APLICAÇÃO DE ALGORITMO DE *MACHINE LEARNING* NA IDENTIFICAÇÃO DE ALUNOS EM RISCO DE EVASÃO

Rafael Baldasso - rafaelobaldasso@gmail.com

Marcelo Cortimiglia - cortimiglia@gmail.com

RESUMO

Este trabalho é um estudo da aplicação de ferramentas de *machine learning em um modelo* de predição de evasão de alunos no contexto de um curso pré-vestibular online. O objetivo é a identificação das principais variáveis indicativas de uma possível evasão, assim como prever quais alunos estão em risco de evadir, tendo como base os dados gerados pelo aluno após quatro semanas de utilização da plataforma de ensino. Pautado na metodologia KDD e se valendo do método *random forest*, trabalhou-se os dados de cadastro e comportamentais dos alunos que utilizaram a plataforma como principal ferramenta de estudo por, pelo menos, cinco meses. Obteve-se como resultado uma lista das variáveis mais influentes na previsão da evasão, assim como um modelo com acurácia de 79% de previsão, identificando 66% dos alunos que vieram a evadir. Conclui-se que o estudo gerou uma ferramenta com capacidade preditiva significativa para o curso pré vestibular, assim como informações úteis e novas em relação às variáveis relevantes na predição da evasão.

Palavras-chave: *machine Learning*[1], árvore de decisão[2], *radom forest*[3], evasão[4], pré-vestibular[5], *knowledge discovery in databases*[6], *educational data mining*[7]

1. INTRODUÇÃO

De acordo com números disponibilizados pelo INEP, a taxa anual de evasão no ensino superior brasileiro tem se mantido em um patamar de 22% entre os anos de 2001 e 2017, sendo esse um problema que afeta grande parte das instituições de ensino. Além disso, o fenômeno da evasão escolar é visto como um dos maiores problemas de qualquer nível de ensino (Lobo, 2012). Sendo assim, a busca de suas causas tem sido objeto de muitos trabalhos e pesquisas educacionais no contexto nacional e internacional, devido à difusão do problema tanto na esfera pública quanto privada (Silva Filho, 2007; Souza, 2012). No ambiente de educação online, em especial, o fenômeno da evasão é ainda mais prevalente: segundo Carr (2000), a evasão na educação à distância (EAD) é de 10% a 20% maior do que em cursos presenciais.

A evasão gera consequências negativas para a manutenção da qualidade do ensino e da saúde financeira das instituições de ensino. Silva Filho et al. (2007) reforçam tal impacto, afirmando que a evasão gera perdas sociais, acadêmicas e econômicas. Para a instituição, o fenômeno acarreta ociosidade e perda de credibilidade. Para os estudantes, a evasão pode representar o atraso ou cancelamento de um sonho, perda de oportunidades de trabalho, de crescimento pessoal e de melhoria de renda, entre outras consequências.

Segundo Muilenburg e Berge (2005), existem oito fatores principais na predição da permanência de alunos em um curso EAD ou um *massive open online course* (MOOC): habilidades acadêmicas e técnicas, motivação, tempo e estímulo para estudos, custo, acesso a internet e problemas técnicos. Apesar da aparente simplicidade destes oito fatores, o grande volume de dados envolvidos em sua constituição torna complexa a análise sobre a compreensão e prevenção da evasão estudantil.

Nesse contexto, Attaran (2018) afirma que *softwares* de análise preditiva são uma das principais ferramentas que podem ser utilizadas para adquirir informação que gere ação a partir dos dados disponíveis. Resumidamente, análise preditiva é a área de mineração de dados preocupada em prever probabilidades e tendências. Um modelo preditivo é baseado em um número de dados e fatores que potencialmente influenciam um comportamento futuro. A saída do modelo é uma lista de fatores que prevê, com certo grau de confiança, o resultado de um evento (Society of Actuaries, 2012).

Segundo a Forrester (2017), a área de análise preditiva terá considerável crescimento nos próximos anos chegando a 11,5% de taxa de crescimento anual composta no período de 2016 a 2021. Esse crescimento pode ser atribuído à ascensão de ferramentas de *Machine Learning* (ML), que se diferenciam dos outros métodos de análise existentes por dois grandes motivos: I) ML realiza predições e calibra modelos em tempo real e automaticamente - enquanto outras ferramentas tendem a perder acurácia conforme mais dados são inseridos; II) ML não precisa de intervenção externa para identificar testar e validar associações de causa e efeito (Kumar, 2018).

Apesar de existirem estudos que abordam os motivos que levam um aluno a evadir, há poucos estudos na literatura nacional com o objetivo de gerar modelos que auxiliem as instituições de ensino na detecção de grupos com risco de evasão por meio de ML (Tontini e Walter, 2014). Com isso em mente, esse artigo busca elaborar um critério pelo qual seja possível se caracterizar a evasão de alunos no contexto do curso pré-vestibular MeSalva! (MS), desenvolver um modelo

preditivo utilizando técnicas de ML para identificar o subconjunto de alunos do MS que se enquadram na definição de evasão elaborada e, por fim, identificar os principais fatores que indicam o risco de um aluno evadir. Não há atualmente um modelo de identificação no MS, não havendo nenhum tipo de processo de reconhecimento e categorização de alunos em grupos de risco, nem ao menos uma definição clara do que é um aluno evadido. Portanto, o desenvolvimento de um modelo preditivo é de grande importância para evitar futuras evasões do curso, pois identificando os padrões de comportamento que sugerem a possibilidade de evasão, o MS, a partir do primeiro mês de estudos do aluno, pode tomar ações preventivas a fim de reduzir este fenômeno.

Este artigo é composto de cinco partes. Após esta introdução está a revisão da literatura, onde os tópicos que abrangem a elaboração de um modelo preditivo com ML, assim como pesquisas já realizadas na área de evasão escolar, serão apresentados. A terceira seção descreve o método empregado no trabalho e a quarta seção, os resultados obtidos. A seção final contempla a conclusão do artigo, onde serão apresentados os principais aprendizados.

