Brazilian Journal of Development

Predição do rendimento dos alunos em lógica de programação com base no desempenho das disciplinas do primeiro período do curso de ciências e tecnologia utilizando técnicas de mineração de dados

Predicting student performance in programming logic based on the performance of first-course science and technology subjects using data mining techniques

DOI:10.34117/bjdv6n1-186

Recebimento dos originais: 30/11/2019
Aceitação para publicação: 16/01/2020

Renata Pitta Barros
[1]Departamento de Engenharia Elétrica e de Computação - Universidade Federal do Rio Grande do Norte (UFRN) – Natal – RN – Brasil
repitta@gmail.com

Orivaldo Vieira de Santana Junior
[2]Escola de Ciência e Tecnologia - Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN – Brasil
orivaldo@gmail.com

Igor Rosberg de Medeiros Silva
[2]Escola de Ciência e Tecnologia - Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN – Brasil
igorosbergster@gmail.com

Luana Fernandes dos Santos
[2]Escola de Ciência e Tecnologia - Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN – Brasil
luanafs.info@gmail.com

Vilson Rodrigues Câmara Neto
[2]Escola de Ciência e Tecnologia - Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN – Brasil
vilsonrodrigues07@gmail.com

RESUMO
Os altos índices de reprovação e evasão de estudantes universitários nas disciplinas iniciais de programação apresentam uma estatística preocupante enfrentada pelos coordenadores dos cursos da área de Tecnologia. O problema da reprovação dos estudantes nessas disciplinas é, muitas vezes, apontado como um fator influenciador da evasão dos cursos. Esta pesquisa propõe a utilização de técnicas de Mineração de Dados Educacionais para tentar predizer o desempenho dos alunos na disciplina de Lógica de Programação, do segundo período do curso de Bacharelado em Ciências e Tecnologia da UFRN, através do desempenho nas disciplinas do primeiro período do curso. Os resultados mostraram que é possível inferir o rendimento dos estudantes com uma acurácia de até 77%, sendo esta informação útil para a realização de ações para evitar a reprovação/evasão e, principalmente, para personalizar o ensino de lógica de programação.

Palavras-chave: ciência de dados educacionais, aprendizado de máquina e dados educacionais

Brazilian Journal of Development

ABSTRACT
The high rates of university student disapproval and dropout in the initial courses of programming present a worrying statistics faced by the coordinators of the Technology programs. The problem of students disapproval is often pointed as an influential factor in dropping out of university programs. This work proposes the use of techniques of Educational Data Mining to predict the performance of students in the course of Programming Logic, of the second period of the Bachelor of Science and Technology program at UFRN, based on performance in the courses of the first period of that program. The results showed that it is possible to infer students performance with an accuracy of up to 77%, this information being useful for planning actions to avoid disapproval/dropout and, especially, to personalize the teaching of programming logic.

Keywords: educational data science, machine learning and educational data

1 INTRODUÇÃO

De acordo com o último Censo da Educação, ano de 2017, disponibilizado pelo Ministério da Educação (MEC), a quantidade de alunos que abandonam seus cursos ou trancam a matrícula é alarmante. Em alguns cursos a taxa de evasão ultrapassa os 50%. [INEP 2019].

Segundo dados fornecidos pela Secretaria Acadêmica da Escola de Ciências e Tecnologia (ECT) da Universidade Federal do Rio Grande do Norte (UFRN), de 2016, o curso tem um fluxo de aproximadamente 5.000 alunos, porém apenas 3.447 matrículas estão, atualmente, ativas no sistema. Além disso, o percentual de formação por turma é de apenas 40% do grupo ingressante, totalizando cerca de 200 alunos colando grau por semestre. Os 60% restantes dividem-se entre abandono total do curso (uma estimativa de 25%) e trancamento ou retardo do tempo de formação (em torno de 35%). Muitas podem ser as razões que implicam nessa realidade. A evasão no ensino superior é um problema social (sob o ponto de vista educacional) e administrativo (devido ao impacto negativo que gera nas universidades) e deve ser combatida veementemente.

