Relatório PIBIC

Wladimir Ganzelevitch Mesquita Gramacho
Orientador: Guilherme Novaes Ramos

December 2018

Resumo

A evasão em Instituições de Ensino Superior (IES) é um grande desafio atualmente, visto que altas taxas de evasão prejudicam a gestão eficiente de recursos nas universidades, sendo necessárias políticas de redução desse problema. Dessa forma, este trabalho tem como objetivo avaliar se é possível usar técnicas de Mineração de Dados para classificação do perfil de alunos em risco de evasão. A partir do histórico dos dois primeiros semestres do estudante, conseguimos obter classificadores com acurácia geral de até 80%.

1 Introdução

A evasão de estudantes nas universidades tem sido motivo de muita discussão e análise entre educadores, gestores e outros agentes da Educação [10] [13]. Nas universidades públicas, as perdas de estudantes que iniciam mas não terminam seus cursos são desperdícios de recursos públicos: sociais, acadêmicos e econômicos [26]. Esses recursos são desperdiçados porque o investimento feito na formação dos ingressantes não é convertido em profissionais formados se tais estudantes evadem ao longo do curso. Objeto de nosso estudo, a Universidade de Brasília possui uma taxa média anual de evasão de aproximadamente 10% [9]. Conseguir prever se um estudante evadirá ou não é importante para poder auxiliar gestores no apoio à diminuição da evasão na Universidade. Descobrir manualmente os elementos que mais influenciam nessa evasão pode se tornar uma tarefa muito difícil e onerosa, mas podemos aplicar modelos de mineração de dados para classificar tais alunos. Este trabalho tem o intuito de mostrar a construção e os resultados da aplicação de métodos de aprendizagem de máquina à evasão de estudantes nas universidades. Dessa forma, nossa hipótese é de que o resultado de um aluno num curso da universidade (formatura ou evasão) pode ser previsto a partir do histórico dele nas matérias cursadas.

A partir de dados dos históricos de estudantes da Universidade de Brasília, processamos os dados, transformamo-los e aplicamos modelos de aprendizagem de máquina neles. Considerando os trabalhos correlatos e seus resultados, neste trabalho utilizamos Árvores de Decisão, Redes Neurais e Regressão Logística como métodos de aprendizagem de máquina. Seguindo a metodologia *Knowledge Discovery Process* (KDP), obtemos conhecimento a partir dos dados. Juntamente com o esclarecimento de cada etapa do KDP, são apresentados os processos e os resultados desenvolvidos ao longo deste trabalho. Por fim, apresentamos os resultados obtidos e comentamos sobre os próximos passos desta pesquisa.

2 Trabalhos Correlatos

Existem diversos trabalhos que se propõem a definir evasão [6] [25] [28], sendo o desenvolvido por Tinto [27] a principal referência nos trabalhos correlatos a este [3] [4] [11]. Sua ideia é a de que o processo de evasão de um aluno é determinado por questões individuais, familiares, sociais, acadêmicas e estruturais (da instituição).

Aulck, Velagapudi, Blumenstock e West construíram classificadores (Regressão Logística, Florestas Aleatórias e k-NN) com acurácias por volta dos 65% analisando somente o primeiro semestre do estudante [4]. Nandeshwar analisou o primeiro ano cursado e observou que informações sobre o contexto familiar e situação sócio-econômica dos estudantes ajuda a identificar melhor os padrões para classificação das entradas [23]. Na Índia, Saurabh Pal usou Naïve Bayes e obteve modelos com mais de 90% de sensibilidade ao analisar o tipo de bairro que o estudante mora e fatores sócio-econômicos como ocupação e nível de escolaridade dos pais [24]. No Brasil, Lucas de Assis estudou evasão no Ensino Superior em relação a cursos, áreas de estudo e Instituições de Ensino Superior e obteve o melhor desempenho (a nível de curso) usando o algoritmo de Árvores de Classificação e Regressão [3]. Delen combinou os históricos de Ensino Médio e Universidade de alunos do primeiro ano para treinar os classificadores e obteve acurácias de 81%, 78% e 74% para Redes Neurais, Árvores de Decisão e Regressão Logística, respectivamente [12].