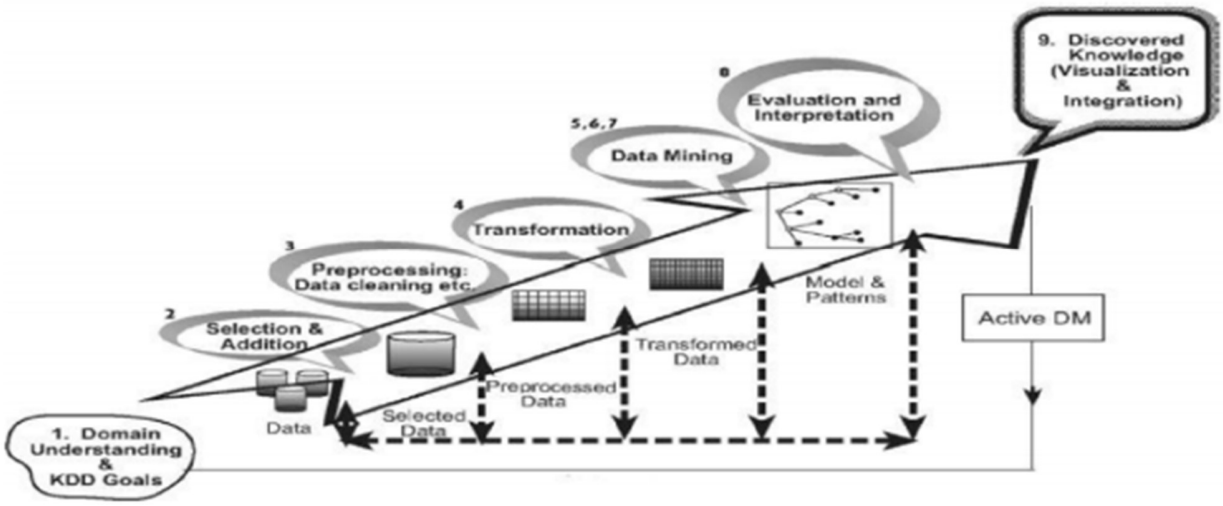
2. REFERENCIAL TEÓRICO

2.1. KDD / Mineração de dados

Diversos autores propuseram modelos de *data mining* (DM) para guiar a sua aplicação em bancos de dados. Entre eles os mais notáveis são o Knowledge Discovery in Databases (KDD), CRISP-DM e SEMMA, os quais têm nove, seis e cinco etapas, respectivamente. Em uma análise comparativa dos métodos, Shafique e Qaiser (2014) concluíram que os três métodos são equivalentes, mas com especificações em diferentes níveis de detalhe. Contudo, a maioria dos estudos realizados e grande parte das aplicações na indústria seguem o modelo (KDD), pois este foi o primeiro a ser desenvolvido e difundido em grande escala.

O KDD foi definido por Fayyad *et al* (1996) como um processo não trivial de identificar compreensíveis, previamente desconhecidos e potencialmente úteis padrões em dados. Não trivial enfatiza a complexidade envolvida neste tipo de problema. A natureza de ser previamente desconhecido implica que o processo deve gerar informações não existentes previamente no sistema. Finalmente, potencialmente úteis indica que a nova informação não será apenas ruído, mas de fato algo relevante para o analista (Prass, 2007). A figura 1 apresenta uma esquematização do processo de KDD para melhor entendimento.

Figura 1 - O ciclo do processo de KDD.



Fonte: Rokach e Maimon (2015)

Segundo Rokach e Maimon (2015), a primeira das nove etapas do modelo KDD é o desenvolvimento e compreensão do domínio de aplicação, que se trata de uma etapa preparatória ao método. Neste momento devem ser definidos quais os objetivos com a elaboração do KDD, assim como a delimitação do ambiente onde o processo de descoberta de informações será executado.

Da segunda à quarta etapa o modelo entra na macro etapa de pré-processamento dos dados, iniciando com a criação do banco de dados onde as descobertas serão feitas. Nesta etapa, devem-se identificar os dados disponíveis, assim como seus possíveis cruzamentos e extensões. Todos os dados devem, então, ser compilados em um banco de dados. Esse processo é de grande importância, pois todo o KDD se baseará nestes dados iniciais, e caso haja uma lacuna de informações relevantes, a validade do modelo como um todo pode ser comprometida. A terceira etapa, nomeada limpeza, busca garantir a robustez do modelo por meio da remoção de amostras incoerentes como ruídos, *outliers* e registros incompletos, garantindo que o modelo fundamentará suas análises apenas em amostras coerentes com as intenções do desenvolvedor. A quarta etapa, a de transformação dos dados, tem o objetivo de atribuir um formato padrão aos dados. Tal processo deve ser feito com grande cuidado, para evitar qualquer transformação que possa afetar as informações que estes dados carregam. Realizadas as primeiras quatro etapas de pré-

processamento, as seguintes quatro etapas têm como foco a definição dos aspectos dos algoritmos a serem implementados no projeto, valendo-se de ferramentas de ML.

A quinta etapa tem o papel de definir a estratégia mais apropriada para a mineração de dados. Munido das informações das etapas prévias e com um objetivo de projeto estabelecido, essa etapa definirá a tarefa principal do modelo, que predominantemente é de classificação, regressão ou clusterização. A sexta parte, escolha do algoritmo de mineração de dados, consiste em um desdobramento da quinta. A definição do algoritmo dependerá da importância dada a fatores como precisão e compreensibilidade da resposta, qualidade e amplitude de dados de entrada, entre outros. A sétima etapa é a execução do algoritmo. Nesse estágio é comum aplicar o algoritmo escolhido diversas vezes até se atingir resultados satisfatórios. Essa iteração é feita concomitantemente com a alteração de parâmetros de controle a fim de se obter o resultado mais acurado possível. A oitava etapa tem como objetivo a avaliação e interpretação dos resultados obtidos em contraste aos objetivos delimitados na primeira etapa. Além disso, esse estágio busca garantir a compreensibilidade e utilidade do modelo, assim como sua documentação.

Finalmente, a nona e última etapa, o uso do conhecimento descoberto, analogamente à primeira, é praticamente uma etapa pós análise, pois consiste na utilização das informações geradas pelo modelo fora de seu ambiente controlado de criação. Esta etapa determina a eficácia do processo como um todo, pois o KDD só tem validade uma vez que o conhecimento gerado é de fato aplicado para a geração de algum valor.

2.2. *Machine Learning* (ML)

ML é considerado uma evolução da adaptação estatística de modelos e, como seu antecessor, tem como objetivo a extração de informação de um banco de dados através de um modelo probabilístico. A evolução do ML se encontra na automatização de grande parte do processo (Baldi e Brunak, 2001). ML pode ser definido como o método pelo qual se faz um algoritmo modificar sua estrutura a fim de adquirir maior acurácia no atingimento de um objetivo previamente definido (Marsland, 2014).

Métodos de ML podem ser segmentados em dois grandes grupos: Supervisionado e não supervisionado. A principal diferença entre os métodos se dá na necessidade de uma rotulação prévia dos dados do supervisionado, o que não é necessário no não supervisionado. Contudo, essa praticidade de não rotular os dados vêm ao custo de uma menor adaptabilidade e menor controle

sobre o desenvolvimento do modelo. Algoritmos não supervisionados são majoritariamente utilizados com fins de clusterização, já os modelos supervisionados, por outro lado, podem ser segmentados em dois principais grupos: Regressão e categorização, sendo o de categorização o mais utilizado e com maior amplitude de aplicações (Rokach e Maimon, 2015).

Apesar do grande interesse em definir o algoritmo mais apropriado para cada situação, na literatura não existe nenhuma regra determinística para sua escolha - essa definição ainda é fortemente dependente de iterações e experiência do analista. Apesar disso, existem algumas heurísticas que podem ser utilizadas para facilitar o processo de definição. Segundo Michalski *et al* (1984), para a escolha do algoritmo, deve-se pensar em alguns fatores como o formato das variáveis do modelo (discretas/contínuas, alta/baixa amplitude), formato das saídas do modelo e fatores ligados a falhas na base de dados (categorização incorreta ou ausência de atributos em parte das amostras).

