Proposta de Ferramenta para Predição de Evasão no Ensino Superior

Josiane Nunes de Araujo Reis, Weverton da Silva Souza, Sônia A. Santana

Bacharelado em Ciência da Computação – Centro Universitário do Triângulo (UNITRI).

josiane-reis@outlook.com, wevertonad@gmail.com,soniaapsasntana@gmail.com

Abstract. The increase in dropout rates in higher education has been of studies and research that aim to identify the factors that lead a student to evade, the use of data mining technologies has been highly effective in identifying patterns that aid the institution in making decisions and actions to mitigate such a problem. The objective of this work is to present a proposal for a tool to forecast cases of evasion in higher education using data mining techniques added to IaaS cloud computing technologies, which have benefits such as flexibility, agility and security in the development of projects. The article presents an overview of the problem of avoidance in educational institutions and describes the technologies used in the creation of the analysis models used in the case study for the predictions that presented levels of accuracy between 0.87 and 0.93 proving to be effective for use with bases of real data from educational institutions.

Resumo. O aumento nos índices de evasão no ensino superior tem sido alvo de estudos e pesquisas que visam identificar os fatores que levam um aluno a evadir, a utilização de tecnologias de mineração de dados tem se mostrado altamente eficaz na identificação de padrões que auxiliem a instituição na tomada de decisões e ações para mitigar o problema. O objetivo deste trabalho é apresentar uma proposta de ferramenta para predição de casos de evasão no ensino superior utilizando técnicas de mineração de dados agregadas a tecnologias de *cloud computing IaaS,* que contam com benefícios como flexibilidade, agilidade e segurança no desenvolvimento de projetos. O artigo apresenta uma visão geral do problema da evasão nas instituições de ensino e descreve as tecnologias usadas na criação dos modelos de predição usados no estudo de caso. Os resultados apresentaram níveis de acurácia entre 0.87 e 0.93 mostrando-se eficaz para utilização com bases de dados reais de instituições de ensino.

1. Introdução

Com a democratização do acesso ao ensino superior, a evasão escolar, um dos maiores desafios encontrados atualmente pelas instituições de ensino superior, tem aumentado cada vez mais, o que caracteriza um grande problema, não só para as instituições mas também para a sociedade no geral.

O sistema educacional superior brasileiro [INEP 2009], possui índices muito altos de estudantes que iniciam um curso e por diversas razoes não conseguem concluí-lo. A identificação de perfis de alunos evasores tem que sido alvo de pesquisas ao longo

do tempo e a aplicação de técnicas de *data mining (mineração de dados)* vem sendo estudadas cada vez mais gerando assim aporte para a identificação de tais alunos, o que possibilita tomada de ações para prevenir ou até mesmo reverter tais casos de evasão.

Seguindo as fases da metodologia *CRISP-DM (Cross-Industry Standard for Data Mining)* este trabalho apresenta uma proposta de ferramenta que utiliza métodos e processos de mineração de dados para prever possíveis casos de evasão identificando o percentual de chance que um aluno tem de evadir do curso a partir da análise dos dados já existentes das universidades e identificação dos padrões que podem acabar levando o aluno a evadir do curso.

Dessa forma, este estudo trabalho foi dividido em 7 sessões. A sessão 2, apresenta uma visão geral sobre o problema evasão nas instituições de ensino e sobre as técnicas de mineração de dados que cada vez mais vem sendo utilizadas em estudos em torno deste assunto. A sessão 3 apresenta um resumos dos principais trabalhos correlatos e de seus resultados. Na sessão 4 é apresentada a metodologia que serviu como base para desenvolvimento deste trabalho.

Na sessão 5 é apresentada a arquitetura proposta para o desenvolvimento deste trabalho, baseada na metodologia escolhida. A sessão 6 apresenta o estudo de caso realizado como prova de conceito da arquitetura proposta, os métodos utilizados e uma análise dos resultados obtidos com ele. Na sessão 7 é apresentada a conclusão que oferece uma relação dos resultados atingidos com os pretendidos e por fim, a sessão 8 apresenta as considerações finais e perspectivas de trabalhos futuros.

2. Evasão e Mineração de Dados Escolares

Em um contexto global, o ensino superior tem atraído cada vez mais os olhares da população em geral. Segundo alguns estudos como Carvalho e Nunes (2007) especificamente no Brasil, o ensino superior tem se expandido de maneira significativa, motivada pelo aumento do número de instituições do tipo, da abertura de novos cursos e também por meio da democratização das formas de ingresso aos cursos ofertados. Por outro lado, apesar da quantidade crescente de vagas para os cursos de graduação que são disponibilizadas anualmente, têm-se observado que uma grande parcela dos ingressantes acabam não concluindo seus estudos, o que tem gerado um grande problema para as instituições de ensino superior (IES), tanto privadas quanto públicas [INEP 2009].

