Estudo e implementação de algoritmos de aprendizagem de máquina supervisionado para a realização de processo de descoberta de conhecimento na base de dados do SIGAA-UFPI

Pedro Antonio F. da Silva[1], Felipe B. Caminha[1], Arthur C. Basílio[1], Vinícius P. Machado[1]

[1]Departamento de Ciência da Computação – Universidade Federal do Piauí (UFPI) – Teresina, PI – Brasil

p.antonio.f.s@gmail.com, felipebarroscaminha@gmail.com, a.basilio.99pause@gmail.com, vinicius@ufpi.edu.br

Abstract: This article proposes an application of supervisioned machine learning methods to identify and predict the students academic performance from the Federal University of Piauí (UFPI). By using of decision tree algorithm, profiles of students profiles of students related to their academic performances were found.

Resumo: Este artigo propõe uma aplicação de métodos de aprendizagem de máquina supervisionado para identificar e prever o desempenho acadêmico de alunos da Universidade Federal do Piauí (UFPI). Através de um algoritmo de árvore de decisão, foi encontrado perfis de alunos relacionados a seus desempenhos acadêmicos.

1. Introdução

Com o propósito de encontrar motivos que levam um aluno de ensino superior a ter baixo rendimento durante a graduação, objetiva-se utilizar ferramentas de Tecnologia da informação (TI) para auxiliar em tomadas de decisão que tratam desse problema. Dado isso, buscou-se encontrar uma aplicação de aprendizado de máquina relacionado a questão.

Análogo ao trabalho realizado por SILVA (2018), tentou-se identificar padrões na base de dados de estudantes de graduação na Universidade Federal do Piauí (UFPI). Esses dados foram obtidos no Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) e contém tanto informações acadêmicas de alunos quanto socioeconômicas e escolares anteriores a graduação. Diferentemente do trabalho citado, foi aplicado algoritmos de aprendizado de máquina supervisionado nos cursos presenciais.

Através de uma abordagem supervisionada, a partir dos dados e de um atributo que classifique cada registro é selecionado uma amostra para treinar o modelo baseado na classe e outra para avaliar através de métricas que comparem o valor real da classe e o valor predito.

2. Metodologia

O presente trabalho utilizou o processo de KDD (do inglês, *Knowledge-Discovery in Databases*) proposto por Fayyad et al. (1996) para extrair conhecimento a partir dos dados do SIGAA, exibido na Figura 1.

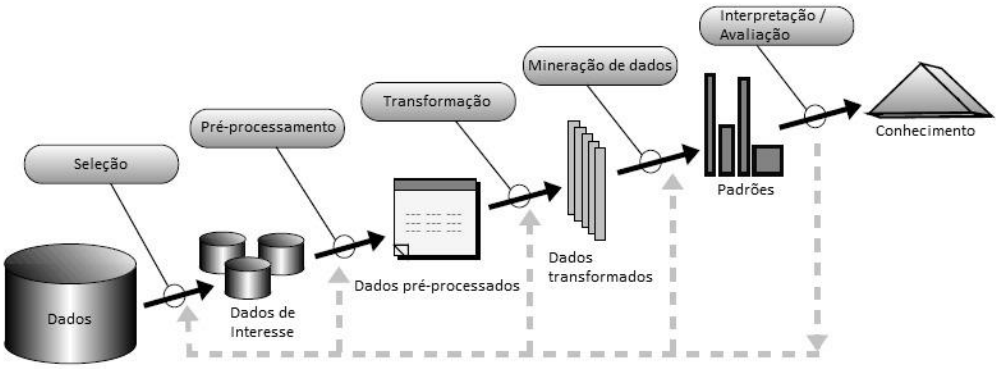

Figura 1: Esquema do Processo de KDD proposto por Fayyad et al. (1996).

A primeira etapa, seleção, trata-se de uma triagem das informações relevantes para o processo, ou seja, aqueles dados que não possuem importância para o alcançar o objetivo são removidos da base. Atributos de identificação do registro e atributos únicos, como CPF e número de matrícula, são exemplos de informações desconsideradas.