A busca de causas para esse problema tem sido objeto de estudo de muitos trabalhos e pesquisas, como pode ser visto em Silva Filho et al. (2017), Santos et al. (2011) e Lobo (2012). Os trabalhos discutem os tipos de causas de evasão. Dentre os motivos temos o despreparo para acompanhar as aulas, a falta de identificação com o curso, desinformação na opção por uma carreira profissional e a inadequação nos métodos de estudo. Esse último motivo que causa evasão deve ser combatido com ações pedagógicas por parte das instituições.

As ações realizadas pelas instituições para ajudar estudantes do primeiro ano são geralmente atividades de tutoria e monitoria, porém corroboramos com Nascimento (2018), o qual sugere que haja um reconhecimento prévio e uma assistência intensiva e personalizada aos alunos que correm risco de retenção ou evasão.

Recentemente, técnicas de Mineração de Dados (do inglês Data Mining, DM) são aplicadas como forma de predição de desempenho dos estudantes. Segundo a literatura essas técnicas dão

suporte aos educadores para tomar decisões estratégicas que colaboram para a redução dos números de evasão nas instituições de ensino.

O objetivo deste trabalho é identificar os estudantes que necessitam de apoio didático na disciplina de Lógica de Programação (LOP), no segundo período do curso da ECT da UFRN. Para realizar as análises dos dados, neste artigo, foi utilizada uma base de dados dos alunos das turmas iniciais da ECT que utilizam um ambiente virtual intitulado de LOP [LOP 2019] para execução de exercícios e provas da disciplina de Lógica de Programação, entre os anos de 2017 e 2018.

Essa base de dados foi escolhida, pois, sendo um dos pilares da implementação da Educação 4.0 [De Trabalho 2018], a disciplina que contempla os conceitos iniciais para o curso de Ciências da Computação, Algoritmos e Lógica de Programação, apresenta, em sua aplicação, os mais altos níveis de evasão e reprovações.

Dessa forma, este trabalho busca analisar a relação entre desempenho nas disciplinas do primeiro período do aluno e o seu desempenho na disciplina de LOP. Os resultados iniciais deram indícios de que essa relação existe, sendo uma informação relevante para que medidas sejam tomadas para a diminuição da retenção ou evasão acadêmica ao suprir deficiências provenientes das disciplinas iniciais do curso.

O artigo está organizado da seguinte maneira: Na Seção 2 apresentamos os conceitos de Mineração de dados e Mineração de dados Educacionais, na Seção 3, os trabalhos relacionados; na Seção 4 descrevemos como os dados foram obtidos e o pré-processamento realizado nos mesmos; os resultados são apresentados na Seção 5 e as considerações finais e trabalhos futuros, na Seção 6.

2 MINERAÇÃO DE DADOS

Mineração de dados (DM) é um conjunto de ferramentas e técnicas que, através do uso de algoritmos de aprendizagem ou classificação baseados em redes neurais e estatística, extraem padrões a partir de dados. A DM é parte de um âmbito mais amplo, conhecido como Descoberta de Conhecimento em Banco de Dados (do inglês Knowledge Discovery in Databases, KDD). KDD é o processo de descobrir conhecimento útil a partir de dados e, além da etapa de DM, envolve mais cinco etapas, como preparação, seleção, limpeza dos dados, incorporação de conhecimento prévio apropriado e interpretação adequada dos resultados da mineração de dados. Tudo isso é realizado no intuito de garantir que resultados úteis sejam derivados dos dados [Fayyad et al. 1996].

Os conhecimentos obtidos através dos dados são subsídio para a melhoria das práticas pedagógicas, tendo como base uma área de pesquisa denominada Mineração de Dados Educacionais (do inglês Educational Data Mining, EDM). EDM está preocupada com o desenvolvimento de métodos para explorar informações coletadas de ambientes educacionais, permitindo compreender o

comportamento dos alunos e o ambiente no qual a aprendizagem ocorre, fornecendo insumos para que professores e alunos otimizem /personalizem o processo de ensino e aprendizagem [Baker et al. 2011].

