9º EnPE
Encontro de Ensino, Pesquisa & Extensão
Patrocínio, MG, outubro de 2022

ALGORITMOS DE MINERAÇÃO DE DADOS PARA A PREVISÃO DO SUCESSO E FRACASSO NA EDUCAÇÃO SUPERIOR

Liliana Terrazas Balderrama, Danielli Araújo Lima
<liliana_tatiana@hotmail.com>,<danielli@iftm.edu.br>
Instituto Federal do Triângulo Mineiro (IFTM) Campus Patrocínio
Laboratório de Inteligência Computacional e Robótica (LICRo)
Modalidade: Pesquisa

Resumo: A educação passou por muitas transformações nas últimas décadas. a compreensão dos atributos e características individuais do estudante ainda é bastante relevante no contexto educacional. Neste sentido, este trabalho tem por objetivo explorar uma base de dados com técnicas de aprendizado de máquina com a finalidade de encontrar os principais atributos que podem influenciar no sucesso e fracasso escolar de estudantes da educação superior. Para essa abordagem, neste trabalho, utilizamos os algoritmos de Árvore de Decisão, k-Nearest Neighbor e Support Vector Machine. O melhor desempenho de precisão foi de 94.483%, enquanto que o pior desempenho foi de 86.207%, superando os resultados dos trabalhos precursores.

Palavras-chave: Aprendizado de máquina. Classificação. Inteligencia artificial. Educação.

Introdução

Sabe-se que a educação é um dos meios mais importantes para o desenvolvimento de uma sociedade. É exercida de forma que o indivíduo desenvolva suas habilidades, adequando-se à sociedade. Educação é uma prática social que visa o desenvolvimento do ser humano, de suas potencialidades, habilidades e competências. A educação, portanto, não se restringe à escola[1]. A educação, direito de todos e dever do Estado e da família, será promovida e incentivada com a colaboração da sociedade, visando ao pleno desenvolvimento da pessoa, seu preparo para o exercício da cidadania e sua qualificação para o trabalho[2]. No entanto, apesar de apoiada pelas leis federais, a grande maioria dos estudantes ainda deixam as escolas com muitas deficiências no conhecimento (CASTRO; GOMES, 2000; OLIVEIRA; BORUCHOVITCH; SANTOS, 2008). Para resolver este problema, algumas estratégias e estudos podem ser realizadas com o objetivo de encontrar parâmetros que afetam o desempenho de estudantes no nível superior.

Dessa forma, uma das maneiras que tem ajudado bastante é o refinamento e análise de dados dos fatores que levam os estudantes à terem sucesso ou insucesso acadêmico (DORNELES; LIMA, 2020). Existem diversas maneiras de se realizar a análise desses dados, podendo

[1] Ministério da Educação, Lei 9.394/96: Lei de Diretrizes e Bases da Educação Nacional. Brasília: Congresso Nacional, 1996. Disponível em: <http://portal.mec.gov.br/arquivos/pdf/ldb.pdf>.
[2] Constituição da República Federativa do Brasil. Brasília: Planalto, 1988.

ser a partir de estatística descritiva (ED) pura ou com o uso de ferramentas e tecnologias de mineração de dados (MD). A partir da MD é possível explorar grandes quantidades de dados com o objetivo de encontrar padrões e relacionamentos entre variáveis, detectando novos subconjuntos de dados. Dessa forma, é possível tomar decisão a partir desses conjuntos de dados ou até mesmo otimizar processos.

Neste trabalho, analisaremos uma base de dados (BD) recentes do mundo real do ensino superior em universidades estrangeiras. Os dados foram coletados do UCI Repository[3], um repositório aberto e online para a ciência de dados (CD). O objetivo é prever o desempenho do estudante de nível superior a partir de algumas variáveis por meio de algoritmos de mineração de dados (MD). As duas classes principais aprovado (A) ou reprovado (R) serão modeladas para a avaliação do aluno no ensino superior usando técnicas de MD a partir de uma análise científica feita pela CD. Essa pesquisa poderá ser aplicada em instituições de ensino superior (IES) brasileiras com o objetivo de melhorar o desempenho dos estudantes universitários.

Fundamentação teórica

Neste trabalho usaremos uma BD com um vetor de características de tamanho 33 (30 parâmetros, 2 classes e 1 identificador de linha), que foram coletados de 145 alunos da Faculdade de Engenharia e da Faculdade de Ciências da Educação em 2019, em duas faculdades distintas. Durante a determinação do sucesso acadêmico, exames escritos, provas e exames orais são considerados no sucesso cognitivo dos alunos e escalas são geralmente usadas para componentes. Os alunos foram avaliados em conceitos, que representam as classes abstraídas do problema, (0): `Fail`, (1): `DD`, (2): `DC`, (3): `CC`, (4): `CB`, (5): `BB`, (6): `BA`, (7): `AA`. No entanto, fizemos a abstração dessas 8 classes para duas classes úteis: $R(0)$: `reprovado` e $A(1)$: `aprovado`.

