].astype(object)
dados.dtypes
dados = pd.get_dummies(dados)
dados.head().T

dados[['TP_CATEGORIA_ADMINISTRATIVA_1', 'TP_CATEGORIA_ADMINISTRATIVA_2',
       'TP_CATEGORIA_ADMINISTRATIVA_4','TP_CATEGORIA_ADMINISTRATIVA_5',
       'TP_TURNO_1.0','TP_TURNO_2.0', 'TP_TURNO_3.0','TP_TURNO_4.0',
       'TP_TURNO_9.0','TP_GRAU_ACADEMICO_1.0', 'TP_GRAU_ACADEMICO_2.0',
       'TP_GRAU_ACADEMICO_3.0','TP_GRAU_ACADEMICO_9.0',
       'gerencia_geografica_1','gerencia_geografica_2',
       'gerencia_geografica_3','gerencia_geografica_5',
       'gerencia_geografica_6','gerencia_geografica_9',
       'gerencia_geografica_10'
     ]] = dados[['TP_CATEGORIA_ADMINISTRATIVA_1',
                 'TP_CATEGORIA_ADMINISTRATIVA_2',
                 'TP_CATEGORIA_ADMINISTRATIVA_4',
                 'TP_CATEGORIA_ADMINISTRATIVA_5',
                 'TP_TURNO_1.0','TP_TURNO_2.0', 'TP_TURNO_3.0',
                 'TP_TURNO_4.0','TP_TURNO_9.0','TP_GRAU_ACADEMICO_1.0',
                 'TP_GRAU_ACADEMICO_2.0','TP_GRAU_ACADEMICO_3.0',
                 'TP_GRAU_ACADEMICO_9.0',
                 'gerencia_geografica_1', 'gerencia_geografica_2',
                 'gerencia_geografica_3','gerencia_geografica_5',
                 'gerencia_geografica_6','gerencia_geografica_9',
                 'gerencia_geografica_10'
     ]].astype(float)
dados.dtypes

## Após o balanceamento do conjunto treinamento e a seleção de
## variáveis, ## vamos retirar tais variáveis do banco completo e
## refeito todo balanceamento e modelagem dos dados

cols=['TP_CATEGORIA_ADMINISTRATIVA_2', 'TP_CATEGORIA_ADMINISTRATIVA_4',
      'TP_CATEGORIA_ADMINISTRATIVA_5','TP_TURNO_2.0', 'TP_TURNO_3.0',
      'TP_TURNO_4.0','TP_TURNO_9.0','TP_GRAU_ACADEMICO_2.0',
      'TP_GRAU_ACADEMICO_3.0','TP_GRAU_ACADEMICO_9.0','IN_RESERVA_VAGAS',
```

```
      'IN_APOIO_SOCIAL', 'IN_ATIVIDADE_EXTRACURRICULAR',
      'IN_INGRESSO_TOTAL','gerencia_geografica_2',
      'gerencia_geografica_3','gerencia_geografica_5',
      'gerencia_geografica_6','gerencia_geografica_9',
      'gerencia_geografica_10','NU_IDADE', 'QT_CARGA_HORARIA_INTEG'
     ]
X=dados[cols]
y=dados['evasao']


os = SMOTE(random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
                                                    random_state=42)
columns = X_train.columns
os_dados_X,os_dados_y=os.fit_resample(X_train, y_train)
os_dados_X = pd.DataFrame(data=os_dados_X,columns=columns )
os_dados_y= pd.DataFrame(data=os_dados_y,columns=['evasao'])

y_train.value_counts()

######## REGRESSÃO LOGÍSTICA ########

logreg = LogisticRegression(penalty='none', solver='newton-cg')
logreg.fit(os_dados_X, os_dados_y.values.ravel())
y_pred = logreg.predict(X_test)
print(f'Acurácia: {metrics.accuracy_score(y_test, y_pred):.2f}')
metrics.plot_confusion_matrix(logreg, X_test, y_test,
                              cmap=plt.cm.Blues, normalize='true')
plt.xlabel('Categoria prevista')
plt.ylabel('Categoria real')
plt.show()

print (metrics.classification_report(y_test, y_pred))

#CURVA ROC

from sklearn.metrics import plot_roc_curve

roc = plot_roc_curve(
```

```
    logreg, X_test, y_test,
    name='Regressão Logística'
)

roc.ax_.set_ylabel('Taxa de verdadeiros positivos')
roc.ax_.set_xlabel('Taxa de falsos positivos')