Observando os resultados destes trabalhos correlatos, utilizaremos neste trabalho os algoritmos de Árvores de Decisão, Regressão Logística e Redes Neurais pois apresentam os melhores resultados e podem trazer conhecimento importante para o usuário final. Vamos avaliá-los em relação à sensibilidade porque assim podemos qualificar a habilidade dos classificadores de encontrar o máximo de evasores possível. Ademais, utilizamos somente os dados dos históricos dos estudantes no seu primeiro ano de curso, desconsiderando para este trabalho informações sócio-econômicas dos estudantes e seus contextos familiares.

3 Fundamentação Teórica

Nesta etapa apresentamos alguns outros conceitos preliminares, como Mineração de Dados e os algoritmos que vamos usar, além de quais métricas foram avaliadas.

3.1 Mineração de Dados e Aprendizagem de Máquina

Mineração de Dados (MD) é definido como o processo de descobrir padrões (automaticamente ou semi automaticamente) a partir de dados [30]. Com os padrões encontrados, fazemos previsões dos nossos dados de forma a obter conhecimento. O processo de mineração de dados faz proveito de algoritmos de aprendizagem de máquina para conseguir obter esses padrões de forma automatizada.

Aprendizagem de Máquina, ou *Machine Learning*, é a área de estudo que pesquisa como computadores conseguem aprender (ou melhorar seu desempenho) baseados em dados, fazendo com que algoritmos reconheçam automaticamente padrões complexos a partir de dados e tomem decisões em relação a esses dados [15]. Quando falamos em Aprendizagem de Máquina, sempre temos que levar em consideração o tipo de problema que temos em mãos e os objetivos que queremos alcançar [22]. Existem três tipos de Aprendizagem de Máquina: aprendizagem supervisionada, não supervisionada e por reforço [21]. Nosso problema trata-se de aprendizagem supervisionada, visto que sabemos o resultado da classificação. Aprendizagem de Máquina supervisionada é a busca por algoritmos que raciocinam a partir de instâncias fornecidas externamente para produzir hipóteses gerais, que então fazem previsões sobre instâncias futuras [21]. Neste trabalho, utilizaremos algoritmos de aprendizagem para classificar os alunos em evadidos e formados.

Tais classificadores devem ser genéricos, que não são sofrem de sobre-ajuste nos dados. Sobre-ajuste é o uso de modelos ou procedimentos que incluem mais termos do que os necessários ou usam abordagens mais complicadas do que as necessárias [16]. Logo, sobre-ajuste ocorre quando treinamos em demasia o nosso classificador e este se torna um classificador perfeito para os dados de treinamento, mas não são generalizados o suficiente para outros dados de entrada [22].

3.2 Classificadores utilizados

Basicamente, uma árvore de decisão é um algoritmo que funciona a partir de comparações *if-then-else*, onde tem-se como objetivo dividir o conjunto de observações pela metade, a fim de minimizar o tamanho da árvore. A partir das respostas desse conjunto de questões, Árvores de Decisão permitem que você classifique um indivíduo em "sucesso" ou "fracasso" [18], ou, levando em consideração nosso problema, em "evadido" ou "formado". Este modelo, usado tanto para classificação como regressão, tem se tornado bastante popular com o avanço da Mineração de Dados [12]. Além disso, após o treino e construção da árvore, podemos analisar cada nó (a comparação do nó e a divisão realizada dos dados) para obter mais conhecimento sobre quais características mais influenciam na trajetória do nosso aluno, por exemplo.

Regressão Logística é uma abordagem para prever o resultado de uma variável dependente categorial baseada em uma ou mais variáveis observadas [1]. Esse algoritmo destina-se a prever a probabilidade de que um evento ocorra. As probabilidades que descrevem as possíveis classificações de uma dada entrada são definidas usando uma função logística, que segue uma curva em forma de "S". A partir do ponto onde a entrada é traçada nessa função logística, uma classe é definida. Após treinar o modelo, ele pode ser usado para prever o sucesso (no caso, evasão) de novos indivíduos [18].