As técnicas utilizadas em EDM são geralmente divididas em dois tipos principais: A análise preditiva ou supervisionada e a análise descritiva ou não supervisionada. Segundo Romero et al. (2010), o aprendizado supervisionado é o mais amplamente utilizado na prática. Tem-se duas tarefas nesse tipo de aprendizado: A classificação e a regressão. A tarefa de classificação é caracterizada quando a variável de saída é do tipo categórica ou nominal. A tarefa de regressão é quando a variável de saída é do tipo numérica. A classificação tem como objetivo aprender um mapeamento de entradas x para uma saída y ∈ {1, …, C} onde C é o número de classes. Temos a classificação binária, a variável alvo possui apenas dois valores e a classificação multiclasses, na qual a variável alvo é uma variável categórica com mais de duas classes.

A tarefa de classificação é geralmente utilizada para classificar perfis de alunos, classificação de estilos de aprendizagem, e previsão de desempenho. Na classificação os algoritmos mais utilizados são árvores de decisão, máquinas de vetores de suporte, naive bayes e redes neurais.

3 TRABALHOS RELACIONADOS

A previsão do desempenho acadêmico é um tema já estudado por diversos pesquisadores. Estudos mais antigos utilizam métodos, dentre eles estatísticos, para compreender o problema. Recentemente, com o advento da mineração de dados educacionais, os estudos têm utilizado técnicas de aprendizado de máquina para buscar soluções para esse tipo de problema.

No estudo de Hämäläinen et al. (2004) eles analisaram duas disciplinas de programação de computadores em um curso on-line. O trabalho utilizou regras de associação e modelos probabilísticos para identificar os fatores mais importantes para predizer os resultados finais nas duas disciplinas. O trabalho de Manhães et al. (2011) utiliza dados acadêmicos de alunos de graduação da UFRJ. Os resultados mostraram que utilizando as primeiras notas semestrais dos calouros é possível identificar com precisão de 80% a situação final do aluno no curso.

Os algoritmos IBk, RandomForest (RF), BayesNET (BNet) e MultilayerPerceptron (MLP) foram utilizados no trabalho de Pascal et al. (2015) e os resultados mostraram, com taxas de acerto superiores a 80%, que foi possível identificar os estudantes com tendência a evasão usando informações de desempenho na prova de ingresso e nas disciplinas pré-requisito da disciplina de programação. De Brito et al. (2014) tiveram como objetivo identificar os estudantes que necessitam de apoio didático. Foi analisado o desempenho acadêmico no primeiro período de um curso de graduação em Ciência da Computação. O algoritmo SMO obteve a maior precisão, com média de

74,7% de acerto. Segundo os autores, os resultados podem ser utilizados para que medidas sejam tomadas para a diminuição da retenção ou evasão acadêmica.

O trabalho de Santos et al. (2016) traz uma análise e classificação de publicações na área de Mineração de Dados Educacionais publicados por pesquisadores brasileiros em eventos e periódicos na América Latina nos últimos anos. Os resultados da classificação das publicações foram apresentados e apresentam informações a respeito da área de pesquisa.

4 COLETA DOS DADOS E PRÉ-PROCESSAMENTO

Para a realização deste estudo utilizou-se um conjunto de dados real de alunos do curso de Ciência e Tecnologia da UFRN, entre os anos de 2017 e 2018. Esses dados foram fornecidos pela Secretaria Acadêmica da ECT da UFRN, e são dados provenientes do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) [SIGAA 2019]. Representam o desempenho dos alunos nas disciplinas do primeiro período do curso e o desempenho na disciplina de Lógica de Programação do segundo período. As disciplinas consideradas na análise são: Pré-Cálculo, Cálculo I, Química Geral, Práticas de Leitura e Escrita I, Ciência, Tecnologia e Sociedade e Vetores e Geometria Analítica. Essas disciplinas pertencem ao primeiro período do curso que historicamente possui uma elevada taxa de reprovação/evasão no curso.

