Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Fábio Junior Rodrigues Moreira

Aprendizagem de máquina na predição da evasão no ensino superior

Curitiba
2020

Fábio Junior Rodrigues Moreira

Aprendizagem de máquina na predição da evasão no ensino superior

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Marco A. Zanata Alves

Curitiba
2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Aprendizagem de máquina na predição da evasão no ensino superior

Fábio Junior Rodrigues Moreira[1]
Marco Antonio Zanata Alves[2]

Resumo

No Brasil a taxa média de evasão escolar do ensino superior tem se mantido em 21% nos último 10 anos. Os problemas gerados pela evasão são diversos, como desperdício de recursos por parte do aluno e também da universidade. Nesse cenário, mesmo coordenadores de curso com experiência não conseguem prever o abandono por parte dos alunos. Essa predição é bastante complexa de ser feito manualmente devido a grande quantidade de alunos geridos pelo mesmo coordenador. Nesse trabalho, foi realizado uma análise de acompanhamento da trajetória de alunos do curso de graduação em Ciência da Computação da UFPR compreendendo os períodos de 2011 á 2019 utilizando duas técnicas de aprendizado de máquina, regressão logística e árvores de decisão, para prever a evasão. Com as técnicas aplicadas podemos prever a evasão de alunos de ciência da computação de uma base de dados real, onde atingimos até 86% de precisão com os algoritmos de aprendizado de máquina avaliados.
Palavras-chave:Evasão escolar, Aprendizado de máquina, Regressão logística, Árvores de Decisão.

Abstract

In Brazil, the average dropout rate in higher education has remained at 21 % in the last 10 years. The problems generated by evasion are diverse, such as wasted resources by the student and the university. In this scenario, even experienced course coordinators are unable to predict students' dropout. This prediction is quite complex to be done manually due to the large number of students managed by the same coordinator. In this work, we performed a follow-up analysis of students' trajectory of the undergraduate course in Computer Science at UFPR, comprising the periods from 2011 to 2019, using two machine learning techniques, logistic regression and decision trees, to predict dropout. The applied techniques show that we can predict the evasion of computer science students from a real database, where we reach up to 86 % accuracy with the evaluated machine learning algorithms.
***Keywrods**: School dropout, Machine learning, Logistic regression, Decision Trees.*

[1]Aluno do programa de Especialização em Data Science & Big Data, fabio_jrmoreira@hotmail.com.
[2]Professor do Departamento de Informática - DINF/UFPR., mazalves@inf.ufpr.br.

1 Introdução

Compreende-se por evasão escolar a condição do aluno que reprovou ou abandonou os estudos e não realizou a matrícula no ano seguinte. A evasão escolar têm sido um problema frequente nas mais diversas modalidades de ensino desde a educação básica até o ensino superior, isso têm-se tornado tema de inúmeros estudos, pois conhecer o perfil do aluno que evade pode auxiliar a encontrar uma solução para esse processo e a criar ações preventivas que minimizem a evasão.

Os gestores educacionais precisam realizar tomadas de decisões precisas no combate à evasão e nesse contexto o apoio da algoritmos de *machine learning* representa uma solução interessante para a criação de modelos analíticos para acompanhar, identificar e predizer estudantes com potencial de evadir-se do curso, proporcionando medidas preventivas para minimizar o problema.

Assim, esse presente trabalho tem por objetivo auxiliar as instituições de ensino superior, através do desempenho acadêmico do aluno ao longo dos anos, prever e aplicar ações paliativas na prevenção da evasão em conjunto com o uso de ferramentas de machine learning que irá permitir uma gestão mais eficiente através dos resultados obtidos, pois sabemos que por falta de motivação, vocação ou orientação os estudantes podem vir a desistir do curso entretanto podem existir outros motivos que serão demonstrados por esse trabalho. Nesse estudo foi utilizado dois modelos: regressão logística e árvore de decisão.

Esse trabalho está estruturado da seguinte forma. A seção 2 contextualiza machine learning. A seção 3 contempla alguns trabalhos relacionados a evasão no ensino superior. A seção 4 contempla a proposta e metodologia de trabalho. A seção 5 contempla os resultados obtidos e por fim a seção 6 contempla a conclusão e sugestões a trabalhos futuros.