Redes Neurais são redes inspiradas na Biologia, com técnicas analíticas altamente sofisticadas e capazes de modelar funções não-lineares extremamente complexas [12]. Formalmente, uma rede neural é um processador distribuído e massivamente paralelo feito de unidades de processamento simples (*perceptrons*) que são propensos a armazenar conhecimento e disponibilizá-lo para uso [17]. Usaremos neste trabalho um *perceptron* multicamadas (MLP), semelhante ao perceptron mas com mais de uma camada de neurônios que são alimentados por "sinapses" com pesos. A partir das entradas (características), o algoritmo constrói uma rede com pesos utilizando o algoritmo

de *backpropagation*, onde o erro encontrado na saída é propagado para a rede de forma a calibrar os pesos da rede e minimizar o erro.

4 Metodologia e Resultados

O *Knowledge Discovery Process* é um modelo de processo que foi estabelecido pela academia para fazer com que projetos de mineração de dados tivessem uma metodologia direcionada à resolução de problemas. O processo consiste em uma série de etapas iterativas e com loops de feedbacks para que o problema seja resolvido incrementalmente de forma contínua. Neste trabalho, utilizaremos o modelo apresentado por Fayyad em 1996 [14], que possui 9 etapas. Cada etapa será apresentada juntamente com o desenvolvimento e os resultados deste trabalho.

4.1 Entendimento do Domínio

A primeira etapa do KDP é entender as nuances do problema que estamos abordando e como podemos ajudar o usuário final com o conhecimento que obtivermos [8], a fim de diminuir a evasão universitária. Altas taxas de evasão [5] [29] prejudicam a imagem da universidade em relação à sociedade, além de causarem perdas de recursos financeiros e humanos [26]. Tendo isso em vista, nosso objetivo com este trabalho é o de identificar padrões para classificar se alunos irão evadir ou não de um determinado curso na universidade.

4.1.1 Evasão

Podemos separar evasão no ensino superior em diferentes níveis: em termos da Instituição de Ensino Superior (IES), em termos da área de estudo, em termos do curso e da matéria [3]. Neste trabalho, vamos focar na evasão a nível de curso, pois queremos que os classificadores sejam usados por coordenadores de curso. Dito isso, usaremos o curso de Engenharia Mecatrônica da Universidade de Brasília para realizar nossos experimentos e colher os resultados. Essa escolha se deu pelo fato de termos os históricos dos alunos desse curso.

Um detalhe importante é o de analisar o desempenho nos dois primeiros semestres dos estudantes pois, dessa forma, os alunos podem ser auxiliados e orientados de forma a prevenir possíveis evasões no terceiro semestre. Na Figura 1 temos a quantidade de evasões por semestre em relação ao fluxo do estudante, de forma que cada coluna i representa o número de estudantes que evadiu no seu i-ésimo semestre. Podemos observar que o número de alunos de Engenharia Mecatrônica que evadem no seu terceiro semestre é o maior entre todos os semestres do curso. Logo, analisar o histórico do primeiro ano de um estudante no curso é relevante para a prevenção da evasão.

Figura 1: Número de evasões por semestre

4.2 Seleção dos dados

Nesta etapa, investigamos os dados de interesse para nosso problema e tentamos encontrar informações que nos ajudem a modelá-los de forma a criar entradas que possam ser classificadas. Para tal, é necessário extrair os dados de interesse para nosso trabalho.

4.2.1 Extração e Seleção

O coordenador do curso de Engenharia Mecatrônica (que orientou este trabalho) obteve os dados do histórico acadêmico dos alunos do curso. Tais dados foram descaracterizados de modo que não fosse possível identificar um aluno. A Tabela 1 apresenta uma descrição das características que obtivemos dos dados e que usaremos para a construção dos modelos de entrada.

Nº	Variável	Tipo	Descrição
1	`IDAluno`	Número	Identificador do aluno
2	`AnoIngresso`	Número	Número de 5 dígitos que representa o semestre de ingresso. Os 4 primeiros dígitos são o ano de ingresso e o último é o semestre
3	`AnoMateria`	Número	Número de 5 dígitos que representa o semestre que o estudante cursou a matéria. Os 4 primeiros dígitos são o ano e o último é o semestre.
4	`CodigoMateria`	Número	Código da matéria que o aluno cursou em um dado semestre
5	`Conceito`	String	Menção do aluno ao cursar aquela matéria naquele semestre
6	`StatusFinal`	String	Situação final do aluno no curso

Tabela 1: Descrição das variáveis dos dados

4.3 Pré-processamento

Na 3ª etapa, o objetivo é limpar os dados e estruturá-los para que possamos criar nossos modelos de entrada. Modelos de entrada são uma forma de estruturar os dados que servirão de insumo para que os classificadores encontrem padrões.