A média final dos alunos na disciplina de LOP foi utilizada como critério para a seleção das classes de desempenho, conforme discutido mais adiante (variável alvo). Uma vez que o fracasso na disciplina de LOP pode estar relacionado com deficiências prévias de conteúdo, optou-se por selecionar as informações relacionadas ao desempenho das disciplinas do primeiro período do aluno, as quais são essenciais para a disciplina de LOP.

Os atributos de entradas utilizados foram as médias das disciplinas obtidas no primeiro período do curso. Uma análise da correlação entre os atributos de entrada foi realizada para identificar a existência de possíveis redundâncias entre os mesmos, proveniente de uma correlação positiva muito alta.

Brazilian Journal of Development

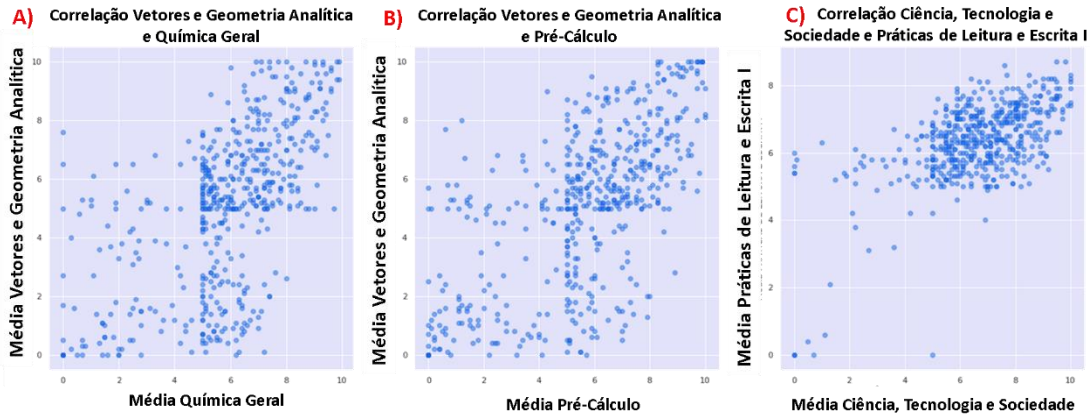

Figura 1. Gráficos de Dispersão para correlação entre os atributos das médias das notas de disciplinas do primeiro período do curso

A Figura 1(A) apresenta o gráfico de dispersão dos atributos Média Química Geral e Vetores e Geometria Analítica, com um coeficiente de correlação r =0,59, o que indica uma correlação positiva moderada. Na Figura 1(B), para os atributos Média Pré-Cálculo e Média de Vetores e Geometria Análitica, tem-se r =0,61, o que indica uma correlação positiva moderada. A Figura 1(C), para os atributos Média Ciência, Tecnologia e Sociedade e Práticas de Leitura e Escrita I temos r =0,65, que representa uma correlação positiva moderada. Nenhum dos possíveis pares de atributos possui uma correlação positiva muito alta (superior a 0,9), indicando a relevância individual dos mesmos na tarefa de classificação dos dados. Além disso, testes foram realizados com a exclusão de cada uma das variáveis, onde foi observada a relevância dos seis atributos na tarefa de classificação.

Os algoritmos de classificação foram testados com 532 instâncias de alunos. Os estudantes foram divididos em duas classes distintas de acordo com situação na disciplina de LOP. Caso o estudante tenha sido aprovado, isto é, média acima de 5,0, é considerado da classe “APROVADO”. Caso o estudante tenha sido reprovado, isto é, média inferior a 5,0 e/ou não atendeu os critérios de assiduidade, é considerado como sendo da classe “REPROVADO”. A divisão dos estudantes de acordo com o tipo de classe é 326 instâncias para a classe Aprovado e 195 instâncias para a classe Reprovado.