Dentre os diversos algoritmos existentes, alguns apresentam menor complexidade que outros. Um modelo simples é a regressão logística, que consiste de um modelo estatístico de análise de dados que funciona a partir de uma base com uma ou mais variáveis independentes. As saídas do algoritmo são calculadas a partir da análise de variáveis dicotômicas, buscando identificar qual categoria se enquadra melhor aos valores das variáveis, assim classificando a base de dados. Outro modelo de classificação é o *naive bayes,* que parte da premissa de que todas as variáveis de predição são independentes, ou seja, qualquer interação entre estas variáveis é ignorada mesmo que exista, o que simplifica drasticamente a construção e execução do modelo. Por fim, outro algoritmo com grande aderência é o de redes neurais, baseado em camadas de nodos que transformam vetores de entrada em vetores de saída, corrigindo o peso dessas transformações a cada iteração a fim de adaptar a rede neural ao problema selecionado (Marsland, 2014).

Árvore de decisão, um dos modelos mais difundidos em ML, trata-se de um fluxograma em formato de árvore que classifica informações numéricas ou simbólicas, baseado nos valores dos atributos de decisão. Ou seja, a árvore de decisão analisa as entradas por meio de múltiplas funções e as categoriza a partir de algum critério (Sharma *et al*., 2013). O algoritmo faz isso através da criação de nós, ramos e folhas que representam funções, critérios de categorização e categorias de resultado, respectivamente (Rokach e Maimon, 2015).

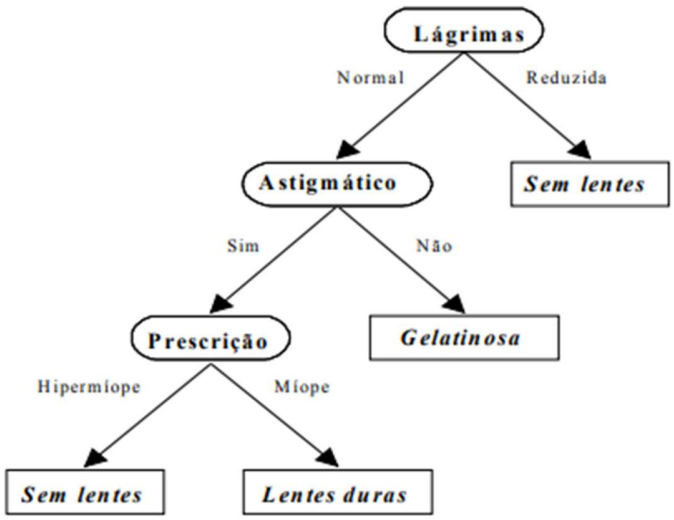
Segundo Dechter e Michie (1985), apesar de sistemas de classificação por árvore de decisão aparentarem servir apenas um pequeno segmento de análises, diversos problemas podem

ser reestruturados a fim de torná-los problemas de classificação. Para ilustrar a utilização em diversos campos de pesquisa: Giasson (2011) utiliza árvores de decisão para o mapeamento digital de solos, Oña (2012) na identificação de fatores que contribuem para a gravidade em acidentes de trânsito e Lee *et al*. (2011) no reconhecimento do estado emocional através da voz.

Segundo Sharma *et al.* (2013), a construção de uma árvore de decisão pode ser segmentada em duas etapas: fase de crescimento e fase de poda. A fase de crescimento representa a etapa em que a árvore é construída, o que inclui a criação de múltiplos níveis de nós, ramos e folhas a fim de maximizar a acurácia de precisão do modelo. Em seguida, múltiplas subárvores são criadas, cada uma delas omitindo um nó, a fim de verificar se sua remoção aumenta ou reduz a taxa de erro. Caso a remoção reduza ou não a altere essa taxa, a nova árvore seguirá sem este nó, pois este provavelmente foi gerado devido a ruídos na base de dados de treinamento. A remoção destes nós, que nomeia a fase, reduz a chance de o algoritmo incorrer em problemas de *overfitting* (adaptação excessiva do modelo aos dados de treino), garantindo assim uma melhor adaptabilidade do modelo a novas entradas de dados.

Um algoritmo de árvore de decisão muito utilizado é o *Classification Regression Tree* (CART). Esse método é característico por sua construção de árvores binárias, ou seja, cada nó tem exatamente dois ramos. O interessante do CART é que ele proporciona a possibilidade de gerar uma árvore de regressão, nas quais as folhas preveem valores contínuos (Rokach e Maimon, 2015). A figura 2 apresenta um modelo de árvore de decisão CART, onde as elipses (nós) testam as entradas através das linhas (ramos), para então, categorizar as entradas em um dos retângulos (folhas). Nessa árvore de decisão, o objetivo é a prescrição correta de lente de contatos a partir da quantidade de lágrimas, presença de astigmatismo e miopia ou hipermetropia.

Figura 2 - Exemplo de árvore de decisão



Fonte: Witten e Frank. (2000)

Algoritmos CART obtiveram grande sucesso devido à sua excelência em interpretação, visualização e habilidade de lidar com grandes complexidades na não linearidade entre as entradas e a resposta (Friedman et al, 2001). Contudo, este sistema ainda sofre de problemas como *overfitting*, pouca robustez a *outliers* e um mau desempenho preditivo quando comparado a outros modelos (Breiman, 2001). Uma estratégia para combater esses problemas é agrupar o resultado de diversas árvores de decisão, o que veio a ser denominado método Random Forest (RF).

.

2.2.1. *Random forest*

RF é um método de aprendizado por agrupamento utilizado para classificação e regressão. Em um modelo de RF, a resposta de cada uma das árvores é levada em consideração e a saída definida como correta será a mais votada em um caso de classificação ou média das respostas em uma regressão. O método RF é mais resiliente a *overfitting*, tem maior precisão e consegue lidar com bancos de dados maiores do que uma árvore de decisão simples (Breiman, 2001). Pode-se dizer que a principal vantagem do RF se dá com a redução de impacto dos ruídos nos dados, pois é menos provável que múltiplas árvores de decisão cometam o mesmo erro, devido ao modo como as diferentes árvores são geradas.

Para garantir a diversidade entre as árvores, o modelo RF aplica duas principais estratégias. Primeira, comum em diversos outros modelos, é denominada *bagging* e consiste em treinar as diferentes árvores em bancos de dados ligeiramente distintos. A segunda estratégia consiste na variação dos atributos dos registros fornecidos a cada árvore. A limitação dos atributos providos às árvores faz com que cada uma delas tenha que elaborar uma lógica distinta para maximizar sua acurácia de resposta a partir dos dados que possui. Um exemplo dessa estratégia seria prover a um modelo de previsão de chuva dados sobre a data do ano, mas nenhum dado sobre a umidade do ar. (Rokach e Maimon, 2015)

2.3. Aplicações de *Data Mining* e *Machine Learning* na Educação

Apesar da mineração de dados ser aplicada com sucesso em diversos setores da indústria, a aplicação de DM no contexto educacional *educational data mining* (EDM) ainda é limitado. Contudo, pesquisadores vem provando que é possível aplicar princípios de mineração de dados em ambientes de educação ricos em dados (Gandhimathi e Gomathi, 2015).

Hussain *et al*. (2018) utilizaram ferramentas de ML para identificar alunos com baixa participação em um curso à distância com o objetivo de auxiliar o corpo docente a analisar as atividades e utilização de informações por parte dos discentes. Já Shawky e Badawi (2018) elaboraram um modelo que, além de adaptar o ensino aos alunos, auxilia os docentes a melhorar os conteúdos, atividades e referências utilizando informações sobre utilização e aprendizado dos alunos.