A partir dos dados, o aluno será considerado como evadido caso não seja listado como formado, nem como ativo (declarado como aluno regular). Os alunos que vieram a falecer durante o curso são ignorados pois não concluíram o curso mas também não evadiram. Portanto, temos as seguintes possíveis situações finais daquele aluno:

1. "FORMADO": aluno formado;
2. "EVADIDO": aluno que não está ativo, nem formado.

Após fazer o recorte por somente alunos não ativos da Engenharia Mecatrônica, obtemos 1177 matrículas diferentes que ingressaram entre o 2º semestre de 1997 e o 1º semestre de 2018, sendo 629 deles considerados formados e 548 evadidos.

4.3.1 Modelos de entrada

Vamos inserir dois modelos de entrada nos nossos classificadores, pois queremos avaliar diferentes níveis de especificidade em relação à avaliação dos estudantes nas matérias. Como não há nenhum estudo atualmente sobre fatores que levam à evasão no curso de Engenharia Mecatrônica na UnB, construímos esse dois modelos de forma arbitrária e hipotética de modo a ter uma base comparativa para futuras análises de outros modelos. A coluna `Conceito` tem os seguintes possíveis valores e seus respectivos equivalentes em notas de 0 a 10: SS (9,0 a 10), MS (7,0 a 8,9), MM (5,0 a 6,9), CC (Crédito Concedido, sem nota vinculada), MI (3,0 a 4,9), II (0,1 a 2,9), SR (zero), TR (trancamento parcial de matéria, não conta como reprovação nem aprovação), TJ (trancamento justificado da matéria, similar ao TR). No Modelo 1, teremos somente a informação se o aluno aprovou ou reprovou cada matéria obrigatória do fluxo de Engenharia Mecatrônica nos dois primeiros semestres. Dessa forma, consideramos SR, II e MI como reprovação e todos os outros como não reprovação. O Modelo 2 é semelhante ao Modelo 1 exceto que usaremos a informação exata do conceito do aluno naquela disciplina, ou seja, a menção que o estudante obteve na disciplina.

4.4 Transformação dos dados

Agora que temos dados limpos e normalizados, podemos transformá-los para que cada observação dos modelos de entrada represente um estudante, de forma que este possa ser classificado. O objetivo dessa transformação é ter toda a informação do histórico de um aluno em somente uma linha no modelo de entrada. Isto posto, a primeira transformação que se faz necessária é a projeção de uma nova coluna chamada `Semestre` que será o resultado do cálculo (usando as colunas `AnoIngresso` e `AnoMateria`) do semestre em que o aluno cursou determinada matéria. Dessa forma, filtramos pelos dados onde o `Semestre` é 1 ou 2, correspondendo às matérias cursadas no 1º e 2º semestre do estudante. Além disso, filtramos pelos códigos das matérias obrigatórias de Engenharia Mecatrônica, segundo a definição do fluxo do curso [1].

Então, concatenamos as colunas `Semestre` e `Materia` gerando uma coluna `Semestre_Materia`. Após isso, transpomos os dados de forma que uma linha corresponda a um único aluno. Nesse processo, surgem as colunas para cada valor de `Semestre_Materia` existente. Essas colunas seguem o formato `S_MMMMMM` onde `S` corresponde ao semestre no qual a matéria foi cursada e `MMMMMM` é o código da matéria, por exemplo `1_113034` representa Cálculo 1 (código 113034) no 1º semestre. Por fim, uma linha no banco de dados transformado tem a estrutura:

IDAluno	StatusFinal	1_113034	1_118001	...	2_118036	2_168874

Para o Modelo 2, faz-se necessário mapear cada possível menção para uma nova coluna. Para cada coluna de `Semestre_Materia`, teremos novas colunas seguindo o formato `Semestre_Materia_Menção`, onde teremos a informação se um aluno tirou aquela menção naquele semestre para aquela matéria. Por exemplo, se um estudante tirou SS em Cálculo 1 no seu 1º semestre, seu valor na coluna `1_113034_SS` será 1. Se o aluno não cursou, mantendo o exemplo, Cálculo 1 em seu 1º semestre, ele terá o valor 1 na coluna `1_113034_NC`.