5 DISCUSSÃO DOS RESULTADOS

Utilizamos neste trabalho os algoritmos de aprendizado de máquina na linguagem Python, implementados pela biblioteca scikit-learn, cujo código é aberto para a linguagem de programação Python. Ela inclui vários algoritmos de classificação, regressão e agrupamento, incluindo máquinas de vetores de suporte, florestas aleatórias, gradient boosting e k-means. É projetada para interagir com as bibliotecas Python numéricas e científicas NumPy e SciPy. Permite também que usuários experimentem e comparem resultados dos diferentes métodos [PYTHON 2019].

Brazilian Journal of Development

Para a realização dos testes foi utilizada também a biblioteca Pandas. Essa biblioteca permite a manipulação e a exploração dos dados. Na biblioteca scikit-learn foram utilizados quatro algoritmos de aprendizado de máquina pertencentes a classes distintas de classificadores, conforme pode ser visto na Tabela 1. Os algoritmos foram treinados com a situação final do aluno nas disciplinas do primeiro semestre para prever a sua situação ou desempenho em uma disciplina do segundo semestre. A seção quatro explica em mais detalhes sobre a situação/desempenho do estudante. A disciplina do segundo semestre escolhida para este estudo foi Lógica de Programação, dada a sua importância em trabalhar conceitos básicos da área de Ciência da Computação, que por sua vez, é um dos pilares da implementação da Educação 4.0.

Para isso, dividimos os dados em dois conjuntos. O primeiro foi o conjunto de treinamento com o qual treinamos os algoritmos para construir um modelo. O segundo foi o conjunto de testes usado para testar nosso modelo e ver quão precisas foram as suas previsões.

Tabela 1. Algoritmos de aprendizado de máquina utilizados

Algoritmo	Classe
MultinomialNB	Métodos Baysianos
KNeighborsClassifier	Métodos de Vizinhos mais próximo
SVM	Máquina de Vetor de Suporte
DecisionTreeClassifier	Árvore de Decisão

A divisão entre o conjunto de treinamento e testes foi feita em 70% dos dados para treinamento e 30% dos dados para teste. Como forma de avaliação do desempenho dos algoritmos, utilizou-se as métricas de acurácia, precisão, recall e verdadeiros positivos das duas classes. O classification report e as matrizes de confusão foram executadas para obter esses resultados. Os dados das execuções podem ser vistos nas Figura 2, 3, 4 e 5.

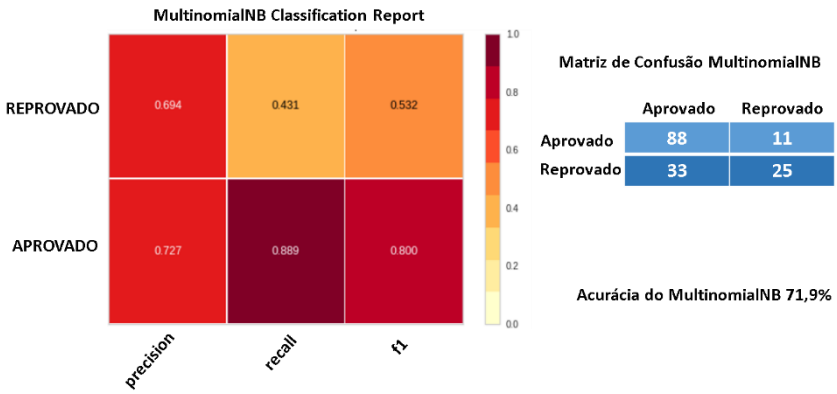

Figura 2. Matriz de confusão e o classification report do algoritmos de aprendizagem de máquina MultinomialNB

Brazilian Journal of Development

A taxa de acurácia do algoritmo MultinomialNB foi de 71,9%. O algoritmo faz a previsão de quase 73% das amostras identificadas como da classe Aprovado, como indica o valor *precision* da Figura 2. Já na classe Reprovado a métrica *precision* apresentou um valor de 0.69 indicando que muitas amostras identificadas como aprovado de fato eram reprovados, o que não colabora muito para a resolução do nosso estudo. A métrica recall apresentou um valor de 0,88 indicando que a maior parte das amostras identificadas como aprovado são de fato aprovadas.