Um estudo realizado por Beck e Woolf (2000) resultou em um modelo capaz de prever o tempo necessário, assim como a acurácia de resposta dos alunos ao resolver problemas de matemática, o que é conhecido como modelagem de alunos, e no ambiente educacional tem diversas aplicações. Um exemplo de sua utilidade é a detecção de características como satisfação, motivação e aprendizado, assim como possíveis problemas no desenvolvimento do aluno, como repetidos erros, baixa utilização de ajuda, tentativas de burlar o sistema ou até ineficiências na utilização dos recursos da disciplina. O estudo de Syed *et al*. (2017) levou o processo de modelagem um passo à frente e utilizou essas informações para gerar recomendações de tarefas e conteúdo mais apropriados aos alunos, de acordo com a etapa de aprendizado que eles se encontram.

Por fim, uma das áreas de pesquisa mais pertinentes no EDM é a predição de desempenho, que tem como objetivo predizer, a partir do perfil do aluno, suas notas finais nas disciplinas e outros resultados, como evasão do curso e futura habilidade de aprendizado. Um ponto interessante desta área é a relevância da utilização de métodos que ofereçam a possibilidade de identificar os fatores críticos na obtenção do resultado, o que facilita a geração de ações corretivas pelas instituições de ensino.

Dentro da área de predição de resultados, um dos mais prevalentes e importantes tópicos de estudo é a previsão de evasão escolar. Pesquisas nesta área comumente se limitam a analisar cursos EAD e resultados em uma só disciplina, devido a praticidade na obtenção de dados. Contudo, pouquíssimos estudos se valem de ferramentas de ML, carência essa que tende ser gradualmente atendida devido ao crescente número de publicações de EDM registradas nos últimos anos (Machado et al., 2015).

Apesar da pequena quantidade de estudos se valendo de ferramentas de ML, dentre os exemplos existentes, boa parte obteve um bom resultado. Manhães *et al*. (2011) utilizou diversos algoritmos de ML na mineração de dados de alunos de graduação da UFRJ, identificando com precisão de 80% a situação final dos alunos no curso. Delen (2010), em um estudo utilizando uma base de dados de mais de cinco anos, conseguiu uma acurácia de aproximadamente 80% na predição de evasão dos calouros utilizando quatro algoritmos distintos. Por fim, Tontini e Walter (2014) identificaram os fatores que influenciam na evasão dos alunos e elaboraram um modelo que atingiu resultados notáveis: 53% dos alunos que se evadiram no semestre subsequente foram identificados anteriormente pelo modelo. Essas análises foram encaminhadas à instituição de ensino superior, que tomou ações preventivas simples, como o estabelecimento de contato com os alunos em busca de entender suas dificuldades, resultando em 18% de redução na taxa de evasão no período subsequente.

Mais especificamente no contexto das MOOCs, a quantidade de estudos nacionais sobre evasão utilizando ML é limitada. Fato que se deve a novidade das plataformas MOOCs assim como dos métodos de análise em ML. Contudo, vem crescendo o volume de pesquisas internacionais publicadas neste tema, com estas se valendo de diferentes algoritmos de predição e obtendo boa acurácia preditiva (Sinha et al, 2014). Xing e Du (2018) geraram um modelo preditivo para a evasão dos alunos de um curso online de gestão de projetos com oito semanas de duração atingindo 95% de acurácia com base em dados comportamentais apenas da primeira semana.

Sharkey e Sanders, ao analisarem um curso de psicologia da Georgia Tech/Coursera, alcançaram 88% de acurácia preditiva utilizando um algoritmo de RF com 15 variáveis para identificar se o aluno assistiria a semana seguinte de aulas se baseando somente em sua atividade na semana passada.

3. METODOLOGIA

3.1. Cenário

Este trabalho foi realizado com base nos dados do MS, plataforma de cursos online fundada em 2012 com sede em Porto Alegre/RS. O MS possui mais de 10.000 vídeos que abrangem cinco grandes áreas: reforço escolar, Exame Nacional do Ensino Médio (ENEM) & vestibular, engenharia, negócios e saúde. Para fins desta pesquisa foi analisada somente a área ENEM & vestibular que pode, por sua vez, ser segmentada em grupos menores, dos quais os seguintes são os mais relevantes em quantidade de alunos: medicina extensivo, medicina semiextensivo, ENEM extensivo, ENEM semiextensivo. Devido à maior quantidade de dados e riqueza de informações, estes quatro grupos serão os únicos analisados nesta pesquisa. Os cursos possuem uma carga horária de aproximadamente 600h, segmentada em 14 núcleos de estudo definidos conforme o escopo da prova do ENEM: Matemática, física, biologia, química, geografia, história, sociologia, filosofia, redação, língua portuguesa, literatura, inglês/espanhol, artes e educação física. Existem também três núcleos interdisciplinares: Ciências da natureza, ciências humanas e linguagens. cada um deles compreende uma parcela da carga horária estipulada, que pode ser dividida conforme o aluno julgar apropriado, dentro do período disponível de preparação para a prova.

Cada aluno do MS fornece e gera diversos dados, que foram utilizados neste estudo como insumo para consolidar uma base de dados sobre o perfil e comportamento dos estudantes. Esta base de dados, após ser pré-processada, serviu para a identificação dos principais fatores de influência na permanência ou evasão dos alunos, assim como na identificação dos grupos de risco. Nesta base serão considerados apenas alunos matriculados com mais de cinco meses disponíveis para preparação e que declararam o MS como principal ferramenta de estudo, de modo a garantir um perfil do qual se possa esperar um engajamento mais elevado e constante. Com isso se reduziu a base de 18.000 alunos para 2.225 alunos, os quais foram caracterizados a partir de 30 variáveis, definidas junto aos funcionários do MS, listadas na Tabela 1.

Tabela 1 - Variáveis do estudo

Variável	Opções	Significado
user_id	-	Key única do usuário
categoria	Básico, Completo e Medicina	Pacote comprado
city	-	Cidade do aluno
tipo_compra	Boleto	Modo de pagamento da compra
regiao	Sudeste, Norte, Centro-Oeste, Sul, Nordeste, não informado	Região do aluno
reg_metropolitana	Região Metropolitana, Interior, não informado	Categoria da região do aluno
semanas_enem	20 a 54	Semanas de antecedência a prova ENEM
etapa_estudos	Colégio, Terceiro, Ensino Médio Concluído, Faculdade, Formado, Pós	Nível de escolaridade do aluno
principal_ferramenta	0 ou 1	Confirmação que o aluno está utilizando MS como principal ferramenta de estudo
curso_desejado	-	Curso que o aluno deseja entrar
publica	0 ou 1	Confirmação se aluno deseja entrar em uma faculdade pública
privada	0 ou 1	Confirmação se aluno deseja entrar em uma faculdade privada
foco_enem	0 ou 1	Confirmação se aluno está dedicado exclusivamente para a prova ENEM