4.4.1 Análise dos Dados

Após fazer as transformações, passa-se à análise estatística dos dados com mais facilidade, visto que cada observação nos modelos de entrada representa um único estudante. Fazemos esta análise a fim de ter um entendimento qualitativo dos dados e possibilitar o cruzamento dessa análise com os resultados dos classificadores.

Após filtrar pelos dados dos dois primeiros semestres, temos 1133 alunos, sendo que 617 são formados e 516 são evadidos. Para cada estudante, avaliamos as menções das 13 matérias que estão nos dois primeiros semestres do fluxo de Engenharia Mecatrônica. Em média, cada estudante tem 9.78 menções gravadas no banco de dados. A moda e a mediana são iguais a 11 matérias cursadas.

Apenas 17.56% (199) do total de estudantes reprovam Cálculo 1 no primeiro semestre. Por outro lado, considerando somente os alunos evadidos, essa proporção cresce para 34.30% (177). Em relação à Física 1, 20.21% (229) de todos os estudantes reprovam no 1º semestre. Levando em conta novamente somente os evasores, 37.01% (191) reprovam a matéria. Essas duas matérias têm os maiores índices de reprovação entre os evadidos.

A menção mais comum é MS, com 36.22% do total de aparições nas menções obtidas, seguida de MM com 30.30%. Para os alunos que cursaram Desenho Mecânico Assistido por Computador 1 no 2º semestre (719), 60.36% (434) deles tiraram MS. Ao observar essa matéria em relação a menções aprovantes, temos que a proporção de aprovados ao cursar essa matéria nesse período é de 96.55%. Se fizermos essa mesma análise para a matéria Introdução à Álgebra Linear (código 113093), temos que 82.10% dos alunos que cursaram tal matéria conseguiram a aprovação nela no segundo semestre. Se formos mais específicos e analisarmos essa matéria em relação aos estudantes que evadiram, 58.41% conseguiram uma nota de aprovação ao cursar a matéria no segundo semestre.

4.5 Mineração dos Dados

Aqui resumimos as etapas 5, 6 e 7. Na etapa 5, definimos nosso tipo de problema como sendo de classificação e na etapa 6 definimos nossos classificadores (Árvores de Decisão, Redes Neurais e Regressão Logística). Na etapa 7, aplicamos de fato os classificadores. Usamos a biblioteca Scikit-learn [2], da linguagem de programação Python para a utilização dos algoritmos de classificação. Além disso, a partir das implementações do Scikit-learn, usamos validação cruzada e *grid search*.

Vamos utilizar a técnica de validação cruzada para generalizar os classificadores, ou seja, para a prevenção de sobre-ajuste. O âmago da validação cruzada está no particionamento dos dados em subconjuntos mutualmente ex-

[1] `https://matriculaweb.unb.br/graduacao/fluxo.aspx?cod=6912`
[2] `https://scikit-learn.org/`

clusivos e de mesmo tamanho, que serão usados para treinamento e teste dos classificadores. Seguindo as referências de trabalhos correlatos e os resultados obtidos por Kohavi [20], utilizamos a validação *10-fold* estratificada.

Para cada tipo de classificador, faremos *grid search* em alguns de seus parâmetros. *Grid search* é a definição de alguns valores para parâmetros de um classificador e a avaliação de todas as combinações possíveis a fim de obter a melhor configuração avaliada de cada classificador dada uma métrica [7].Para Árvores de Decisão e a partir da implementação do Scikit-learn, os parâmetros que vamos ajustar no algoritmo são: critério de ganho de informação, estratégia de divisão nos nós, máxima profundidade da árvore e quantidade mínima de exemplares para fazer a divisão do nó. Os parâmetros que vamos ajustar na execução da Regressão Logística são: algoritmos para o problema de otimização e o nível de regularização. No algoritmo de Redes Neurais, vamos ajustar os parâmetros: algoritmo para o problema dos pesos das camadas, quantidade de penalidade nos erros e algoritmo de ativação para a camada oculta.