## Odds Rattio

logit_model=sm.Logit(os_dados_y.values.ravel(),os_dados_X)
result=logit_model.fit()
print(np.exp(result.params))

print((np.exp(result.params[1:]) - 1) * 100)

## Para melhor análise
(np.exp(result.params[1:]) - 1) * 100

######## KNN ########

dados = pipeline_Alunos_ES('banco_modelo.csv')

dados = dados.drop(columns=['CO_UF_NASCIMENTO'])
dados = dados.drop(columns=['CO_MUNICIPIO_NASCIMENTO'])
dados = dados.drop(columns=[ 'IN_MOBILIDADE_ACADEMICA'])
dados = dados.drop(columns=['IN_FINANCIAMENTO_ESTUDANTIL'])

dados['TP_TURNO'].fillna(9, inplace = True)
dados['TP_GRAU_ACADEMICO'].fillna(9, inplace = True)

dados.head()

cols=['TP_CATEGORIA_ADMINISTRATIVA', 'TP_TURNO','TP_GRAU_ACADEMICO',
      'IN_RESERVA_VAGAS', 'IN_APOIO_SOCIAL',
      'IN_ATIVIDADE_EXTRACURRICULAR', 'IN_INGRESSO_TOTAL',
      'gerencia_geografica', 'NU_IDADE', 'QT_CARGA_HORARIA_INTEG'
     ]
X=dados[cols]
y=dados['evasao']
```

```
os = SMOTE(random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.3,
                                                    random_state=42)
columns = X_train.columns
os_dados_X,os_dados_y=os.fit_resample(X_train, y_train)
os_dados_X = pd.DataFrame(data=os_dados_X,columns=columns )
os_dados_y= pd.DataFrame(data=os_dados_y,columns=['evasao'])

#### K = 2 ####

knn = KNeighborsClassifier(n_neighbors=2)
knn.fit(os_dados_X, os_dados_y.values.ravel())
y_pred = knn.predict(X_test)

## Métricas de desempenho

print(f'Acurácia: {metrics.accuracy_score(y_test, y_pred):.2f}')
metrics.plot_confusion_matrix(knn, X_test, y_test, cmap=plt.cm.Blues,
                              normalize='true')
plt.xlabel('Categoria prevista')
plt.ylabel('Categoria real')
plt.show()

## Curva ROC

roc = plot_roc_curve(
    knn, X_test, y_test,
    name='KNN'
)

roc.ax_.set_ylabel('Taxa de verdadeiros positivos')
roc.ax_.set_xlabel('Taxa de falsos positivos')

#### K = 5 ####

knn = KNeighborsClassifier(n_neighbors=5)
```

```
knn.fit(os_dados_X, os_dados_y.values.ravel())
y_pred = knn.predict(X_test)

## Métricas de desempenho
print(f'Acurácia: {metrics.accuracy_score(y_test, y_pred):.2f}')
metrics.plot_confusion_matrix(knn, X_test, y_test, cmap=plt.cm.Blues,
                              normalize='true')
plt.xlabel('Categoria prevista')
plt.ylabel('Categoria real')
plt.show()

## Curva ROC

roc = plot_roc_curve(
    knn, X_test, y_test,
    name='KNN'
)

roc.ax_.set_ylabel('Taxa de verdadeiros positivos')
roc.ax_.set_xlabel('Taxa de falsos positivos')

#### K = 9 ####

knn = KNeighborsClassifier(n_neighbors=9)
knn.fit(os_dados_X, os_dados_y.values.ravel())
y_pred = knn.predict(X_test)

## Métricas de desempenho
print(f'Acurácia: {metrics.accuracy_score(y_test, y_pred):.2f}')
metrics.plot_confusion_matrix(knn, X_test, y_test, cmap=plt.cm.Blues,
                              normalize='true')
plt.xlabel('Categoria prevista')
plt.ylabel('Categoria real')
plt.show()

print (metrics.classification_report(y_test, y_pred))

## Curva ROC
```

```
roc = plot_roc_curve(
    knn, X_test, y_test,
    name='KNN'
)

roc.ax_.set_ylabel('Taxa de verdadeiros positivos')
roc.ax_.set_xlabel('Taxa de falsos positivos')