Muitas das abordagens educacionais que não se baseiam em determinados critérios de dados e pesos causam subjetividade durante o processo de avaliação e, assim, falsas avaliações podem ser realizadas (YILMAZ; SEKEROGLU, 2019). Cada um desses atributos podem interferir diretamente na taxa de sucesso e insucesso escolar de alunos na educação superior.

Neste caso, temos as seguintes informações dos 30 parâmetros conjunto de dados. As perguntas referentes aos dados (Q_1 - Q_{10}) são as **perguntas pessoais**: (1) idade, (2) sexo, (3) tipo de escola do ensino médio, (4) tipo de bolsa de estudo, (5) emprego adicional, (6) esportes e artes, (7) relacionamento, (8) salário, (9) transporte, (10) acomodação. As perguntas referentes aos dados (Q_{11} - Q_{16}) são as **perguntas familiares**: (11) nível de escolaridade da mãe, (12) nível de escolaridade do pai, (13) número de irmãos, (14) relacionamento dos pais, (15) tipo de trabalho da mãe, (16) tipo de trabalho do pai. As perguntas referentes aos dados (Q_{17} - Q_{30}) são as **perguntas educacionais**: (17) quantidade de horas de estudo diárias, (18) leitura não científica, (19) leitura científica, (20) participação em seminário e conferencia, (21) efeito dos projetos e atividades, (22) participação em leituras, (23) tipo de estudo (grupo | individual), (24) tipo de estudo (regular | semana passada), (25) tirando notas, (26) escrita e leitura, (27) efeito da discussão em sala, (28) efeito da aula invertida, (29) GPA semestre passado, (30) CGPA esperado na graduação.

O aprendizado de máquinas no inglês, Machine Learning (ML) pode ser entendido como máquinas com a capacidade de aprenderem sozinhas a partir de volumes de dados, reconhecendo padrões e criando relações entre estes, este campo de estudo é um subconjunto de algoritmos de Inteligência Artificial (IA). Neste trabalho, usaremos o ML com o objetivo é prever o desempenho final dos alunos do ensino superior. Um dos usos mais comuns para as técnicas de aprendizado de máquina é utilizar os algoritmos de ML para, a partir de situações

[3] UCI Repository - Higher Education Students Performance Evaluation Dataset Data Set <https://archive.ics.uci.edu/ml/datasets/Higher+Education+Students+Performance+Evaluation+Dataset>.

já conhecidas, prever ou classificar novas situações dentro do mesmo contexto trazendo novas informações (YILMAZ; SEKEROGLU, 2019).

Existem quatro tipos de técnicas de aprendizado de máquinas: (1) Aprendizado Supervisionado, (2) Aprendizado semi-supervisionado, (3) Aprendizado não-supervisionado e o (4) aprendizado por reforço. Neste trabalho focaremos no aprendizado supervisionado que tem como objetivo fundamental aprender uma função que mapeia uma entrada para uma saída com base em exemplos de pares entrada-saída. Os métodos de aprendizado supervisionado tentam inferir uma função a partir de dados de treinamento rotulados que consistem em um conjunto de exemplos de treinamento, sistematizando e analisando dados que trazem novas constatações.

Dentre as técnicas de aprendizado supervisionado temos: K-Nearest Neighbor (KNN), Support Vector Machine (SVM) e Decision Tree (DT). Neste sentido, o algoritmo DT é uma técnica usa a estratégia de dividir e conquistar para classificar atributos. O nó inicial, os nós internos e os nós folha representam o ponto inicial, atributos e rótulos de classe, respectivamente. O KNN considera um conjunto de $k = 3$ vizinhos mais próximos para fazer a votação majoritária, assim, a classe cujos elementos estão mais próximos do novo elemento é atribuída ao novo elemento a ser classificado. O SVM, por sua vez, é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para desafios de classificação ou regressão. Existem diferentes ferramentas para a mineração de dados e neste trabalho usaremos o KNIME® Analytics Platform (LIMA; ZATI; SILVA, 2017).