4.5.1 Métricas

Um dos objetivos deste trabalho é servir de insumo para que coordenadores de cursos de graduação na UnB possam identificar quais são os alunos que têm risco de evadir e poder auxiliá-los e reverter essa situação. Portanto, é necessário definir as métricas dos nossos classificadores. Nosso foco é o de que, dentre as pessoas que irão evadir, consigamos identificar o maior número delas.

Sensibilidade (*recall*) é a habilidade do classificador de encontrar todas as instâncias relevantes para o problema. Dessa forma, sensibilidade será a proporção de entradas que foram classificadas corretamente como positivas dentre todas as entradas que de fato são positivas. Sua fórmula é dada por:

$$sensibilidade = \frac{VP}{VP + FN}$$

onde VP é o número de verdadeiros positivos. No nosso problema, isso representa os alunos que foram classificados como evadidos que realmente evadiram. FN é o número de falsos negativos, ou seja, número de instâncias que foram classificadas erroneamente como formados e que de fato evadiram. Como queremos ajudar coordenadores de curso e gestores da educação a tomarem providências em relação à evasão de estudantes, queremos achar todos os alunos que vão, de fato, evadir, maximizando a sensibilidade dos classificadores.

Outra métrica que vale a pena observar é a acurácia. Acurácia é a capacidade do modelo de conseguir classificar corretamente as entradas em relação aos seus valores verdadeiros. Dessa forma, a fórmula da acurácia se dá por:

$$acurácia = \frac{VP + VN}{VP + FP + FN + VN}$$

onde VN representa o número de verdadeiros negativos e FP é o número de falsos positivos.

4.6 Avaliação e Interpretação

Após execução dos algoritmos com *grid search*, apresentamos aqui os resultados dos melhores classificadores de cada algoritmo para cada modelo de entrada. Além de analisá-los em relação à sensibilidade, observamos também a acurácia dos algoritmos. Com o objetivo de observar a consistência dos classificadores, executamos os experimentos 10 vezes e obtivemos as seguintes médias e seus desvios-padrão:

	sensibilidade		acurácia	
	Modelo 1	Modelo 2	Modelo 1	Modelo 2
Árvores de Decisão	64.71 % ± 0.00 %	**71.76% ± 0.0%**	75.13 % ± 0.00 %	**80.21% ± 0.0%**
Redes Neurais	56.62 % ± 1.01 %	70.09 % ± 1.95 %	73.91 % ± 0.51 %	72.03 % ± 3.23 %
Regressão Logística	59.41 % ± 0.00 %	68.82 % ± 0.00 %	69.41 % ± 0.00 %	79.41 % ± 0.00 %

Tabela 2: Resultados dos classificadores em relação aos modelos de entrada

Em todos os resultados, exceto em relação à acurácia do classificador de Redes Neurais, o Modelo 2 obtém melhores resultados que o Modelo 1. Além disso, podemos observar pelos resultados que Árvores de Decisão se mostram como o melhor classificador tanto em termos da sensibilidade quanto em termos da acurácia, ao encontrar 71.76% dos estudantes que irão de fato evadir e conseguir classificar os estudantes em evadidos e formados com mais de 80% de acurácia. Como não há trabalhos relacionados ao uso de classificadores para evasão nos cursos da UnB, estes classificadores possibilitam a identificação mais acurada dos evasores.

Na Figura 2 podemos ver parte da estrutura da Árvore de Decisão ao utilizar o Modelo 2 como entrada. Apesar de Cálculo 1 e Física 1 terem os maiores índices de reprovação entre os evasores, o desempenho nessas matérias não é avaliado nos três primeiros nós da árvore. A primeira divisão se dá ao observarmos se o estudante cursou a disciplina Física Experimental 2 (código 118036) no 2º semestre. Se o estudante tiver cursado, seu valor na coluna `2_118036_NC` será zero e ele seguirá para o ramo da esquerda, tendo aproximadamente 74.22% de probabilidade de formar. Se o valor nessa coluna for 1, ou seja, o estudante não tiver cursado Física Experimental 2 no 2º semestre, segue para o ramo da direita e provavelmente evadirá (aproximadamente 78.89%). Dessa maneira, coordenadores do curso de Engenharia Mecatrônica podem acompanhar os estudantes de forma a ter uma maior atenção para os que não a tenham cursado no 2º semestre.