metodo_estudo	Múltipla escolha: Ainda não achei a melhor forma de estudar; Pratico com exercícios fora do Me Salva!; Costumo pausar os vídeos para fazer anotações e resumos pessoais; Minha principal fonte de estudos são as aulas; Revejo algumas aulas mais de uma vez para fixar o conteúdo.	Principal método de estudo empregado pelo aluno
internet_acess	1 a 5	Nível de acesso a internet, 5 sendo o melhor
study_platform	Pelo Computador/Notebook, Pelo Celular, Pelo Tablet	Principal plataforma de estudo utilizada
percentual_consumo_ios	0 a 1	Porcentagem do uso da plataforma a partir de um dispositivo IOS
total_semanas_ativo	20 a 54	Semanas ativo desde o primeiro login até o ENEM
pct_semanas_ativo	0 a 1	Porcentagem de semanas ativo desde o primeiro login até o ENEM
pct_cons_manha	0 a 1	Porcentagem das utilizações da plataforma no turno da manhã
pct_cons_tarde	0 a 1	Porcentagem das utilizações da plataforma no turno da tarde
a_s1	0 a ∞	Aulas assistidas na primeira semana de uso
a_s2	0 a ∞	Aulas assistidas na segunda semana de uso
a_s3	0 a ∞	Aulas assistidas na terceira semana de uso
a_s4	0 a ∞	Aulas assistidas na quarta semana de uso
ex_s1	0 a ∞	Exercícios realizados na primeira semana de uso
ex_s2	0 a ∞	Exercícios realizados na segunda semana de uso

ex_s3	0 a ∞	Exercícios realizados na terceira semana de uso
ex_s4	0 a ∞	Exercícios realizados na quarta semana de uso
fez_simulado	0 ou 1	Confirmação se o aluno fez o simulado

3.2. Método de pesquisa

A criação de um modelo preditivo em *ML* da evasão dos alunos do MS se caracteriza como uma pesquisa de natureza aplicada, pois é dedicada à geração de conhecimento para solução de um problema específico (Nascimento e Sousa, 2016). A pesquisa possui uma abordagem quantitativa devido à ênfase na utilização de dados para identificar as possíveis causas de evasão. Além disso, em relação aos objetivos de pesquisa, o estudo é classificado como explicativo, na medida em que tem como principal foco identificar os fatores que determinam e contribuem para a evasão. (GIL, 2008).

3.3. Método de trabalho

A estrutura dos resultados deste artigo será dividida em três etapas. A primeira consistirá na definição dos critérios a serem utilizados para categorizar uma evasão na base de dados, seguida de duas macro etapas fundamentadas no KDD, cunhado por Fayyad *et al*. (1996). Das nove etapas do modelo de Fayyad serão realizas todas menos a última, pois esta acarretaria uma necessidade de envolvimento mais intenso do MS, o que não é viável para este projeto devido a limitação de tempo. Além disso, a primeira etapa não será discutida na seção de resultados, pois sua execução faz parte na definição de objetivos da pesquisa, etapa já realizada neste artigo. A abordagem, então, consistirá de duas macro etapas: I. Pré-processamento de dados; II. Aplicação e avaliação do algoritmo de ML.

A etapa de pré-processamento dos dados tratará de três sub-etapas: Criação do banco de dados; Limpeza dos dados; e Transformação dos dados. Na criação do banco de dados será realizada uma revisão das informações fornecidas pelo MS para garantir que não há nenhum outro registro em relação ao desempenho ou perfil dos alunos que possa ser acrescido à base de dados exemplificada na tabela 1. Uma vez validada a base, será feita a limpeza dos dados, removendo todos os alunos que apresentam comportamento incompatível com o perfil a ser analisado, como a não utilização da plataforma como método de estudo principal ou aquisição de acesso a plataforma com tempo insuficiente para consumo do conteúdo disponível, definido como 5 meses

pelos funcionários do MS. Também serão removidos os registros incompletos devido ao não preenchimento dos formulários iniciais de definição de perfil.

A última etapa, consistirá na aplicação de um método supervisionado de categorização devido a qualidade de adaptação desse tipo de modelo ao problema desta pesquisa. Para tanto, serão utilizados princípios de árvore de decisão através de um modelo de categorização baseado em RF, algoritmo escolhido devido à sua resiliência a bases de dados com ruídos, boa capacidade preditiva e capacidade de indicar a importância das variáveis envolvidas na predição dos resultados. Todas essas análises serão realizadas em código Python devido ao fácil acesso a conhecimento nesta ferramenta e sua transparência, por ser código aberto, ao contrário de ferramentas *black-box* como SPSS. Para facilitar as análises, será utilizada a biblioteca scikit-learn, a mais popular biblioteca direcionada para aprendizado de máquina em Python.

Por fim, será identificada a precisão do modelo e os principais fatores indicativos de uma possível evasão. Além disso, serão discutidas as novas informações geradas pelo modelo KDD, sua validade, utilidade e novidade.

4. RESULTADOS

4.1. Definição dos critérios de evasão

Juntamente com funcionários do MS foi elaborada uma definição do que seria considerado uma evasão, visto que a empresa não tinha um critério definido para isso. A definição elaborada para este estudo consiste no atendimento de pelo menos um de dois critérios:

I. Aluno esteve ativo em menos de 50% de semanas;

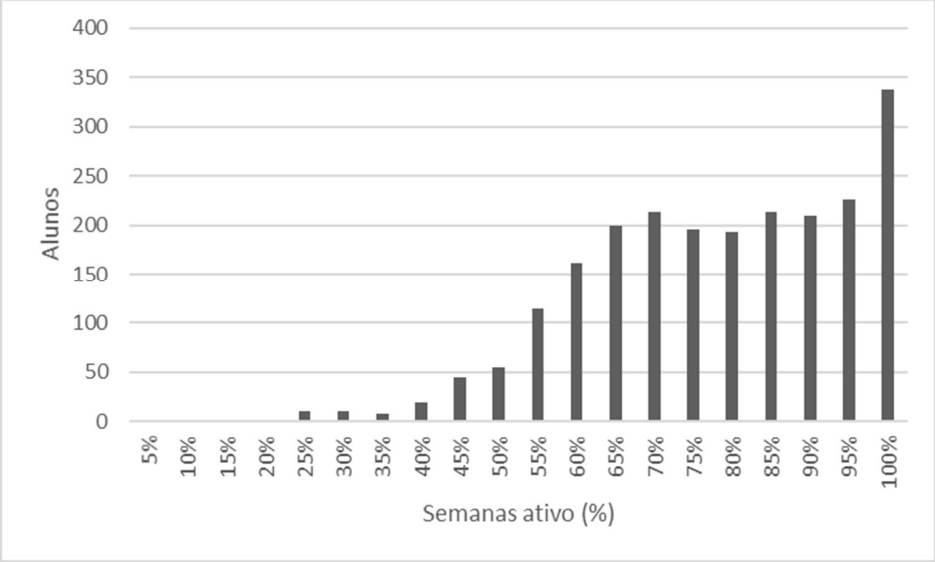
II. Aluno esteve inativo no último mês precedente à prova do ENEM.

Entende-se atividade como a realização de pelo menos um exercício ou a visualização de pelo menos um vídeo. Estes critérios, então, classificaram a base em 2.022 alunos que permaneceram e 183 alunos que evadiram.

O primeiro critério foi introduzido para identificar os alunos que não se comprometeram ao curso ao longo do ano e, portanto, não utilizaram o curso conforme o previsto pelo MS. A figura 3 demonstra a distribuição dos alunos por porcentagem de semanas ativo ao longo do ano a fim de ilustrar a grande gama de alunos com atividade abaixo de 100%, assim como a baixa representatividade dos alunos com menos de 50% das semanas ativo. O segundo critério, por sua vez, foi adicionado para categorizar como evadidos os alunos que, independentemente de sua

atividade ao longo do ano, não estudaram durante o último mês de sua preparação para o vestibular, comportamento que indicaria evasão segundo os funcionários do MS.