Proposta

Utilizando a plataforma KNIME® para análise de dados e construção de relatórios, foram elaborados 3 workflows, o DT, SVM e o KNN, conforme é apresentado na Figura 1. Inicialmente,

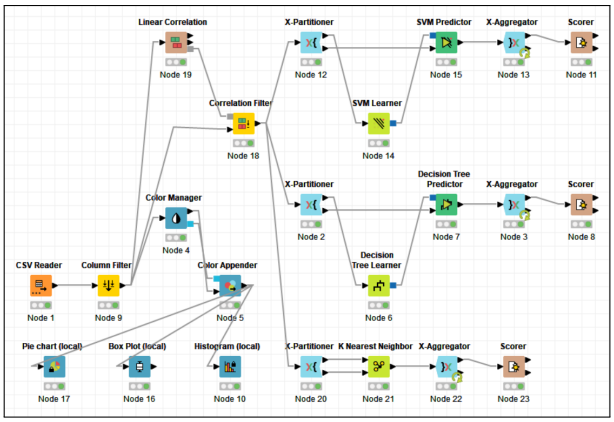


Figura 1 – Workflow no KNIME usado para a visualização e classificação de dados.

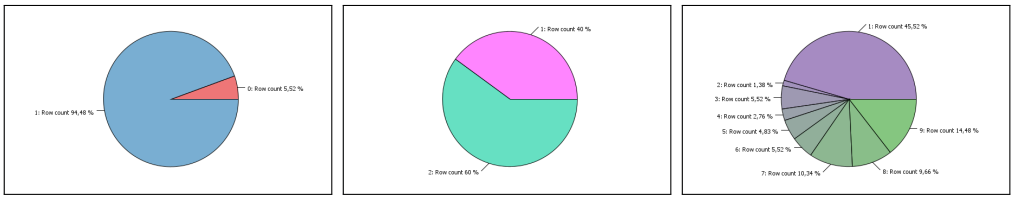
os dados do tipo (`.csv`) foram lidos pelo `CSV Reader`. Posteriormente, algumas colunas foram filtradas com o `Column Filter`, para filtrar os dados das 8 classes (YILMAZ; SEKEROGLU, 2019) e usamos somente as duas classes aprovado (1) e reprovado (0). Posteriormente fizemos algumas visualizações de dados usamos para isso os nós `Color Manager` e `Color Apprender` para a coloração de dados. Os gráficos que realizamos `Pie Chart`, `Box Plot` e `Histogram` para as visualizações de dados. Outro nó usado foi o `Linear Correlation` e o `Correlation Filter` para a filtragem em que as colunas mais correlacionadas sobrevivem, enquanto que todas as colunas correlacionadas são filtradas.

O nó `X-Agregator` é o primeiro em um loop de validação cruzada. No final do loop deve haver um `X-Aggregator` para coletar os resultados de cada iteração. Todos os nós entre

esses dois nós são executados quantas vezes as iterações devem ser executadas, neste caso, o valor de $x = 20$ iterações. Os nós `SVM Learner` e `SVM Apprender` servem para fazer a predição do algoritmo SVM, enquanto que os nós `Decision Tree Learner` e `Decition Tree Apprender` servem para fazer a predição e classificação dos dados a partir do algoritmo DT. Por fim, o nó `K Nearest Neighbor` faz a predição do KNN, com $k = 3$ vizinhos mais próximos. Ademais, o nó `Scorer` faz a avaliação de cada um dos algoritmos de predição que foram aqui usados por meio da acurácia e é apresentada uma matriz de confusão.

Resultados

Nesta seção serão apresentados os dados coletados a partir da visualização de dados. Neste sentido, nos gráficos de setores da Figura 2 temos os resultados referentes: (i) à quantidade de alunos reprovados (5.52%) e a quantidade de alunos aprovados (94.48%), ver Figura 2(a). Adicionalmente, (ii) o gênero dos estudantes entrevistados: (1) 40% do sexo feminino e 60% do

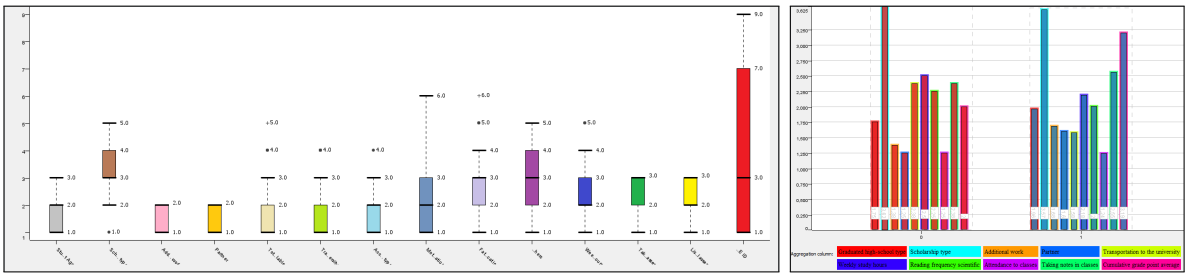


(a) Aprovação e reprovação. (b) Gênero. (c) Curso.