No segundo nível da árvore, no ramo da esquerda temos a pergunta de se o estudante teve MI (nota de reprovação) como menção em Introdução a Álgebra Linear (código 113093) no segundo semestre. No ramo da direita, a pergunta se trata de avaliar se o estudante obteve MS (nota de aprovação) em Desenho Mecânico Assistido por Computador 1 (código 168874) também no segundo semestre. Dessa forma, podemos observar que as três primeiras perguntas feitas na estrutura da árvore de decisão são relacionadas ao segundo semestre do curso de Engenharia Mecatrônica. Com isso, gestores do curso de Engenharia Mecatrônica podem criar alertas em caso de alunos que se encontrem no ramo da direita para poder auxiliá-los e evitar prováveis evasões.

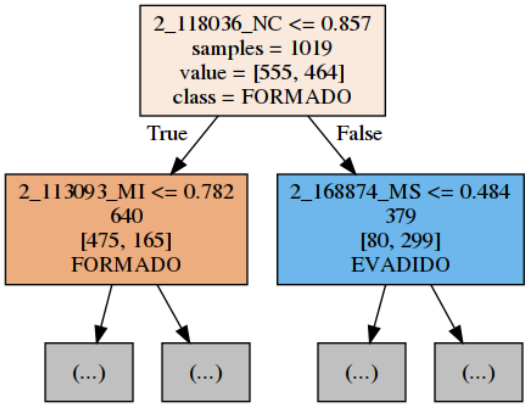

Figura 2: Parte da Árvore de Decisão do Modelo 2

4.7 Consolidando o conhecimento

Na última etapa, o pesquisador deve reportar e documentar o conhecimento adquirido durante o processo inteiro de forma a atingir o objetivo inicial, além de, se possível, disponibilizar os classificadores para proveito do usuário final. Desta maneira, a aplicação dos classificadores está disponível em um site [3] para que qualquer um possa, a partir dos dados dos históricos de estudantes, usar os resultados dos classificadores para tomar ações em relação à evasão em seu curso. O código e a documentação deste estão disponíveis em um repositório público [4].

5 Conclusão

Evasão nas universidades é um tema de preocupação de diversos atores da Educação Superior. A partir da hipótese de que podemos classificar se um aluno irá evadir a partir do seu histórico no curso, obtivemos classificadores com mais de 80% de acurácia. Dessa forma, concluímos que este trabalho atingiu resultados consideráveis em relação à evasão de alunos no curso de Engenharia Mecatrônica, além de não haver ferramenta assim disponível. A partir deste trabalho, avaliamos se é possível utilizar classificadores para o problema da evasão nos cursos e disponibilizamos os classificadores, de forma que gestores de educação e coordenadores de cursos de graduação possam investir tempo no auxílio de estudantes em risco de desligamento.

Este trabalho foi o início de um projeto de pesquisa acerca da classificação da evasão nas universidades. Outros classificadores podem ser analisados, como os algoritmos Floresta Aleatória e Apriori. Além disso, podemos avançar na otimização de parâmetros dos classificadores, a fim de investigar se é possível melhorar os resultados obtidos

[3] `https://evasion-app.herokuapp.com/`
[4] `https://github.com/wladimirgramacho/evasion_unb`

neste trabalho. Podemos também gerar outros modelos de entrada para testar novas hipóteses, como avaliar se o aluno teve reprovação, aprovação ou trancamento na matéria ou, desconsiderando o semestre no qual foi cursado a matéria, avaliar quantas reprovações o aluno teve por matéria.