Logo após, no pré-processamento, pretende-se melhorar a qualidade da base de dados. “Informações ausentes, errôneas ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de KDD” (BOENTE; GOLDSMIDT; ESTRELA. 2006). Remoção de atributos com muitos valores ausentes, utilização de regressão para preenchimento de informações ausentes, busca e retirada de valores incoerentes são exemplos de correções aplicadas.

A seguir, na transformação, são feitas alterações nos dados para melhor adequá-los, facilitando etapas posteriores. Excluir, mesclar ou aplicar funções em atributos são exemplos de operações que podem ser empregadas. Mais especificamente para nosso caso, foi feita uma discretização na propriedade alvo da classificação, neste caso o IRA (Índice de Rendimento Acadêmico), transformando valores contínuos que variam entre zero e dez para valor 0 quando abaixo de sete (exclusive) e valor 1 caso contrário, criando assim o atributo classe desempenho acadêmico com valor 0 quando não satisfatório e 1 quando satisfatório.

Passadas as etapas anteriores, agora na mineração de dados, é possível aplicar algoritmos de aprendizado de máquina para obter padrões e correlações entre as características acadêmicas do universitário adquiridas no SIGAA e o desempenho obtido através do IRA como exposto anteriormente.

Durante a mineração, assumindo o desempenho acadêmico como atributo classe, foi utilizado o método “*DecisionTreeClassifier*” da biblioteca *scikit-learn*. Segundo Géron (2017), a biblioteca implementa o algoritmo CART (do inglês, *Classification And Regression Tree)*. Dado que o algoritmo CART representa uma árvore de decisão, temos que “Uma árvore de decisão é uma estrutura de árvore semelhante a um fluxograma, onde cada nó interno (nó não folha) denota um teste em um atributo, cada ramo representa um resultado do teste, e cada nó folha (ou nó externo) contém um rótulo de classe. O nó mais alto de uma árvore é o nó raiz” (HAN; KAMBER; PEI. 2012).

Ao final da aplicação do algoritmo, o modelo geral de toda a universidade obteve 78.68% de acurácia, isto é, 78.68% de acerto ao comparar os valores previstos com os reais, além variar entre 78.52% a 81.93% ao analisar individualmente cada centro e campus. Comparando com o resultado encontrado por SILVA (2017), percebe-se uma queda de 15.82% no modelo geral e uma diferença entre 15.98% e 12.57% nos modelos específicos, baseado nas métricas finais utilizadas no artigo citado. É uma redução aceitável dada a maior diversidade nos registros e atributos da base de dados utilizada por este trabalho.

Por último, através de todos os resultados obtidos, na fase de interpretação e avaliação, é possível inferir padrões que identificam um futuro desempenho acadêmico de um aluno como não satisfatório (valor 0) ou satisfatório (valor 1). Utilizar imagens e textos que representam a estrutura de árvore facilitou o processo, através da identificação visual das regras definidas em cada nó.

3. Resultados

Analisando alunos durante o primeiro semestre da graduação de diversos campi e centros, em geral, eles têm baixo rendimento acadêmico. Porém, ao observar de forma específica cada centro e campus, encontram-se peculiaridades neles. Carga horária total do curso, grau acadêmico (licenciatura ou bacharelado) e a diferença entre ingresso no ensino superior e conclusão do ensino médio são atributos que variam sua influência na classificação dos alunos nos campi e centros da universidade. Apesar disso, alunos que tenham aproveitado disciplinas de alguma graduação estão propensos a ter bom rendimento, variando pouquíssimo o mínimo de 314.75 horas integralizadas exigido pelo modelo geral para classificar o discente assim.