2 Contexto de Machine Learning

Contida na área de inteligência artificial, machine learning teve seu início nos anos 40, contudo foi realmente evidenciado seu uso nos anos 90 com a aplicação dessa técnica em filtros de spam mas machine learning não se resume em um simples filtro de e-mails, ela esta desde

análise climática realizando previsões á reconhecimento facial em vias públicas.

E devido ao grande volume de dados gerados diariamente pelas organizações a serem analisados, percebeu-se que os dados tornaram-se um produto valorizado e para trabalhar com esse volume, machine learning possui capacidade de aprendizado e desempenho preditivo, automatizando modelos analíticos a partir de um método de análise identificando padrões e tomadas de decisões e quanto maior o volume de dados for inserido melhor será o desempenho do modelo.

Nesse contexto machine learning possui três categorias:

Aprendizado não supervisionado: Utiliza algoritmos que não necessita de intervenção humana e rotulagem.
Aprendizado semi-supervisionado: Utiliza uma mescla do aprendizado não supervisionado com aprendizado supervisionado tendo maior utilização com grande volume de dados.
Aprendizado supervisionado: Utiliza algoritmo rotulados para treinamento dos modelos preditivos.

Nesse trabalho foi utilizado o algoritmo supervisionado pois possui classificação e regressão para estimar valores desconhecidos de variáveis a partir de valores conhecidos de outras variáveis.

- A classificação tem como objetivo analisar o dado de entrada e aplicar um rótulo a ela, seu uso é basicamente quando a previsão é simples, se o aluno evadiu ou não.
- A regressão tem como objetivo determinar como o rótulo de dados será modificado à medida que os valores previstos divergem de zero e um.

Nesse estudo foi utilizado dois modelos: regressão logística e árvore de decisão. Tendo em vista as informações coletadas os melhores algoritmos escolhidos foram a regressão logística e árvore de decisão por se tratarem de algoritmos de aprendizado supervisionado.

2.1 Regressão Logística

A regressão logística e comumente utilizada para estimar a probabilidade de uma instância pertencer a uma determinada classe (por exemplo, qual é a probabilidade desse e-mail ser um spam?). Se a probabilidade estimada for maior que 50%, então o modelo prevê que a instância pertence a essa classe (isto é, pertence á classe negativa, rotulada “0”). Isso o transforma em um classificador binário[3].

A regressão logística foi utilizada para classificar os alunos que abandonaram e os alunos que se formaram no curso de graduação em Ciência da Computação contendo todo seu histórico de notas, frequência e desempenho por disciplina, na matriz curricular de 2011. Tendo esses dados para preparação foi utilizado a linguagem de programação Python e aplicado as seguintes bibliotecas Pandas, Numpy, Scikit-Learn e Matplotlib.

2.2 Árvores de Decisão

Uma árvore de decisão usa uma estrutura de árvore para representar um número de possíveis caminhos de decisão e um resultado para cada caminho [6].

Permite que um indivíduo ou organização compare possíveis ações com base em probabilidades, podem ser utilizadas para mapear um algoritmo que prevê a melhor escolha, geralmente a árvore começa com um nó em que o resultado se distribui em nós subsequentes e por consequência ramificam-se em novas possibilidades com nós adicionais. Categorizando-se em nós de probabilidade, nós de decisão e nós de término.

3 Trabalhos Relacionados

Nessa seção apresentamos um levantamento de pesquisas bibliográficas para conhecer os modelos relacionados para prever o perfil do aluno que irá abandonar os estudos.

Focado em prever a taxa de aprovação nos dois primeiros anos do curso de graduação, Karamouzis [7] fazendo o uso de uma rede neural Multiplayer Perceptron (MLP) e fazendo uso de dados demográficos e acadêmicos criaram um conjunto de dados, o estudo mostrou uma média de acerto de 72%.

Jadric [5] analisando estudantes no início do curso de graduação em Economia considerando apenas os primeiros 2 anos compararam os seguintes modelos: árvore de decisão e regressão logística e redes neurais e fazendo uso da metodologia SEMMA (Sampling, Exploring, Modifying, Modelling and Assessment) os autores optaram por utilizar variáveis relacionadas à inscrição do candidato e atributos referentes ao processo de estudos, os resultados mostraram que o modelo de rede neural evidenciou um melhor desempenho apontando que 36% dos estudantes poderão evadir.

Analisando apenas informações acadêmicas, Manhães [8] fazendo uso de algoritmos de classificação e dados de seis cursos da Universidade Federal do Rio de Janeiro conseguiu uma taxa de pelo 87% para cada curso em uma taxa de verdadeiro positivo que varia 66,8%.