O algoritmo KNeighborsClassifier foi executado com vários parâmetros k (quantidade de vizinhos analisados) diferentes, porém a melhor taxa de acurácia do algoritmo foi de 76,4% com o valor de k =17. Apesar da taxa de acurácia e a taxa de recall para a classe Reprovado tenham melhorado, concluímos, após uma análise da matriz de confusão, que ainda não temos valores satisfatórios para a previsão da classe dos reprovados, conforme pode ser observado na Figura 3.

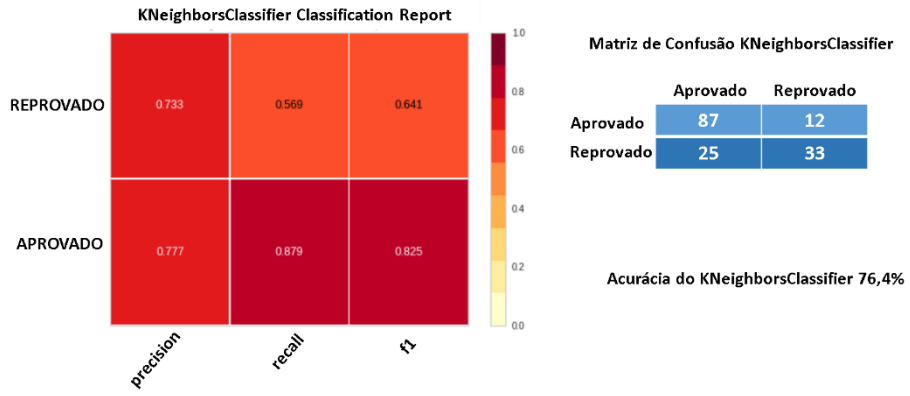

Figura 3. Matriz de confusão e o classification report dos algoritmos de aprendizagem de máquina KNeighborsClassifier

O algoritmo LinearSVC apresentou uma taxa de acurácia de 77% essa taxa foi alcançada com os parâmetros: random_state=3, tol=0,002. O algoritmo faz a previsão de quase 77% das amostras identificadas como da classe Aprovado, como indica o valor *precision* da Figura 4. Já na classe Reprovado a métrica *precision* apresentou um valor de 0.77 indicando que muitas amostras identificadas como aprovado de fato eram reprovados, o que não colabora muito para a resolução do nosso estudo. A métrica recall apresentou um valor de 0,90 indicando que a maior parte das amostras identificadas como aprovado são de fato aprovadas. Estes são resultados relevantes que podem contribuir no planejamento de matrícula do aluno para otimizar seu tempo no curso na identificação dos alunos aprovados, porém não muito eficiente para a previsão dos alunos reprovados.

Brazilian Journal of Development

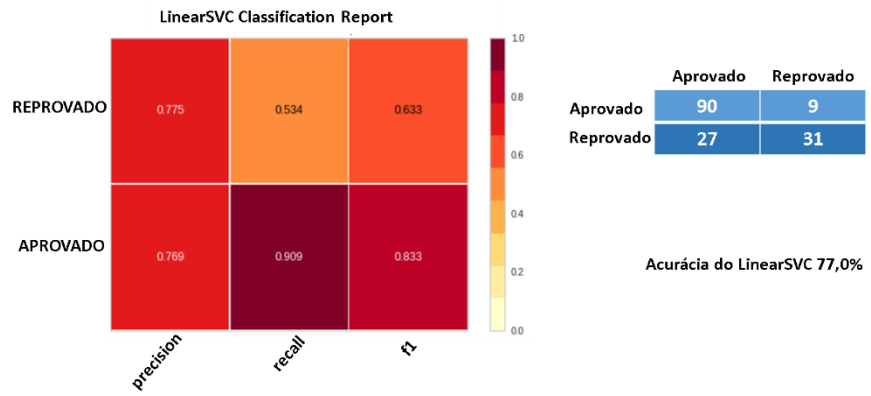

Figura 4. Matriz de confusão e o classification report dos algoritmos de aprendizagem de máquina LinearSVC