Figura 3 - Distribuição dos alunos por porcentagem de semanas ativo



A utilização destes dois critérios, ao invés de apenas uma variação do segundo critério apresentado, se deve ao comportamento dos alunos no contexto de cursos pré-vestibular, onde muitos voltam a estudar nas últimas semanas antecedendo a prova apesar de já terem passado diversas semanas sem atividade na plataforma. Deste modo, a aplicação dos dois critérios acima busca adaptar o conceito de evasão ao contexto do pré-vestibular para garantir que este comportamento de retorno nas últimas semanas não omita o histórico comportamental do aluno em questão.

4.2. Pré-Processamento de dados

Inicialmente, foi realizada uma filtragem na base de dados existentes no MS a fim de selecionar somente os alunos que estivessem dentro do perfil esperado. Com isso, reduziu-se a base de 18.000 alunos para 2.225. Posteriormente, removeram-se da base de dados 12 registros duplicados e 8 registros em branco, reduzindo, então, a base para 2.205 registros.

Para a consolidação dos dados existentes na base, as variáveis discretas: “categoria”, “city”, “regiao”, “reg_metropolitana”, “etapa_estudos”, “principal_ferramenta”, “método_de_estudo”, “study_platform” e “curso_desejado” foram segmentadas em variáveis *dummies,* ou seja, foram

geradas colunas auxiliares de modo que cada coluna apresentasse uma das opções destas variáveis e possuísse valores apenas de 1 ou 0. Tal processo aumentou o número de colunas de 29 para 1.015.

Além disso, foram adicionadas quatro novas variáveis a partir da junção e manipulação das variáveis já existentes: “ritmo_a_m1”, “ritmo_ex_m1”, “dificuldade” e “indeciso”. As duas primeiras são variáveis que buscam introduzir no modelo a necessidade dos alunos que adentraram o curso mais tarde de assistir aulas e fazer exercícios mais rápido que seus colegas que começaram seus estudos anteriormente. A variável “dificuldade” está relacionada ao curso de escolha do aluno e consiste em uma nota em uma escala de 1 a 6, de acordo com a nota de corte média para o curso escolhido no ENEM 2018. Os cursos de mais difícil entrada assumiram então a nota máxima 6. Por fim, a variável “indeciso” serve exclusivamente para destacar os alunos que não haviam decidido para que curso prestariam o vestibular no momento em que preencheram seu cadastro no MS.

Uma vez concluída a adição de variáveis pertinentes para a análise, foram desconsideradas todas as colunas que não trariam nenhum dado relevante, como: “key” e “user_id” por serem únicas de cada aluno, e todas as colunas diretamente ligadas ao critério de definição de evasão, como: “pct_semanas_ativo” e “total_semanas_ativo”, pois estas direcionariam o modelo diretamente à resposta certa. Por fim, removeu-se as colunas que possuíam valores iguais para todos os alunos, como: ”tipo_compra” e “principal ferramenta”.

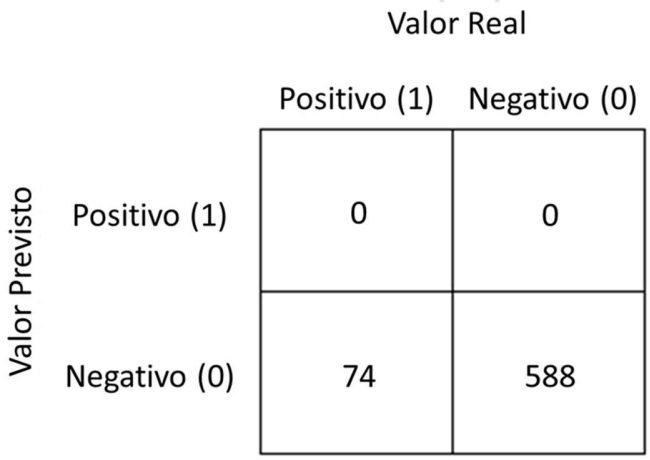
Como última etapa do pré-processamento dos dados, normalizaram-se todas as colunas com valores contínuos para evitar que a diferença de magnitude entre as variáveis inflasse a importância de algumas em detrimento de outras. Sendo assim, chegou-se na base de dados final, a qual será utilizada para a aplicação da ferramenta de classificação por RF.

4.3. Aplicação e avaliação do Algoritmo de ML

Com o pré-processamento de dados concluído, a base de dados foi segmentada em base de treino e base de teste, representando 70% e 30% do total, respectivamente. Através de um processo iterativo, identificou-se o número ótimo de árvores a serem utilizadas no problema levando em consideração precisão e capacidade computacional. Optou-se então por um modelo com 500 árvores.

A acurácia do modelo testado foi aproximadamente 89%. Contudo, ao analisar sua matriz de confusão, figura 4, fica claro que o modelo desenvolveu uma simplificação brusca enquanto buscava a solução mais simples possível que gerasse uma boa acurácia, o que nesse caso era classificar todos os alunos como não evadidos. Esse erro ocorre porque o modelo não dá nenhuma prioridade aos acertos nos valores positivos, buscando apenas o maior acerto possível em ambas as categorias juntas.

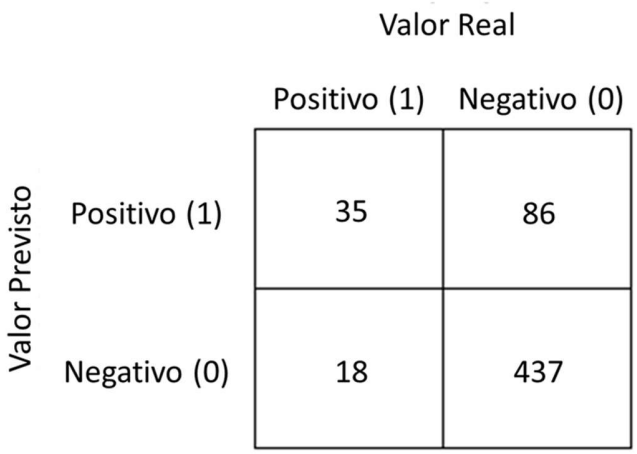
Figura 4 - Matriz de confusão I



Para evitar que o modelo tome este tipo de atalho, alterou-se o peso das categorias de 1:1 para 1:11 a fim de penalizar mais severamente os erros de categorização dos alunos que evadiram, ou seja, os falsos negativos. Além disso, aproveitou-se para já prevenir o modelo contra *overfitting*. Para tanto, elevou-se o número mínimo de amostras necessárias para subdividir uma folha de 1 para 6, o número mínimo de amostras para subdividir um nodo de 2 para 12 e a profundidade máxima da árvore foi alterada de ∞ para 8, pois, com um número mínimo de amostras para subdivisão elevado e uma profundidade moderada, o modelo tem menor chance de se adaptar excessivamente à base amostral.

Após os ajustes, o programa foi rodado novamente atingindo acurácia de 79%. A queda de acurácia era esperada devido alteração nos pesos das categorias. Contudo, esta queda ocorreu em troca de um aumento drástico na capacidade de identificação do modelo, agora prevendo em torno de 66% dos alunos que vieram a evadir, como pode ser visto na nova matriz de confusão, figura 5.