Vendo uma alta taxa de evasão no curso de Engenharia, Dekker [1] elaborou um estudo experimental na busca de um modelo preditivo para predizer a evasão dos calouros fazendo uso de arvore de decisão e um classificador bayesiano os resultados mostraram uma precisão de 68% quando analisado somente os dados antes do ingresso na universidade, ao verificar o conjunto completo de dados a precisão fica entre 75% e 80% na identificação de uma possível evasão.

Em sua pesquisa Fu [2] aplicou um questionário para criar um data-set com as características da personalidade dos estudantes aplicando a regressão de vetores de suporte para construir um modelo de desempenho, o modelo obteve uma precisão próxima de 80% na previsão do desempenho acadêmico dos alunos.

Utilizando algoritmo de classificação popular e propondo um algoritmo genético que faz uso de cost-

sensitive learning e técnicas de balanceamento Márquez-Vera [9] explorou a evasão no ensino médio nas cidades das províncias mexicanas obteve uma taxa de verdadeiros positivos e verdadeiro negativos entre 93,4% e 88,3% respectivamente.

Tendo como base o curso de Zootecnia e tendo as informações de desempenho acadêmico dos alunos Hoffmann [4], aplicando técnicas de mineração de dados obteve um resultado de 98% de acurácia na previsão e mais de 70% na previsão de alunos que abandonaram o curso.

4 Proposta e Metodologia

Nessa seção iremos apresentar a metodologia utilizada nesse estudo para alcançar o objetivo proposto, os procedimentos para a coleta e análise dos dados e as ferramentas e metodologias utilizadas.

4.1 Metodologia CRISP

Para esse trabalho foi implementado a metodologia CRISP DM, essa metodologia criada a mais de 20 anos visa auxiliar na estruturação dos projetos de mineração de dados (CHAPMAN et al., 2000). A metodologia tem por base um processo cíclico e interativo. CRISP DM está dividido em 6 fases que serão descritas a seguir:

1) Entendimento do problema: Essa fase visa a entender de fato qual o problema a ser resolvido, buscando todos os fatos e quais objetivos serão alcançados em relação ao trabalho.

2) Compreensão dos dados: Essa fase visa a organizar e documentar os dados que se tem acesso. É nsssa fase que começa de fato o trabalho de mineração de dados, pois e nesse ponto que o analista reconhece os dados importantes para resolução do problema.

3) Preparação dos dados: Após a identificação dos dados, documentados e analisados esta fase tem por objetivo preparar os atributos para que seja criado um conjunto de dados adequados para as ferramentas de mineração de dados e durante esse processamento identificar e excluir as variáveis que não vem agregar no processamento.

4) Modelagem: Essa fase visa selecionar e aplicar os algoritmos de machine learning buscando o melhor algoritmo para a solucionar o problema, documentando todos os passos do processo da geração do modelo, a avaliação e seleção das melhores métricas.

5) Avaliação: Essa fase busca validar os melhores modelos coletando para que seja feita a implantação, busca nessa fase recolher os conhecimentos que cada modelo agrega no processo.

6) Implementação: Essa fase busca descrever como o conhecimento adquirido com o projeto de mineração de dados pode ser aplicado na organização em suas atividades corriqueiras.

Nesse trabalho o uso do CRISP se fez necessário tendo em vista a fragmentação dos dados disponibilizados para esse estudo. Na preparação dos dados, CRISP auxiliou no ajuste das informações contidas no arquivo origem onde constava os dados acadêmicos dos alunos do curso Ciência da Computação, visto compreender melhor os dados foi consolidado as informações de cada ano em que o aluno possuía uma matrícula ativa, sendo assim o que possibilitou uma modelagem mais assertiva, visto que se criou agrupamentos ano a ano do desempenho do aluno consolidando assim uma base de dados ao contrario da sua origem mais condizendo para a aplicação dos modelos.

4.2 Preparação dos Dados

Os dados que foram utilizados nesse estudo foram extraídos do sistema acadêmico da UFPR e antes de ser disponibilizados passaram por um processo de mascaramento de informações sensíveis ou seja anonimizados, impossibilitando a identificação de qualquer aluno.

E para atingir o resultado final desse trabalho foi necessário seguir algumas etapas:

Coleta dos dados: nessa etapa será feito o levantamento dos registros dos alunos quanto à informação do curso, matrícula, situação e dados do aluno, pois esses dados irão determinar o quão preditivo o modelo será.