######## ÁRVORE DE DECISÃO ########

dados = pipeline_Alunos_ES('banco_modelo.csv')

dados = dados.drop(columns=['CO_UF_NASCIMENTO'])
dados = dados.drop(columns=['CO_MUNICIPIO_NASCIMENTO'])
dados = dados.drop(columns=[ 'IN_MOBILIDADE_ACADEMICA'])
dados = dados.drop(columns=['IN_FINANCIAMENTO_ESTUDANTIL'])

dados['TP_TURNO'].fillna(9, inplace = True)
dados['TP_GRAU_ACADEMICO'].fillna(9, inplace = True)

dados.head()

cols=['TP_CATEGORIA_ADMINISTRATIVA','TP_TURNO','TP_GRAU_ACADEMICO',
      'IN_RESERVA_VAGAS', 'IN_APOIO_SOCIAL',
      'IN_ATIVIDADE_EXTRACURRICULAR','IN_INGRESSO_TOTAL',
      'gerencia_geografica', 'NU_IDADE', 'QT_CARGA_HORARIA_INTEG'
     ]
X=dados[cols]
y=dados['evasao']



os = SMOTE(random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.3,
                                                    random_state=42)
columns = X_train.columns
os_dados_X,os_dados_y=os.fit_resample(X_train, y_train)
os_dados_X = pd.DataFrame(data=os_dados_X,columns=columns )
os_dados_y= pd.DataFrame(data=os_dados_y,columns=['evasao'])
```

```
tree = DecisionTreeClassifier(class_weight="balanced", max_depth=7,
                              random_state=42)
tree.fit(os_dados_X, os_dados_y.values.ravel())
y_pred = tree.predict(X_test)

## Métricas de desempenho
print(f'Acurácia: {metrics.accuracy_score(y_test, y_pred):.2f}')
metrics.plot_confusion_matrix(tree, X_test, y_test, cmap=plt.cm.Blues,
                              normalize='true')
plt.xlabel('Categoria prevista')
plt.ylabel('Categoria real')
plt.show()

print (metrics.classification_report(y_test, y_pred))

## Curva ROC

roc = plot_roc_curve(
    tree, X_test, y_test,
    name='Árvore de decisão'
)

roc.ax_.set_ylabel('Taxa de verdadeiros positivos')
roc.ax_.set_xlabel('Taxa de falsos positivos')

######## RANDOM FOREST ########

dados = pipeline_Alunos_ES('banco_modelo.csv')

dados = dados.drop(columns=['CO_UF_NASCIMENTO']).
dados = dados.drop(columns=['CO_MUNICIPIO_NASCIMENTO'])
dados = dados.drop(columns=[ 'IN_MOBILIDADE_ACADEMICA'])
dados = dados.drop(columns=['IN_FINANCIAMENTO_ESTUDANTIL'])

dados['TP_TURNO'].fillna(9, inplace = True)
dados['TP_GRAU_ACADEMICO'].fillna(9, inplace = True)

dados.head()
```

```
cols=['TP_CATEGORIA_ADMINISTRATIVA','TP_TURNO','TP_GRAU_ACADEMICO',
      'IN_RESERVA_VAGAS', 'IN_APOIO_SOCIAL',
      'IN_ATIVIDADE_EXTRACURRICULAR','IN_INGRESSO_TOTAL',
      'gerencia_geografica', 'NU_IDADE', 'QT_CARGA_HORARIA_INTEG'
     ]
X=dados[cols]
y=dados['evasao']


os = SMOTE(random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.3,
                                                    random_state=42)
columns = X_train.columns
os_dados_X,os_dados_y=os.fit_resample(X_train, y_train)
os_dados_X = pd.DataFrame(data=os_dados_X,columns=columns )
os_dados_y= pd.DataFrame(data=os_dados_y,columns=['evasao'])

rf = RandomForestClassifier(class_weight="balanced",
                            n_estimators=300, random_state=42)
rf.fit(os_dados_X, os_dados_y.values.ravel())
y_pred = rf.predict(X_test)

# Métricas de desempenho
print(f'Acurácia: {metrics.accuracy_score(y_test, y_pred):.2f}')
metrics.plot_confusion_matrix(rf, X_test, y_test, cmap=plt.cm.Blues,
                              normalize='true')
plt.xlabel('Categoria prevista')
plt.ylabel('Categoria real')
plt.show()

print (metrics.classification_report(y_test, y_pred))

## Curva ROC

roc = plot_roc_curve(
    rf, X_test, y_test,
    name='Random Forest'
```

```
)

roc.ax_.set_ylabel('Taxa de verdadeiros positivos')
roc.ax_.set_xlabel('Taxa de falsos positivos')
```