Figura 5 - Matriz de confusão II



Finalmente, um dos resultados mais importantes desta análise é a identificação do peso atribuído a cada uma das variáveis, ou seja, a importância dela na predição final. Podemos ver como se deu a alocação de importância na tabela 2, onde estão ranqueados os principais fatores envolvidos da categorização dos alunos.

Tabela 2 - Importância das variáveis

N°	**Variável**	**Importância**	**Importância acumulada**
1	semanas_enem	12%	12%
2	ex_s4	5,8%	18%
3	ex_s3	5,2%	23%
4	a_s4	4,7%	28%

5	ritmo_ex_m1	4,4%	32%
6	ex_s2	4,0%	36%
7	ex_s1	4,0%	40%
8	ritmo_a_m1	4,0%	44%
9	a_s3	3,8%	48%
10	a_s2	3,8%	52%
11	pct_cons_tarde	3,7%	56%
12	a_s1	3,6%	59%
13	pct_cons_manha	2,5%	62%
14	internet_acess	2,3%	64%
15	study_platform_Pelo Computador/Notebook	1,8%	66%
16	etapa_estudos_Colégio	1,5%	67%
17	metodo_estudo_Revejo algumas aulas mais de uma vez para fixar o conteúdo	1,5%	69%
18	metodo_estudo_Costumo pausar os vídeos para fazer anotações e resumos pessoais	1,4%	70%
19	dificuldade	1,3%	72%
20	study_platform_Pelo Celular	1,3%	73%
21	fez_simulado	1,3%	74%
22	percentual_consumo_ios	1,3%	75%

23	regiao_Sul	1,3%	77%
24	metodo_estudo_Revejo algumas aulas mais de uma vez para fixar o conteúdo	0,7%	77%
25	etapa_estudos_Faculdade	0,7%	78%
26	categoria_Básico	0,7%	79%
27	privada	0,7%	79%
28	metodo_estudo_Ainda não achei a melhor forma de...	0,6%	80%

Das variáveis utilizadas no modelo, percebe-se que a de maior importância para a previsão de evasão dos alunos foi a "semanas_enem" com 12%, indicando a importância do tempo disponível para o preparo pré-vestibular. Nota-se que um maior tempo de estudo disponível pode gerar uma maior taxa de evasão, possivelmente indicando que alunos que iniciam seus estudos com maior antecedência perdem sua motivação de estudos com maior frequência que seus colegas que adentraram na plataforma mais próximos da data da prova.

Outro fator notável é a prevalência das variáveis de engajamento do aluno na plataforma (a_,ex_,ritmo_) representando, em conjunto, 43% da importância geral. Este resultado era esperado e reforça a importância do acompanhamento destas métricas de atividade na plataforma.

Também é perceptível a maior importância atribuída a variáveis de engajamento relacionadas aos exercícios, em detrimento às relacionadas a visualização de aulas. Com exceção de "a_s4", todas as variáveis de exercício ficaram acima das variáveis de visualização, informação que indica que uma ação de maior engajamento, como exercícios, tende a possuir um comportamento mais discrepante entre alunos que permanecem e que evadem do que ações de menor engajamento, como visualização de aulas, contudo, a diferença de importância é pequena, 0.2%. Além disso, percebe-se uma relevância maior nas variáveis de engajamento nas últimas semanas de análise s3 e s4, o que pode indicar uma tendência de desengajamento na plataforma com o tempo, mesmo em um período curto de um mês, como o utilizado nesta pesquisa.

Ademais, surpreende a importância atribuída à variáveis relacionadas ao turno de estudo, "pct_cons_tarde" e "pct_cons_manha", juntas somando 6,2% da importância geral. Pode-se inferir

que estas variáveis não têm relação causal com a evasão, ou seja, alterar o turno de estudo não é uma estratégia válida para reduzir a evasão. Contudo, sua correlação indica uma menor taxa de evasão dos alunos que fazem a maior parte de seus estudos no período do dia, o que pode ser reflexo de outros fatores, como tempo disponível e nível de importância atribuído aos estudos.

Além disso, é interessante ressaltar a importância de fatores como método e qualidade de acesso à plataforma, “study_platform_Pelo Celular”, “study_platform_Pelo Computador/Notebook” e ”internet_acess”, que juntos somam 5%, conforme esperado de acordo com a literatura. Por fim, foi inesperada a relevância de variáveis ligadas ao método de estudo, juntas somando aproximadamente 4%, o que indica que uma maior orientação de boas práticas nos estudos pode vir a ter resultados positivos na redução de evasão.

5. CONCLUSÃO

Esse artigo buscou elaborar uma definição de evasão no contexto do curso pré-vestibular online do MS e, a partir desta definição, desenvolver um modelo preditivo utilizando técnicas de ML para identificar o subconjunto de alunos do MS em risco de evasão; por fim, foram identificados os principais fatores que indicam o risco de um aluno evadir.

Para a construção da base de dados deste estudo foram utilizadas variáveis numéricas e não numéricas disponíveis ao MS após um mês da inscrição do aluno, refletindo fatores comportamentais e de perfil dos estudantes. Para tanto, foram considerados apenas alunos com mais de cinco meses disponíveis para preparação e que declararam o MS como principal ferramenta de estudos, de modo a homogeneizar a base de dados e evitar ruídos.

Na análise, foi utilizado o método de categorização de RF que gerou predições com acurácia de 79%, assim como uma predição correta de 66% dos alunos que vieram a evadir, o que é considerado um bom resultado quando comparado a outros estudos nacionais na área de evasão. Dentre as variáveis utilizadas na predição, percebeu-se uma forte influência de fatores de engajamento na plataforma, antecedência da compra, turno de estudo, acesso à plataforma e método de estudo, representando 43%, 12%, 7%, 5%, 4% do todo, respectivamente, o que está condizente com a literatura.

Com as informações descobertas nos resultados da pesquisa, compreende-se que o KDD proporcionou os resultados esperados, trazendo informações úteis e novas para empresa. Esta adquiriu um melhor conhecimento dos indicadores a serem avaliados na previsão da evasão dos

alunos, assim como um maior entendimento da importância de variáveis como o engajamento nos exercícios disponíveis na plataforma e o método de estudo empregado pelo aluno.

Compreende-se como uma limitação do estudo a necessidade da elaboração de uma definição de evasão através do engajamento ao longo do ano, o que pode ter acarretado numa categorização incorreta de alunos que adquiriram a plataforma cedo sem intenção de se dedicar aos estudos imediatamente. Como potencial trabalho futuro, pode-se recriar o modelo preditivo a partir dos resultados de uma pesquisa com ex-alunos, de modo a garantir a categorização correta em evasões e permanências, podendo assim expandir a análise para toda a base de alunos do MS.

REFERÊNCIAS

ATTARAN, M., ATTARAN, S. "Opportunities and Challenges of Implementing Predictive Analytics for Competitive Advantage". International Journal of Business Intelligence Research archive, Vol.9, No2, 2018. p.1-26

BALDI, P., BRUNAK, S., "Bioinformatics: The Machine Learning Approach Second Edition". 2001

BARROSO M. F.; FALCÃO, E. B. M., "Evasão universitária: o caso do Instituto de Física da UFRJ". In: ENCONTRO NACIONAL DE PESQUISA EM ENSINO DE FÍSICA, 9., 2004, Jaboticatubas. Anais... Jaboticatubas: Sociedade Brasileira de Física, 2004. p. 1-14.