Compreender o fenômeno da evasão nas IES é algo complexo. Segundo Gaioso (2005) a evasão escolar pode ser definida como a interrupção no período de estudos, seja de nível básico ou superior. Especificamente, a nível superior, o termo evasão é muito utilizado para fazer referência à perda, fuga ou abandono das instituições pelos alunos segundo Kira, (2002) e deve ser analisado a partir de vários aspectos, visto que

fatores como contexto histórico, cultural, situação socioeconômica influenciam de diversas maneiras para seu aumento segundo Barroso e Falcão (2004).

Segundo Hämäläinen et al. (2004) os índices de evasão podem ser medidos a partir de diversos níveis de abrangência, como por exemplo, um determinado período de tempo, um determinado curso de graduação e por diversas outras situações e contextos que a propiciam.

Vinculado ao Ministério da Educação, o INEP (Instituto Nacional de Educação e Pesquisa) aponta que, em 2010, houve uma média de evasão de 11,4% nas IES brasileiras sendo que em 2014, esse número chegou a 49%, embora seja visto como um fator crucial para o desenvolvimento tanto das universidades quanto para a população em geral, atualmente, dos poucos estudos em torno das causas e métodos para prevenção ou reversão dos casos de evasão nas IES segundo Saraiva e Masson (2003), a maioria se restringe à identificação das causas que contribuem para evasão utilizando métodos estatísticos que independem de aplicações e/ou sistemas de informação.

A utilização de técnicas de mineração de dados ou *Data mining*, têm possibilitado uma ampliação nas formas de análise das causas e métodos para prevenção de evasão por se tratar de um processo que possibilita a identificação de grandes volumes de dados e a descoberta de padrões úteis para um determinado fim segundo Fayyad et al. (1996).

Segundo Fayyad et al. (1996) e Ng; Han, (2002), *Data Minin*g é uma etapa importante do processo *KDD (Knowledge Discovery in Databases*) e consiste na análise e classificação de grandes quantidades de dados utilizando algoritmos baseados em técnicas estatísticas e de inteligência artificial, a fim de descobrir, de forma automatizada, relações ou padrões implícitos, potencialmente compreensíveis e úteis para comprovar alguma hipótese a partir de resultados até então não perceptíveis nos dados, a fim de trazer algum beneficio ou auxiliar no desenvolvimento de alguma atividade.

Dentro da mineração de dados existe uma diferença entre tarefas e técnicas de mineração. Segundo Tan et al. (2006) as técnicas de mineração especificam os métodos que possibilitam descobrir os padrões desejados dentro dos dados analisados

Já as tarefas especificam as categorias de padrões a serem encontrados e são classificadas como preditivas – supervisionadas e descritivas - não supervisionadas. As tarefas descritivas, são basicamente de associação, agrupamento e sumarização, e tem a finalidade de realizar a busca de padrões com base na correlação existente entre os dados. As tarefas preditivas por sua vez, como classificação e regressão tem a finalidade de prever o valor de uma determinada variável baseado no valor de outras.

De acordo com Fayyad et al. (1996) essas técnicas ou métodos são classificados de acordo com a tarefa realizada. Tal classificação segue uma linha bem tênue já que alguns métodos preditivos podem realizar tarefas descritivas e vice-versa.

3. Estudos Relacionados

No contexto educacional, segundo Baker et al. (2011) a utilização de técnicas de mineração de dados utilizando algoritmos aplicados a dados educacionais ou *EDM (Educational Data Mining*) tem crescido cada vez mais devido à facilidade propiciada pela utilização de softwares educacionais que proporcionam cada vez mais um fácil acesso aos dados para análises por pesquisadores.

Trata-se de uma linha de pesquisa que oferece grande potencial para melhorias tanto na qualidade do ensino quanto para a realização de descobertas no âmbito educacional como identificação de melhores abordagens instrucionais para determinadas situações, identificação de padrões de comportamentos de alunos e até mesmo personalização do ambiente de estudo e dos métodos utilizados no ensino [Baker et al. 2011].

Dos diversos estudos em torno da área de EDM pode-se destacar por exemplo o estudo de caso realizado por Santos et al. (2012) que permite identificar alunos que têm maior risco de reprovação. Resultados preliminares mostram que os modelos criados permitem a identificação da propensão à reprovação com taxa de acerto em torno de 69%.

Martins et al. (2012) buscam aprimorar um Assistente de Predição de Evasão (APE) por meio da aplicação do processo KDD acrescentando novas variáveis obtidas com os dados do Sistema Tutor Inteligente (STI) e de outras fontes. No processo de KDD foram utilizados os algoritmos OneR e NNge presentes na ferramenta *Weka*. Os autores observaram que o modelo baseado no processo de *KDD* resultou em uma melhor identificação e maior precisão nos resultados.