A Figura 5 apresenta os valores das métricas do algoritmo DecisionTreeClassifier. A taxa de acurácia foi de 72,6% com os parâmetros criterion='entropy', max_depth=1, min_samples_leaf=3. Esteve bem próximo da taxa do algoritmo MultinomialNB. Esse algoritmo fez a previsão de quase 86% dos aprovados. Esse modelo gerou as melhores taxas de recall, 0,81, para a classe dos reprovados. Isso mostra que este modelo trabalhou bem para a classe dos reprovados, contribuindo significativamente com o nosso estudo.

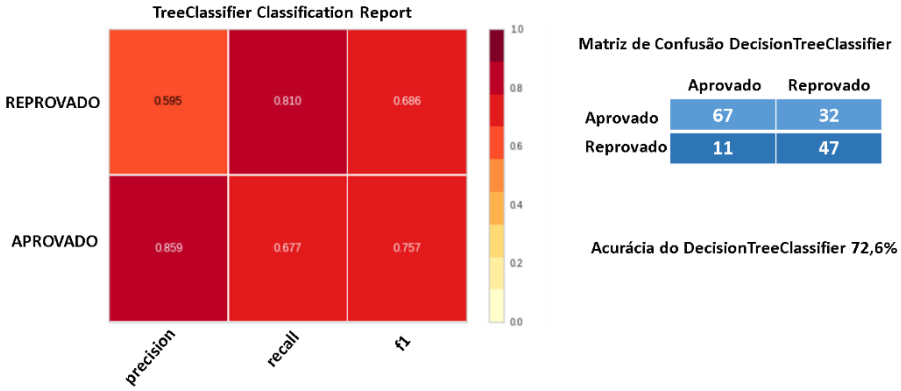

Figura 5. Matriz de confusão e o classification report dos algoritmos de aprendizagem de máquina DecisionTreeClassifier

A seleção de qual algoritmo deve ser utilizado para a previsão da evasão vai além do valor da acurácia, um algoritmo pode diferir do outro nos valores das taxas de acerto e erro na classificação dos exemplos positivos e negativos. Um classificador que possui uma elevada taxa de erro para falso positivo não é adequado para a solução do problema. Neste caso, considera-se um erro grave do algoritmo classificar um aluno com risco de evasão como sem risco. O erro do algoritmo de classificar um aluno no grupo de risco de evasão sem de fato ocorrer à evasão, falso negativo, é considerado um erro brando, menos grave.

Brazilian Journal of Development

Diante das análises apresentadas observamos que o algoritmo de classificação MultinomialNB apresenta um erro grave menor (taxa de erro falso positivo) para a classe de Aprovados e o algoritmo DecisionTreeClassifier apresenta um erro grave menor (taxa de erro falso negativo) para a classe de Reprovados.

6 CONSIDERAÇÕES FINAIS

Esta análise teve como objetivo encontrar a relação entre o desempenho nas disciplinas dos primeiros períodos dos estudantes e o seu desempenho na disciplina de Lógica de Programação do segundo período do curso de Ciência e Tecnologia da UFRN. Através da linguagem Python e suas bibliotecas Pandas e scikit-learn, obteve-se precisão de até 72%, utilizando um conjunto de seis atributos de entrada: As médias das notas das disciplinas de Pré-Cálculo, Cálculo I, Química Geral, Práticas de Leitura e Escrita I, Ciência, Tecnologia e Sociedade e Vetores e Geometria Analítica obtidas no primeiro período do curso.

Os resultados obtidos apontam indícios de que é viável realizar a predição do desempenho da disciplina de LOP baseado em suas notas das disciplinas do primeiro período, o que permite aos professores e tutores tomarem ações pedagógicas com o objetivo de contornar os altos índices de reprovação e evasão da disciplina e, ainda mais, personalizar o ensino da lógica de programação para os diferentes perfis de alunos. Tem-se conhecimento da existência de outras variáveis que podem influenciar o desempenho do aluno durante a sua jornada na disciplina, porém estas são muitas vezes subjetivas e difíceis de serem recuperadas, como motivação do aluno no curso, taxa de aprovação da turma, situação socioeconômica, entre outras.