BECK, J. E., WOOLF, B. P,. "High-Level Student Modeling with Machine Learning". ITS 2000: Intelligent Tutoring Systems pp 584-593

BREIMAN, L. "Random Forest ". Machine Learning archive, Vol.45 No.1, 2001. p 5-32

CHI-CHUN LEE , EMILY MOWER , CARLOS BUSSO , SUNGBOK LEE , SHRIKANTH NARAYANAN, "Emotion recognition using a hierarchical binary decision tree approach", Speech Communication, v.53 n.9-10, 2011. p.1162-1171.

DE OÑA, J., LÓPEZ, G., ABELLÁN, J "Extracting decision rules from police accident reports through decision trees", Accident Analysis & Prevention, Vol. 50, No., 2013, p. 1151-1160.

DELEN, D. "A comparative analysis of machine learning techniques for student retention management." Decision Support Systems Vol.49 No.4, 2010. p.498-506.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, E P. “From data mining to knowledge discovery in databases”. Data Mining to Knowledge Discovery in Databases. AI Magazine. vol.17, No.1, 1996. p.37-54

FORRESTER, by ADAMS, J., SEHGAL, V., KUMAR, S. “Data World Business Intelligence And Analytics Software Forecast, 2016 To 2021 (Global)”. 2017. Disponível em: https://www.forrester.com/report/Forrester+Data+World+Business+Intelligence+And+Analytics+Software+Forecast+2016+To+2021+Global/-/E-RES137027 Acesso em 10 de outubro de 2018

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. “The elements of statistical learning”. Springer Series in Statistics. Vol. 1, No.10, 2001

GAIOSO, Natalicia Pacheco de Lacerda. “O fenômeno da evasão escolar na educação superior no Brasil”. 2005. 75 f. Dissertação (Mestrado em Educação) – Programa de Pós-Graduação em Educação da Universidade Católica de Brasília, Brasília, 2005.

GANDHIMATHI, D., GOMATHI, S. “A Survey of Approaches and Tools Used in Educational Data Mining”, IJIRCCE v.3 n.10. 2015

GIASSON, E., SARMENTO, E. C., WEBER, E., FLORES, C. A., HASENACK, H. “Árvores de decisão para o mapeamento digital de solos em encostas basálticas subtropicais”. Scientia Agricola. Piracicaba. Vol. 68, n.2, 2011. p. 167-174.

GIL, A. C. Métodos e técnicas de pesquisa social. 5. ed. São Paulo: Atlas, 1999

HART, C. (2012). Factors Associated With Student Persistence in an Online Program of Study: A Review of the Literature. Journal of Interactive Online Learning, vol.11, No.1, 2012. p.19-42.

HUSSAIN, M., ZHU, W., ZHANG, W., ABIDI, S., M. R. “Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores”. Computational Intelligence and Neuroscience, vol. 2018, Article ID 6347186, 21 pages. 2018

KUMAR, S. “The Differences Between Machine Learning And Predictive Analytics”. 2018, Disponível em <https://www.digitalistmag.com/digital-economy/2018/03/15/differences-between-machine-learning-predictive-analytics-05977121 > Acesso em: 20 de novembro de 2018.

LOBO, M. B. C. M. “Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções”. Associação Brasileira de Mantenedoras do Ensino Superior, 25, 2012

MACHADO, R. D., NARA, E. O. B., SCHREIBER J. N. C., SCHWINGEL G. A. “Estudo bibliométrico em mineração de dados e evasão escolar”. In XI Congresso Nacional de Excelência em Gestão, 2015

MANHÃES, L. M. B.; CRUZ, S. M. S.; COSTA, R. J. M.; ZAVALETA, J.; ZIMBRÃO, G. “Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados”. In XVII Simpósio Brasileiro de Informática na Educação, 2011.

MARSLAND, S. “Machine Learning An Algorithmic Perspective Second Edition”. 2014

MICHALSKI, R. S., CARBONELL, J. G., MITCHELL, T. M. ”Machine Learning: An Artificial Intelligence Approach”. Artificial Intelligence Vol.25, No.2, 1984. Pages 236-238

MUILENBURG, L. Y., e BERGE, Z. L. "Student barriers to online learning: A factor analytic study. Distance Education". vol.26, No.1, 2005. p.29-48

NASCIMENTO, F. P. do ; SOUSA, F. L. L. “Metodologia da Pesquisa Científica: teoria e prática – como elaborar TCC”. ed. Brasília: Thesaurus Editora, 2015. vol. 1

PRASS, F. S. “KDD – Uma Visão Geral Do Processo”. FP2 Tecnologia, 2012. Disponível em: http://www.fp2.com.br/blog/index.php/2012/artigo-kdd-uma-visao-geral-do-processo. Acesso em: 11 de outubro de 2018.

ROKACH, L., MAIMON, O. “Data mining with decision trees: theory and applications (2nd edition)”. Series in Machine Perception and Artificial Intelligence, vol.81, 2015

SHAFIQUE, U., QAISER, H. “A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)”. International Journal of Innovation and Scientific Research. Vol. 12 No.1, 2014, p.217-222

SHARKEY, MIKE & SANDERS, ROBERT “A Process for Predicting MOOC Attrition”. EMNLP, 2014 p.50-54.

SHARMA, S., AGRAWAL, J., SHARMA, S. “Classification Through Machine Learning Technique: C4.5 Algorithm based on Various Entropies”. International Journal of Computer Applications. Vol.82, 2013. p28-32.

SHAWKY, D., BADAWI, A. “A Reinforcement Learning-Based Adaptive Learning System”. AMLTA, 2018. p.221-231.

SILVA FILHO, R.L.L., MOTEJUNAS, P.R., HIPÓLITO, O., LOBO, M.B.C.M. “A evasão no ensino superior brasileiro”. Cad. Pesqui. [online], vol.37, No.132, 2007. pp.641-659.

SINHA, T., JERMANN, P., LI, N., DILLENBOURG, P. “Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions”. Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 2014. p3-14

SOCIETY OF ACTUARIES. “Report of the Society of Actuaries Predictive Modeling Survey Subcommittee”. 2012

SOUZA, C. J.; PETRÓ, C.S.; GESSINGER, R. M, “Um estudo sobre evasão no ensino superior do Brasil nos últimos dez anos”. Rio Grande do Sul (2011).

SYED, T.A., PALADE, V., IQBAL, R., NAIR, S. S. K. “A Personalized Learning Recommendation System Architecture for Learning Management System”. 9th International Conference on Knowledge Discovery and Information Retrieval, 2017

TONTINI, G., WALTER, S. A. “Pode-se identificar a propensão e reduzir a evasão de alunos? Ações estratégicas e resultados táticos para instituições de ensino superior”. Revista da Avaliação da Educação Superior (Campinas), vol.19, No.1, 2014. p.89-110.

WITTEN, IAN, H., FRANK, EIBE. “Data Mining: Practical machine learning tools and techniques with java implementations”. 2000

XING, WANLI & DU, DONGPING. “Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention”. Journal of Educational Computing Research. vol 57, No.3, 2018.