Gotardo et al. (2013) apresentam uma proposta que utiliza algoritmos de aprendizagem acoplados integrando diversas técnicas de aprendizagem e Mineração de Dados. Segundo os autores, os resultados encontrados nos experimentos demonstraram que a técnica proposta auxilia no oferecimento de recomendações sobre o desempenho do aluno através da geração de modelos de classificação.

No trabalho realizado por Digiampietri; Nakano; Lauretto, (2016) foi realizada a classificação de alunos do primeiro ano do curso quanto ao risco de evasão utilizando uma metodologia que se baseou em mineração de dados, utilizando exclusivamente o histórico escolar de tais alunos. Após serem analisados os históricos escolares de mais

de mil alunos de um determinado curso de graduação, foi possível predizer, com índices de acerto superiores a 90% os alunos com grande potencial a evadir do curso.

No recente trabalho de Schreiber et al. (2017) é apresentado um software que visa mostrar preventivamente casos de evasão discente no ensino superior. O software apresenta a probabilidade individual e os possíveis e mais fortes motivos que influenciam os discentes a evadir, utilizando um dos métodos de classificação de *Data Mining*, as redes bayesianas. Foram criados dois modos de operação onde o primeiro calcula a probabilidade de um único aluno evadir e o segundo modo analisa um conjunto de discentes identificando em cada um a probabilidade de evadir.

Kantorski et al. (2015) explicam uma ferramenta de predição de evasão no ensino superior utilizando uma adaptação da metodologia *CRISP-DM (Cross Industry Standard Process for Data Mining)* proposta por Chapman (2000). Foram realizados estudos de caso em dois cursos de graduação de uma determinada instituição, demonstrando bastante eficiência na proposta, contando com uma acurácia media de 96% na previsão, e mais de 73% de acerto na previsão de alunos que evadiram dos cursos.

Na proposta realizada por Lykourentzou et al. (2009) é mostrado um método de predição de possíveis evasões de alunos de ambientes online que se baseia na utilização da combinação de três técnicas bastante populares de aprendizado de máquina: redes neurais *feed-forward*, máquinas de vetores de suporte e conjunto probabilístico simplificado ARTMAP difuso. Os resultados demonstrados variaram entre 73% e 94% de acerto.

Em geral, os trabalhos que fazem uso de EDM podem se dividir em duas linhas de estudo principais: uma delas visa identificar os atributos e fatores relevantes que possam caracterizar os estudantes e assim traçar um perfil, e a outra visa utilizar e comparar o desempenho de algoritmos de classificação e assim encontrar os algoritmos mais apropriados para a solução do problema. Na sessão a seguir serão apresentadas as características da metodologia e os passos realizados para construção do estudo de caso proposto, detalhando os métodos utilizados e os resultados obtidos a partir dos testes realizados.

4. Metodologia

A metodologia escolhida para se aplicar neste trabalho gira em torno de 6 etapas principais representadas na figura 1, chamada metodologia CRISP-DM (*Cross Industry Standard Process for Data Minning)* que segundo Chapman et al., (2000) pode ser traduzida como Processo Padrão Inter-Indústrias para Mineração de Dados.

A escolha da metodologia se deu principalmente por se tratar de uma das metodologias mais utilizadas nos processos de Mineração de dados nos principais

estudos em torno do assunto e possuir muitas vantagens como uma sequência de fases flexíveis e por oferecer a possibilidade de ser aplicada em uma gama muito grande de negócios sem depender de uma ferramenta ou tecnologia específica para ser executada conforme destacado por Azevedo (2008).

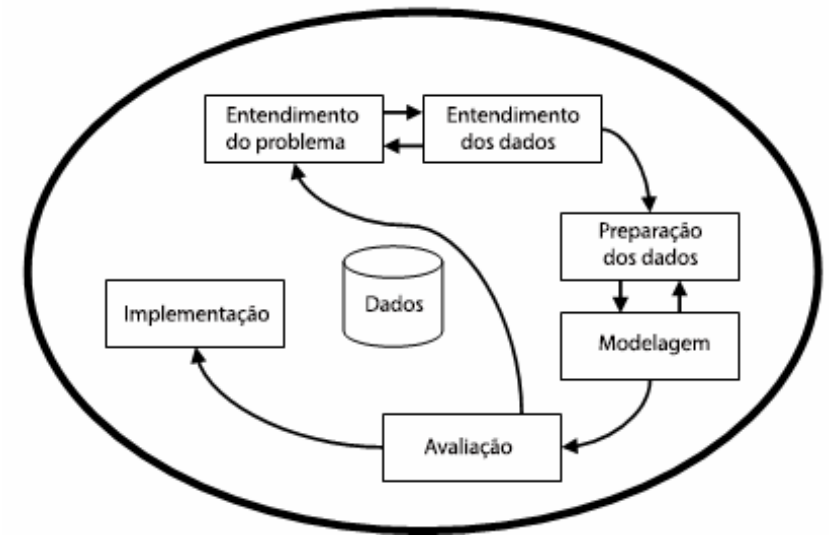


Figura 1: Etapas da metodologia CRISP-DM

A primeira etapa consiste no entendimento do problema a ser tratado. O objetivo é conhecer e entender as características do negócio em que será aplicado os processos de mineração de dados.