Engenharia de recursos nessa etapa foi calculado a média do primeiro e segundo semestre de cada aluno no período que este se apresentava matriculado afim de apresentar o desempenho em cada período e assim criar novos atributos a serem utilizados nos modelos.

Escolha do modelo nessa etapa visto a grande gama de modelos de machine learning disponíveis, cada um com seu objetivo. Desta maneira foi escolhido o modelo que mais se adequou no objetivo do trabalho, ou seja, a classificação binária entre evasão e a formatura.

Treinamento e avaliação nessa etapa os dados já devidamente trabalhados serão divididos entre teste e treino, onde o primeiro será utilizado para validar o treinamento e o segundo em treinar o modelo propriamente dito; dessa forma a máquina terá efetivamente aprendido com seus erros, aplicando a cada evolução os dados tanto de teste quanto de treino.

Predição nessa etapa o algoritmo está para efetivamente responder as perguntas, visto os dados que foram e serão imputados a ele.

4.3 Parametrização dos Algoritmos

Nesse trabalho foi escolhido a linguagem de programação Python, visto ser uma linguagem de programação amplamente utilizada na comunidade de ML, também foi utilizado os seguintes pacotes:

Pandas Biblioteca que proporciona estruturas de dados voltadas para a análise e manipulação de dados

Tensorflow: Plataforma open source para ML, fornecendo ferramentas flexíveis, bibliotecas e uma comunidade de pesquisadores e desenvolvedores.

Scikit-Learn: Biblioteca que disponibiliza ferra-mentas eficientes para mineração e análise dedados

Seaborn: biblioteca de visualização de dados do Python baseado no Matplotlib. Ele provê uma interface de alto nível para construção de gráficos estatísticos atrativos e informativos.

Os modelos adotados para esse trabalho foi árvore de decisão e regressão logística em que para ambos os modelos foi feito balanceamento dos dados aplicando Oversampler (Cria novas observações da classe minoritária a partir das informações contidas nos dados originais) seguindo os seguintes parâmetros:

Árvore de decisão:

- ***random_state*** = 0: Semente usada para a geração de números aleatórios.
- ***min_impurity_decrease*** = 0.09: O nó será dividido se essa divisão induzir uma diminuição da impureza.
- ***criterion*** = gini: Função para medir a qualidade de um split. Valores aceitos são *"gini"* e *"entropy"*.
- ***max_depth*** = 3: Profundidade máxima da árvore, limitando o crescimento da árvore.
- ***min_samples_leaf*** = 1: Número mínimo de exemplares que são necessários em um nó da folha.

Regressão logística:

- ***random_state*** = 0: Semente usada para a geração de números aleatórios.
- ***max_iter*** = 100: Número máximo de interações para a convergência.

4.4 Métricas de Desempenho

Para verificar o desempenho são acurácia, AUC (área abaixo da curva) e matriz de confusão.

Acurácia do modelo: É a proximidade de um resultado com o seu valor de referência real dessa forma quanto maior for a acurácia, mais próximo da referência ou valor real é o resultado encontrado.

AUC: O valor do AUC varia de zero a um e o limiar entre classe é meio. Ou seja, acima desse limiar o algorítimo classifica em uma classe e abaixo em outra. Quanto maior for o AUC melhor.

Matriz de confusão: É a tabela que mostra a frequência para cada classe do modelo, ela mostra a frequência de verdadeiros positivos, falsos positivos, falso negativo e falso verdadeiro.

- Verdadeiro positivo: Ocorre quando um conjunto real, a classe que estamos buscando foi prevista corretamente.
- Falso positivo: Ocorre quando no conjunto real, a classe que estamos buscando prever foi prevista incorretamente.
- Falso negativo: Ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista incorretamente.
- Falso verdadeiro: Ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista corretamente.