Outro caminho de desenvolvimento deste projeto é o de explorar outras informações que estejam disponíveis nos dados da UnB. Analisar a forma de ingresso e que tipo de escola estudou podem trazer muito conhecimento para os modelos, como visto em trabalhos correlatos [2] [24]. Podemos também analisar o gênero dos estudantes. Por exemplo, no curso de Computação, a retenção das alunas tem diminuído nos últimos anos e podemos fazer uma análise desse perfil a fim de aumentar a retenção [19]. Ademais, podemos usar o CEP para calcular a distância do estudante até a universidade e ver se essa característica nos ajuda a classificar melhor os alunos. Ainda, podemos computar o IRA (Índice de Rendimento Acadêmico) parcial de cada semestre do indivíduo, de forma a ter uma noção de como está o rendimento geral do estudante a cada semestre.

Referências

[1] Charu C Aggarwal. *Data classification: algorithms and applications*. CRC press, 2014.

[2] Sattar Ameri, Mahtab J Fard, Ratna B Chinnam, and Chandan K Reddy. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 903–912. ACM, 2016.

[3] Lucas Rocha Soares de Assis. Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados. 2017.

[4] Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*, 2016.

[5] Marta F Barroso and Eliane BM Falcão. Evasão universitária: o caso do instituto de física da ufrj. *IX Encontro Nacional de Pesquisa em Ensino de Física*, 9:1–14, 2004.

[6] John P Bean. Interaction effects based on class level in an explanatory model of college student dropout syndrome. *American educational research journal*, 22(1):35–64, 1985.

[7] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[8] Krzysztof J Cios, Witold Pedrycz, Roman W Swiniarski, and Lukasz Andrzej Kurgan. *Data mining: a knowledge discovery approach*. Springer Science & Business Media, 2007.

[9] Comissão Própria de Avaliação UnB. Relatório final de autoavaliação institucional 2019. `http://www.dpo.unb.br/index.php?option=com_phocadownload&view=category&download=862:relatorio-de-autoavaliacao-institucional-2018&id=91:autoavaliacao-institucional&Itemid=674`. (Acessado em 08/07/2019).

[10] Decanato de Planejamento e Orçamento. Relatório de gestão de 2018. `http://www.dpo.unb.br/images/phocadownload/documentosdegestao/relatoriogestao/2018/Relatrio_de_Gesto_UnB_2018.pdf`. (Acessado em 09/06/2019).

[11] Dursun Delen. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506, 2010.

[12] Dursun Delen. Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, 13(1):17–35, 2011.

[13] Cristiane Aparecida dos Santos Baggi and Doraci Alves Lopes. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior*, 16(2), 2011.

[14] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.

[15] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[16] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[17] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[18] Anne-Sophie Hoffait and Michaël Schyns. Early detection of university students with potential difficulties. *Decision Support Systems*, 101:1–11, 2017.

[19] Maristela Holanda, Marília Dantas, Gustavo Couto, Jan Mendonça Correa, Aleteia Patrícia F de Araújo, and Maria Emília T Walter. Perfil das alunas no departamento de computação da universidade de brasília. In *11º Women in Information Technology (WIT 2017)*, volume 11. SBC, 2017.

[20] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[21] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[22] Stephen Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2011.

[23] Ashutosh Nandeshwar, Tim Menzies, and Adam Nelson. Learning patterns of university student retention. *Expert Systems with Applications*, 38(12):14984–14996, 2011.

[24] Saurabh Pal. Mining educational data using classification to decrease dropout rate of students. *arXiv preprint arXiv:1206.3078*, 2012.

[25] Ernest T Pascarella. Student-faculty informal contact and college outcomes. *Review of educational research*, 50(4):545–595, 1980.

[26] Roberto Leal Lobo Silva Filho, Paulo Roberto Motejunas, Oscar Hipólito, and Maria Beatriz Carvalho Melo Lobo. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, 37(132):641–659, 2007.

[27] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1):89–125, 1975.

[28] Vincent Tinto and Brian Pusser. Moving from theory to action: Building a model of institutional action for student success. *National Postsecondary Education Cooperative*, pages 1–51, 2006.

[29] Tereza Christina MA Veloso and Edson Pacheco de Almeida. Evasão nos cursos de graduação da universidade federal de mato grosso, campus universitário de cuiabá–um processo de exclusão. *Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB*, (13), 2013.

[30] Ian H Witten, Eibe Frank, and Mark A Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.