Os resultados obtidos neste estudo podem ajudar os educadores uma vez que é possível obter estimativas sobre o desempenho dos alunos. É possível que esses resultados sirvam de base para o planejamento de estratégias e materiais de estudos personalizados que visam diminuir o número de reprovações, reduzindo, como consequência, a evasão dos alunos do curso de Ciência e Tecnologia.

Como trabalhos futuros, considera-se expandir este estudo para todas as disciplinas do curso, fornecendo assim uma ferramenta de apoio pedagógico aos coordenadores e professores. Será realizada também uma análise mais detalhada no processo de seleção dos atributos de entrada, investigando a existência de outras variáveis que possam influenciar no desempenho do aluno na disciplina de LOP. Estudar a relação entre o desempenho dos alunos na disciplina de LOP e o desempenho do aluno nas atividades da disciplina nas primeiras semanas de aula também se faz necessário. Além disso, acredita-se que é possível realizar pesquisa semelhante para alunos de cursos diferentes, verificando se os resultados são semelhantes para as disciplinas iniciais de programação.

Brazilian Journal of Development

REFERÊNCIAS

Baker, R., Isotani, S. e Carvalho, A. (2011) “Mineração de Dados Educacionais: Oportunidades para o Brasil”. Revista Brasileira de Informática na Educação, v. 19, n. 02, pp. 3-13.

De Trabalho, Grupo, and Temas Emergentes da Educação Básica. (2018) "O Dilúvio Digital e seus Impactos na Educação 4.0 e na Indústria 4.0." Investigação em Governança Universitária: Memórias: 188.

Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P. (1996) “From data mining to knowledge discovery in databases”. Al magazine, v. 17, n. 3, pp. 37-54.

Hämäläinen, W., Suhonen, J., Sutinen, E., and Toivonen, H. (2004) “Data mining in personalizing distance education courses”. In world conference on open learning and distance education, Hong Kong, pp. 1–11

INEP- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Último acesso em 27.06.2019. Disponível em : http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2018/censo_da_educacao_superior_2017-notas_estatisticas2.pdf

Lobo, M. B. C. M. (2012). "Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções." Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos 25

LOP - Plataforma open-source de gerenciamento de exercícios de programação - Último acesso em 02.07.2019. Disponível em: http://lop.ect.ufrn.br/

Manhães, L. M. B., Da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., & Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie) (Vol. 1, No. 1).

Nascimento, P. B. D. (2018). Recomendação de ação pedagógica no ensino de introdução à programação por meio de raciocínio baseado em casos.

Pascoal, Pascoal, T. A., Brito, D. M., & Rêgo, T. G. (2015). Uma abordagem para a previsão de desempenho de alunos de Computação em disciplinas de programação. Nuevas Ideas en Informática Educativa TISE, 2015, 454-458.

PYTHON - scikit-learn: machine learing in Python – Último acesso em 03.07.2019. Disponível em: https://scikit-learn.org/stable/

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2010). Handbook of educational data mining. CRC press,.

Santos, R., Pitangui, C., Vivas, A., & Assis, L. (2016, November). Análise de trabalhos sobre a aplicaçao de técnicas de mineraçao de dados educacionais na previsao de desempenho acadêmico. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação (Vol. 5, No. 1, p. 960).

Santos Baggi, C. A., & Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. Avaliação: revista da avaliação da educação superior, 16(2)..

SIGAA - Sistema Integrado de Gestão de Atividades Acadêmicas. Último acesso: 02.07.2019. Disponível: https://sigaa.ufrn.br/sigaa/public/home.jsf

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., & Lobo, M. B. C. M. (2007). A evasão no ensino superior brasileiro. Cadernos de pesquisa, 37(132), 641-659.