5 Resultados e Análise

Nessa seção iremos apresentar os resultados obtidos pela regressão logística e árvore de decisão previamente apresentadas. As principais métricas utilizadas serão acurácia, AUC (área abaixo da curva) e matriz de confusão:

- Acurácia: Razão entre as predições corretas e o total de elementos na base de teste.
- Precisão: Proporção de observações positivas preditas corretamente em relação ao total de observações positivas preditas.
- Recall: Proporção de observações positivas preditas corretamente para todas as observações na classe atual.
- F1-score: A pontuação F1 é a média ponderada de precisão e recuperação. Portanto, essa pontuação leva em consideração os falsos positivos e os falsos negativos. Intuitivamente, não é tão fácil de entender quanto a precisão, mas F1 geralmente é mais útil do que a precisão, especialmente se você tiver uma distribuição de classes desigual. A precisão funciona melhor se falsos positivos e falsos negativos tiverem custos semelhantes. Se o custo de falsos positivos e falsos negativos são muito diferentes, é melhor olhar para Precisão e Recuperação.

Os resultados a seguir foram obtidos aplicando os modelos de arvore de decisão e regressão logística do primeiro ano ao quarto ano de graduação do curso de Ciências de Computação.

Foi considerado o primeiro ano com ingresso em 2011 onde 40.39% terminaram o curso e 59.61% abandonaram. A execução da regressão logística apresentou os seguintes resultados com uma acurácia de 0.75, a precisão para a classe evasão foi de 0.76 e para o contrário foi de 0.73, nessa mesma perspectiva tivemos um valor de recall de 0.85 para o evasão e 0.61 para não evasão, o F1-score para as classes ficou em 0.80 para o evasão e 0.66 para não evasão.

Na matriz de confusão houve uma discriminação de 46 casos como verdadeiro positivo, 8 casos como falso positivo, 14 como falso negativo e 22 casos como verdadeiro negativo conforme apresentado na Figura 1.

A Tabela 2 apresenta os resultados obtidos ao aplicar os dados ao modelo de árvore de decisão com os seguintes resultados com uma acurácia de 0.74, a precisão para a classe evasão foi de 0.86 e para o contrário foi de 0.64, nessa mesma perspectiva tivemos um valor de recall de 0.69 para o evasão e 0.83 para não evasão, o F1-score para as classes ficou em 0.76 para o evasão e 0.72 para não evasão.

O grafo da figura 3 apresenta a descriminação entre as classes 0 para evasão e 1 para não evasão. O valor para a média do primeiro ano é menor ou igual a 48.05, ou seja, a linha de corte para a evasão e não evasão foi acima citada.

Na aplicação dos modelos de árvore de decisão e regressão logística, para o segundo ano foram consideradas as médias gerais do primeiro ano e segundo ano.

Relatório de Classificação:

	precision	recall	f1-score	support
0	0.7667	0.8519	0.8070	54
1	0.7333	0.6111	0.6667	36
accuracy			0.7556	90
macro avg	0.7500	0.7315	0.7368	90
weighted avg	0.7533	0.7556	0.7509	90

Acurácia: 0.7556

AUC: 0.7315

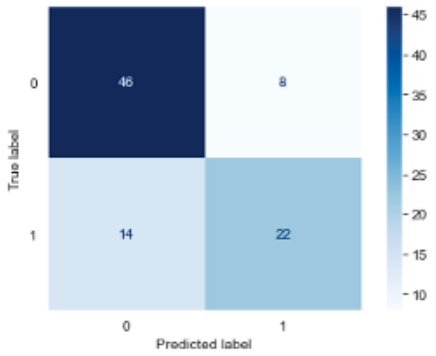

Figura 1: Relatório de classificação e matriz de confusão para o ingresso em 2011 onde evasão é igual a 0 e não evasão é igual a 1.

MEDIA_PRIMEIRO_ANO:1.0
COD_GRR:0.0

	precision	recall	f1-score	support
0	0.86	0.69	0.76	54
1	0.64	0.83	0.72	36
accuracy			0.74	90
macro avg	0.75	0.76	0.74	90
weighted avg	0.77	0.74	0.75	90

Figura 2: Relatório de classificação.

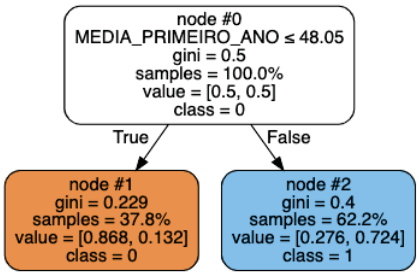

Figura 3: Árvore de decisão onde evasão é igual a 0 e não evasão é igual a 1.

Para esse cenário tivemos uma taxa de evasão de 47.04% e para o alunos que concluíram foi de 52,96%.