Figura 2 – Gráficos de setores realizados a partir dos 145 estudantes entrevistas.

sexo masculino, ver Figura 2(b). Por fim, na Figura 2(c), temos os 9 cursos em que os alunos estão matriculados.

Na Figura 3 temos a amplitude e mediana de alguns dos atributos listados nos boxplots da Figura 3(a) e o histograma agrupado pelas classes $A(1)$ e $R(0)$, sendo que os atributos de tipo de escola de ensino médio, média de notas anteriores, transporte até a universidade são os que mais interferem em média no desempenho do estudante universitário. Enquanto que os



(a) Boxplots para amplitude dos atributos. (b) Histograma médias aprovados.

Figura 3 – Boxplots para descobrir amplitudes e medianas dos atributos e histogramas das médias dos atributos que levam ao sucesso e fracasso escolar.

atributos de tipo de bolsa de estudos e presença em classes parecem ser os atributos que menos interferem, pois possuem médias similares em ambos os casos, conforme Figura 3(b).

Por fim, na Tabela 1 é possível perceber as acurácias para os três algoritmos de classificação usados neste trabalho, pela variação de filtros. Neste sentido, 6 filtros de correlação foram usados, e os melhores valores atingidos foram com os filtros $\{0.1, 0.2, 0.3\}$. O algoritmo

que encontrou o pior desempenho foi o SVM, com 86.207% de acurácia o o melhor resultado 94.483% foram observados pelos algoritmos DT com filtro 0.1, o KNN com filtros $\{0.1, 0.2\}$ e o SVM com filtros $\{0.1, 0.2, 0.3\}$. Para a DT e os demais atributos, percebemos que foi possível

Tabela 1 – Acurácia para cada um dos modelos de classificação com os respectivos filtros.

Filter	**Decision Tree**	**K-Nearest Neighbor**	**Support Vector Machine**
1.0	90.345%	93.103%	86.207%
0.5	90.345%	93.103%	86.897%
0.4	93.793%	93.103%	91.034%
0.3	92.414%	94.483%	94.483%
0.2	91.724%	94.483%	94.483%
0.1	94.483%	93.103%	94.483%

encontrar 4 atributos que influenciam diretamente no rendimento escolar: `Educação do pai`, `Relacionamento dos pais`, `Discussão em seminários` e `Tipo de Curso`. Sendo assim, esses parâmetros devem ser considerados pelos gestores e pedagogos para a avaliação de cursos.

Considerações finais

Neste trabalho foram analisados dados de estudantes da educação superior cujo foco era encontrar parâmetros que afetam o desempenho escolar de estudantes. Neste sentido, algoritmos de mineração de dados (DT, KNN e SVM) foram usados com a finalidade de encontrar os parâmetros. O pior resultado foi de 86.207% para o SVM, e os três algoritmos conseguiram encontrar a melhor acurácia de 94.483 porém com valores de filtros diferentes. Em trabalhos futuros esperamos rodar os algoritmos com $x = 10^2$ iterações (cross-validation). Além de testar outros valores de k para o algoritmo KNN e e fazer uma normalização de dados para ver se resultados melhores de acurácia poderão ser alcançados

Referências

CASTRO, S. L.; GOMES, I. Dificuldades de aprendizagem da língua materna: aprender a literacia. Universidade Aberta, 2000. 1

DORNELES, J. N.; LIMA, D. A. Arvore de decisao para a predicao dos principais fatores que afetam o aprendizado dos estudantes na disciplina de lingua portuguesa. In: **Congresso de Pos-Graduacao em Analise de Desenvolvimento de Sistemas**. [S.l.: s.n.], 2020. 1

LIMA, D. A.; ZATI, A. F.; SILVA, E. C. Análise de dados no google classroom para auxiliar na diminuição do distanciamento transacional nas disciplinas da área de lnformática. In: **TISE Conferência Internacional sobre Informática na Educação**. [S.l.: s.n.], 2017. 3

OLIVEIRA, K. L. de; BORUCHOVITCH, E.; SANTOS, A. A. A. dos. Leitura e desempenho escolar em português e matemática no ensino fundamental. **Paidéia (Ribeirão Preto)**, v. 18, n. 41, p. 531–540, 2008. 1

YILMAZ, N.; SEKEROGLU, B. Student performance classification using artificial intelligence techniques. In: SPRINGER. **International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions**. [S.l.], 2019. p. 596–603. 2, 3