Campus Ministro Reis Velloso (CMRV), Campus Senador Helvídio Nunes de Barros (CSHNB) e Centro de Ciências da Natureza (CCN) são casos bem nítidos da importância da carga horária no desempenho acadêmico durante o primeiro semestre, exemplificados no Quadro 1. Expresso no quadro, cursos do CCN com carga horária total até 2617.5 horas ou superior a 3007.5 horas costumam ter melhores desempenhos que aqueles em faixas diferentes de carga horária.

Estudantes do CMRV de cursos com carga horária entre 2872.5 (exclusive) e 3585 (inclusive) horas ou superior a 3817.5 horas são propensos a um desempenho satisfatório. Nas demais cargas horárias, é semelhante ao comportamento geral da universidade. Já no CSHNB, cursos com carga horária superior a 3885 horas pedem a ter melhores desempenhos classificados.

Quadro 1: Desempenho acadêmico para alunos do primeiro período de cursos do CCN

```
if Integralizada <= 959.1847839355469:
    if Tempo decorrido no curso <= 3.5:
        if Integralizada <= 599.9184875488281:
            if Tempo decorrido no curso <= 1.5:
                if Integralizada <= 297.7989196777344:
                    if Integralizada <= 0.6521739363670349:
                        if CH total <= 3007.5:
```

```
                    if CH total <= 2520.0:
                        Não satisfatório
                    else:
                        if CH total <= 2617.5:
                            Satisfatório
                        else:
                            Não satisfatório
                else:
                    Satisfatório
(...)
```

A diferença entre os anos de ingresso na graduação e de término do ensino médio possui grande peso na performance acadêmica de educandos do Centro de Tecnologia (CT) durante o primeiro período. Como exposto no Quadro 2, quando o resultado dessa subtração não ultrapassa 1.35, aproximadamente, classifica-se o desempenho como satisfatório.

Quadro 2: Desempenho acadêmico para alunos do primeiro período de cursos do CT

```
if Ingresso - Conclusão Ensino Médio <= 1.3461538553237915:
    if Integralizada <= 763.6956481933594:
        if Tempo decorrido no curso <= 1.5:
            if Integralizada <= 382.5:
                if Tempo decorrido no curso <= 0.5:
                    Satisfatório
    (...)
else:
    if Integralizada <= 1754.7554321289062:
        if Tempo decorrido no curso <= 4.5:
            if Integralizada <= 772.5:
                if Tempo decorrido no curso <= 1.5:
                    if Integralizada <= 352.5:
                        Não satisfatório
(...)
```

Também há campi e centros que são influenciados por múltiplos atributos, como o Centro de Ciências da Saúde (CCS), Campus Amilcar Ferreira Sobral (CAFS), Centro de Ciências Humanas e Linguagens (CCHL) e Centro de Ciências Agrárias (CCA), exemplificados no Quadro 3 pelo CCS. Nota-se no quadro que nesse centro para classificar um desempenho acadêmico como satisfatório no semestre inicial é necessário

que o curso tenha mais de 3877.5 horas totais e o discente tenha no máximo 4.15 anos desde a conclusão do ensino médio.

No CAFS, durante o primeiro período, alunos em cursos de bacharelado com carga horária entre 3090 (exclusive) e 3165 (inclusive) horas tendem a ter um melhor desempenho, já estudantes de licenciatura tendem a ter notas ruins, independente da carga horária.

No CCHL, a exigência para classificar seu desempenho como satisfatório é a carga horária ser superior a 3015.0 horas e que tenha passado no máximo 2.69 anos desde sua conclusão do ensino médio.

No CCA, educandos do primeiro semestre da graduação classificados com desempenho satisfatório nesse centro estão em cursos com mais de 4425 horas totais e não tem mais 1.2 anos desde o término do ensino médio.