A execução da regressão logística apresentou os seguintes resultados com uma acurácia de 0.79, a precisão para a classe evasão foi de 0.87 e para o contrário foi de 0.75, nessa mesma perspectiva tivemos um valor de recall de 0.66 para o evasão e 0.91 para não evasão, o F1-score para as classes ficou em 0.75 para o evasão e 0.82 para não evasão.

Na matriz de confusão houve uma discriminação de 21 casos como verdadeiro positivo, 11 casos como falso positivo, 3 como falso negativo e 33 casos como verdadeiro negativo conforme apresentado na Figura 4.

Relatório de Classificação:

	precision	recall	f1-score	support
0	0.8750	0.6562	0.7500	32
1	0.7500	0.9167	0.8250	36
accuracy			0.7941	68
macro avg	0.8125	0.7865	0.7875	68
weighted avg	0.8088	0.7941	0.7897	68

Acurácia: 0.7941

AUC: 0.7865

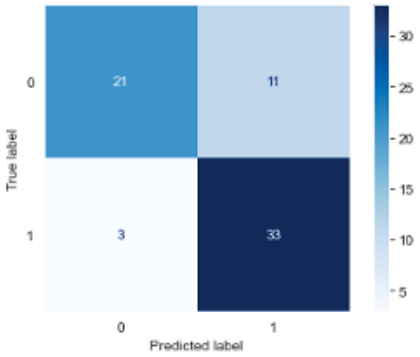

Figura 4: Relatório de classificação e matriz de confusão média do primeiro e segundo ano com ingresso em 2011 onde evasão é igual a 0 e não evasão é igual a 1.

A Tabela 5 apresenta o desempenho do modelo de árvore de decisão ao se aplicar os mesmos dados aplicados a regressão logística.

Acurácia de 0.79, a precisão para a classe evasão foi de 0.78 e para o contrário foi de 0.81 nessa mesma perspectiva tivemos um valor de recall de 0.78 para o evasão e 0.81 para não evasão, o F1-score para as classes ficou em 0.78 para o evasão e 0.81 para não evasão.

MEDIA_PRIMEIRO_ANO:0.0
MEDIA_SEGUNDO_ANO:1.0
COD_GRR:0.0

	precision	recall	f1-score	support
0	0.78	0.78	0.78	32
1	0.81	0.81	0.81	36
accuracy			0.79	68
macro avg	0.79	0.79	0.79	68
weighted avg	0.79	0.79	0.79	68

Figura 5: Relatório de classificação.

No grafo da figura 6 veremos que os alunos que obtiveram uma média superior a 52 concluíram o curso.

Ao se analisar o terceiro ano dos alunos que ingressaram em 2011 a taxa de alunos que concluíram fica em 65.26% e os que evadiram do curso ficou em 34.74%.

Para esse ano obtivemos os seguintes resultados da aplicação da regressão logística, acurácia de 0.9074, a precisão para a classe evasão foi de 0.85 e para o contrário foi de 0.94, nessa mesma perspectiva tivemos um valor de recall de 0.89 para o evasão e 0.91 para não evasão,

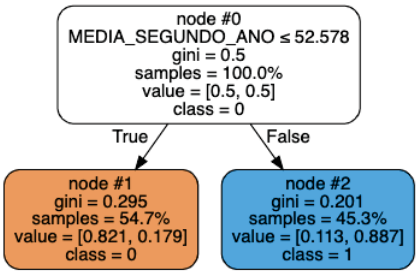

Figura 6: Árvore de decisão onde evasão é igual a 0 e não evasão é igual a 1.

o F1-score para as classes ficou em 0.87 para o evasão e 0.92 para não evasão.

Na matriz de confusão temos os seguintes valores 17 casos como verdadeiro positivo, 2 casos como falso positivo, 3 como falso negativo e 32 casos como verdadeiro negativo conforme apresentado na Figura 7.

Relatório de Classificação:

	precision	recall	f1-score	support
0	0.8500	0.8947	0.8718	19
1	0.9412	0.9143	0.9275	35
accuracy			0.9074	54
macro avg	0.8956	0.9045	0.8997	54
weighted avg	0.9091	0.9074	0.9079	54

Acurácia: 0.9074

AUC: 0.9045

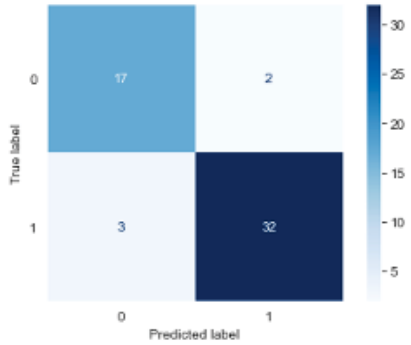

Figura 7: Relatório de classificação e matriz de confusão média do primeiro, segundo e terceiro ano com ingresso em 2011 onde evasão é igual a 0 e não evasão é igual a 1.