Quadro 3: Desempenho acadêmico para alunos do primeiro período de cursos do CCS

```
if CH total <= 3877.5:
    if Integralizada <= 1795.0271606445312:
        if Tempo decorrido no curso <= 4.5:
            if Integralizada <= 1372.5:
                if Integralizada <= 404.75543212890625:
                    if Tempo decorrido no curso <= 0.5:
                        Não satisfatório
    (...)
else:
    if Integralizada <= 1437.7988891601562:
        if Tempo decorrido no curso <= 3.5:
            if Integralizada <= 433.4510803222656:
                if Tempo decorrido no curso <= 0.5:
                    if Ingresso - Conclusão Ensino Médio <= 4.148351669311523:
                        if CH total <= 6547.5:
                            Satisfatório
                        else:
                            Satisfatório
                    else:
                        Não satisfatório
(...)
```

4. Conclusão

Ao analisar centros e campi separadamente, na definição desses perfis encontra-se influência da carga horária total no curso no CMRV, CSHNB, CCN, CAFS, CCHL, CCS e CCA, da espera do ingresso no ensino superior desde o fim do ensino médio no CT, CCHL, CCS e CCA e do grau acadêmico CAFS. Apesar de os campi e centros terem perfis bem diferentes, excetuando CCA e CCE todos possuem algo constante, alunos que devido a graduações anteriores, integralizaram a partir de um determinado valor tem desempenho classificado como satisfatório, esse valor raramente difere muito 314.75 horas identificados pelo modelo geral. Nos períodos seguintes as classificações passam a ser consequência dos semestres anteriores.

Considerando métricas elevadas na avaliação do modelo criado e apontada na metodologia a comparação com outro trabalho semelhante realizado, entende-se que os resultados e as análises elaboradas a partir deles são satisfatórias para o objetivo deste trabalho, pois tendem a condizer com a realidade. Além de, geralmente, apresentar regras bem nítidas. Sendo assim, chegou-se bem próximo de uma conclusão ideal para o que foi proposto.

5. Referências

BOENT, A.N.P; GOLDSCHIMIDT, R.R; ESTRELA, V.V. **Uma metodologia de suporte ao processo de conhecimento em bases de dados**. In V SIMPÓSIO DE EXCELÊNCIA EM GESTÃO E TECNOLOGIA, 2008. Resende (RJ). Disponível em: www.boente.eti.br/publica/seget2008kdd.pdf. Acesso em: 19 ago. 2019.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. (1996). **From Data Mining to Knowledge in Databases**. AI Magazine, Menlo Park, v.17, n.3, mar. 1996. Disponível em: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230. Acesso em: 19 ago. 2019.

GOLDSCHMIDT, R. R.; PASSOS, E. **Data Mining: Um Guia Prático**. Rio de Janeiro: Elsevier, 2005.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn and TensorFlow**: CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLINGENT SYSTEMS. 1. Ed. Sebastopol: O'Reilly Media, 2017.

HAN, J.; KAMBER, M.; PEI, J. **DATA MINING: Concepts and Techniques**. 3. ed. Waltham: Elsevier, 2012.

SILVA, A. M. L**. Descoberta de Conhecimento através de Métodos de Aprendizagem de Máquina Simbólicos aplicados ao Ensino a Distância da Universidade Federal do Piauí**. 2018. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Piauí, Teresina, 2018. Disponível em: http://repositorio.ufpi.br/xmlui/bitstream/handle/123456789/1506/Descoberta%20de%20Conhecimento%20atrav%C3%A9s%20de%20M%C3%A9todos%20de%20Aprendizagem%20de%20M%C3%A1quina%20Simb%C3%B3licos%20aplicados%20ao%20.pdf?sequence=1. Acesso em: 19 ago. 2019.

SILVA, A. M. L, et. al. **DESCOBERTA DE CONHECIMENTO ATRAVÉS DE MÉTODOS DE APRENDIZAGEM DE MÁQUINA SUPERVISIONADOS APLICADOS AO SIGAA/UFPI.** Revista de Sistemas e Computação, Salvador, v.7, n.1, jan/jun. 2017. Disponível em: https://revistas.unifacs.br/index.php/rsc/article/view/4953. Acesso em: 19 ago. 2019.