A Tabela 8 apresenta o desempenho do modelo de árvore de decisão ao se aplicar os mesmos dados aplicados a regressão logística.Acurácia de 0.94, a precisão para a classe evasão foi de 0.94 e para o contrário foi de 0.94 nessa mesma perspectiva tivemos um valor de recall de 0.89 para o evasão e 0.97 para não evasão, o F1-score para as classes ficou em 0.92 para o evasão e 0.96 para não evasão.

Para o terceiro ano o grafo da figura 9 mostra que os alunos que ficaram com media abaixo de 43.6 abandonaram o curso.

Ao se considerar o quarto ano do curso de ciência da computação se levou em consideração as médias do primeiro, segundo, terceiro e quarto ano com ingresso

MEDIA_PRIMEIRO_ANO:0.0
MEDIA_SEGUNDO_ANO:0.0
MEDIA_TERCEIRO_ANO:1.0
COD_GRR:0.0

	precision	recall	f1-score	support
0	0.94	0.89	0.92	19
1	0.94	0.97	0.96	35
accuracy			0.94	54
macro avg	0.94	0.93	0.94	54
weighted avg	0.94	0.94	0.94	54

Figura 8: Relatório de classificação.

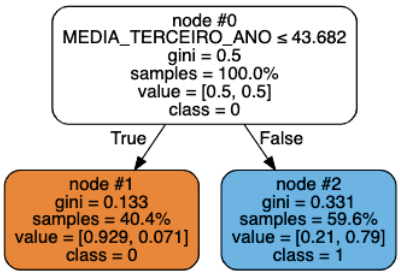

Figura 9: Árvore de decisão onde evasão é igual a 0 e não evasão é igual a 1.

em 2011 e nesse ano houve uma taxa de formandos de 76.27% e uma taxa de 23.73% de evasão. Nessa cenário foi aplicado o modelo de regressão logística senda as métricas obtidas, uma acurácia de 0.95, uma precisão de 0.9091 para a classe evasão e de 0.9706 no caso contrário para o recall a classe evasão ficou em 0.9091 e para não evasão em 0.9706 o f1-score para os que não se evadiram ficou em 0.9706 e para os que se evadiram do curso 0.9091. A matriz de confusão para o quarto ano no houve a seguinte descriminação, 10 casos de verdadeiro positivo, 1 caso de como falso positivo 1 como falso negativo e 33 de verdadeiro negativos conforme apresentado na Figura 10.

A Tabela 11 apresenta o desempenho do modelo de árvore de decisão ao se aplicar os mesmos dados aplicados a regressão logística. Acurácia de 0.91, a precisão para a classe evasão foi de 0.77 e para o contrário foi de 0.97 nessa mesma perspectiva tivemos um valor de recall de 0.91 para o evasão e 0.91 para não evasão, o F1-score para as classes ficou em 0.83 para o evasão e 0.94 para não evasão.

Para o quarto ano o grafo da figura 12 mostra que os alunos que ficaram com média abaixo de 53 abandonaram o curso.

6 Conclusões e Trabalhos Futuros

Nesse trabalho não tentamos encontrar uma causa que leva o aluno a evadir mas sim analisar o desempenho do aluno ano a ano e através de técnica de machine learning prever uma possível evasão, em suma o que tentamos fazer nesse trabalho foi analisar somente o desempenho do aluno até sua formatura ou evasão.

Relatório de Classificação:

	precision	recall	f1-score	support
0	0.9091	0.9091	0.9091	11
1	0.9706	0.9706	0.9706	34
accuracy			0.9556	45
macro avg	0.9398	0.9398	0.9398	45
weighted avg	0.9556	0.9556	0.9556	45

Acurácia: 0.9556

AUC: 0.9398

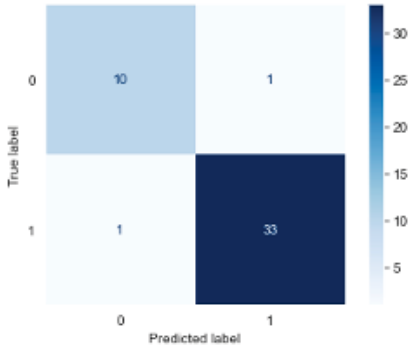

Figura 10: Relatório de classificação e matriz de confusão média do primeiro, segundo, terceiro e quarto ano com ingresso em 2011. Onde evasão é igual a 0 e não evasão é igual a 1.

MEDIA_PRIMEIRO_ANO:0.0
MEDIA_SEGUNDO_ANO:0.0
MEDIA_TERCEIRO_ANO:0.0
MEDIA_QUARTO_ANO:1.0
COD_GRR:0.0

	precision	recall	f1-score	support
0	0.77	0.91	0.83	11
1	0.97	0.91	0.94	34
accuracy			0.91	45
macro avg	0.87	0.91	0.89	45
weighted avg	0.92	0.91	0.91	45

Figura 11: Relatório de classificação.

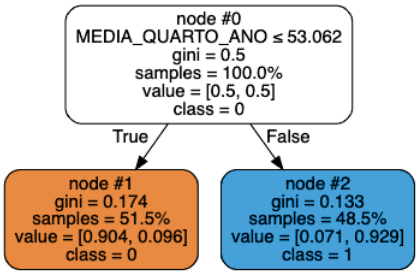

Figura 12: Árvore de decisão onde evasão é igual a 0 e não evasão é igual a 1.

Como trabalho futuro considera-se realizar uma análise considerando o peso de cada disciplina, ou conjunto de disciplinas por área, para o cenário de evasão ou formatura.

Aproveitando esse tópico que conclui o estudo, alguns agradecimentos se tornam necessários, em primeiro lugar minha esposa Karla Ursola com o seu apoio esse trabalho não teria sido concluído e ao meu orientador Marco A. Zanata Alves com sua análise criteriosa e inserções cirúrgicas de tão precisa tornou o andamento desse trabalho prazeroso só tenho a dizer a ambos muito obrigado.

Referências

[1] DEKKER, G. W. *Predicting students drop out: A case study. International Working Group on Educational Data Mining*, ERIC, 2009.

[2] FU, J.-H. *A support vector regression-based prediction of students' school performance. In: IEEE. 2012 International Symposium on Computer, Consumer and Control*, [S.l.], 2012. p. 84–87. Citado na pa´gina 42.

[3] Géron Aurélien, *Mãos à obra: Aprendizado de Máquina com Scikit-Learn & TensonrFlow*, (Alta Books, Rio de Janeiro, 2019), 1ª ed, Prefácio XIII, pag 139.

[4] HOFFMANN, G. Z. K. I. L. *Uma abordagem para previsa˜o de evasa˜o em cursos de graduaça˜o presenciais*,2016

[5] JADRIC´, M, *Student dropout analysis with application of data mining methods. Management: journal of contemporary management issues, Sveuc˘ilis˘te u Splitu, Ekonomski fakultet*, v. 15, n. 1, p. 31–46, 2010.

[6] Grus Joel, *Data Science do Zero*, (Alta Books, Rio de Janeiro, 2016), 1ª ed, pag 201.

[7] KARAMOUZIS S. T., *An artificial neural network for predicting student graduation outcomes. In: Proceedings of the world congress on engineering and computer science.*, [S.l.: s.n.], 2008. p. 991–994.

[8] MANHA˜ES, L. M. B.; CRUZ, S. M. S. da; ZIMBRA˜O, *G Evaluating performance and dropouts of undergraduates using educational data mining. In: Proceedings of the Twenty-Ninth Symposium on Applied Computing*, (S.l.: s.n.), 2014.

[9] MA´RQUEZ-VERA, C., *Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Applied intelligence*, Springer, v. 38, n. 3, p. 315–330, 2013.

[10] Martinho,V. R. D. C., Nunes, C., Minussi, C. R. , *An Intelligent System for Prediction of School Dropout Risk Group in Higher Education Classroom Based on Artificial Neural Networks*, (2013 IEEE 25th International Conference on Tools with Artificial Intelligence, Herndon, VA, 2013, pp. 159-166).

[11] MOURA, D.; ZIVIANI, F.; OLIVEIRA, L. C. V. , *Utilização do design instrucional em curso ead: análise do ambiente virtual de aprendizagem de curso técnico à distância de uma instituição pública de ensino.* , (Educação Tecnologia, v. 21, n. 1, 2018.)