A segunda etapa, etapa de compreensão dos dados visa coletar, organizar e descrever os dados disponíveis e que serão trabalhados no processo. É nesta etapa que é feita, basicamente, a avaliação dos dados que podem ser relevantes na solução do problema. A ideia é utilizar dados de alunos verificando suas propriedades e qualidades. A análise destes dados envolve a observação de características como intervalos de valores, quantidade de atributos e seus significados e qual relevância cada um possui para a solução do problema.

A terceira etapa, preparação dos dados, é a etapa em que os dados são manipulados de maneira técnica, onde é gerado um novo modelo de base adequada ao modelo de descoberta ou mineração de será utilizado. A geração desses novos modelos de dados envolve etapas tais como análise, seleção, limpeza, construção e formatação dos dados, isso serve para combinar diversas fontes de dados e realizar alterações de maneira sintática, ou seja, sem modificar seu significado.

Na etapa de modelagem, quarta etapa, são escolhidas e aplicadas as técnicas de mineração de dados mais adequadas para a situação, a partir dai é gerado um conjunto de dados para teste que depende é claro, dos objetivos identificados na primeira etapa do processo.

A quinta etapa, etapa de avaliação, consiste na revisão do modelo obtido. É onde é analisado se os resultados obtidos estão de acordo com os objetivos propostos na fase

de análise do negócio. Esta etapa é de suma importância para o projeto, pois é nela que é decido se o modelo final precisará de correções ou terá continuidade.

A sexta e última etapa desta metodologia consiste no conjunto de ações que serão realizadas para explicar os resultados obtidos no projeto e no monitoramento da implementação realizada.

5. Arquitetura proposta

Seguindo as fases da metodologia escolhida, durante a fase de entendimento do problema, foram realizados diversos estudos e pesquisas relacionados a evasão escolar, suas definições, os métodos utilizados para medição de seus índices e os principais estudos e pesquisas que buscam aplicar técnicas e conceitos de mineração de dados para solução do problema.

Ainda na etapa de entendimento do negócio, foram definidos os métodos e tecnologias a serem utilizadas no desenvolvimento deste trabalho, uma vez que a escolha dos métodos e tecnologias de mineração de dados esta relacionada aos dados que serão trabalhados e os objetivos pretendidos.

Na etapa de compreensão dos dados, o tratamento pode ser feito de maneira relativa, de acordo com a base de dados fornecida pelas instituições de ensino visto que cada instituição de ensino dispõem de diferentes critérios para julgar a importância dos dados que são armazenados.

A figura 2 descreve as etapas da arquitetura proposta, onde optou-se pela utilização do modelo de serviço *IaaS (Infraestructure as a Service)* que faz parte de um dos 3 modelos presentes na abordagem de *cloud computing,* um modelo de serviço que possibilita o compartilhamento dinâmico de recursos, entregando serviços pela rede na forma de *web services*, o que proporciona uma flexibilidade muito grande na alocação de recursos, baixo custo, disponibilidade e escalabilidade [Voorsluys et al. 2011].

Dos diversos tipos de serviços do modelo *Iaas* disponíveis atualmente, optou-se por utilizar o *Amazon S3 (Amazon Simple Storage Service), o Amazon SageMaker e o Amazon ML (Amazon Machine Learning),* todos pertencentes a plataforma *AWS (Amazon Web Services)* que foi escolhida para ser utilizada neste trabalho principalmente pode ser considerada hoje, um do maiores provedores de serviço em nuvem e um dos mais utilizados devido aos benefícios oferecidos como um ambiente de trabalho altamente escalável, amplo acesso aos recursos, disponibilizar uma infraestrutura com vários níveis de processamento, executando desde tarefas simples até tarefas de alto desempenho de acordo com a demanda do usuário.

Como ilustrado pela figura 2, e etapa 1 do processo compreende a fase entendimento dos dados que visa coletar os dados das instituições de ensino que serão trabalhados durante as análises e predições, os dados são enviados para um *bucket* de armazenamento em nuvem que visa unificar a forma com que os dados são armazenados visto que cada instituição de ensino pode trabalhar com bancos de dados diferentes e arquivos de dados em formatos e tamanhos diferentes.

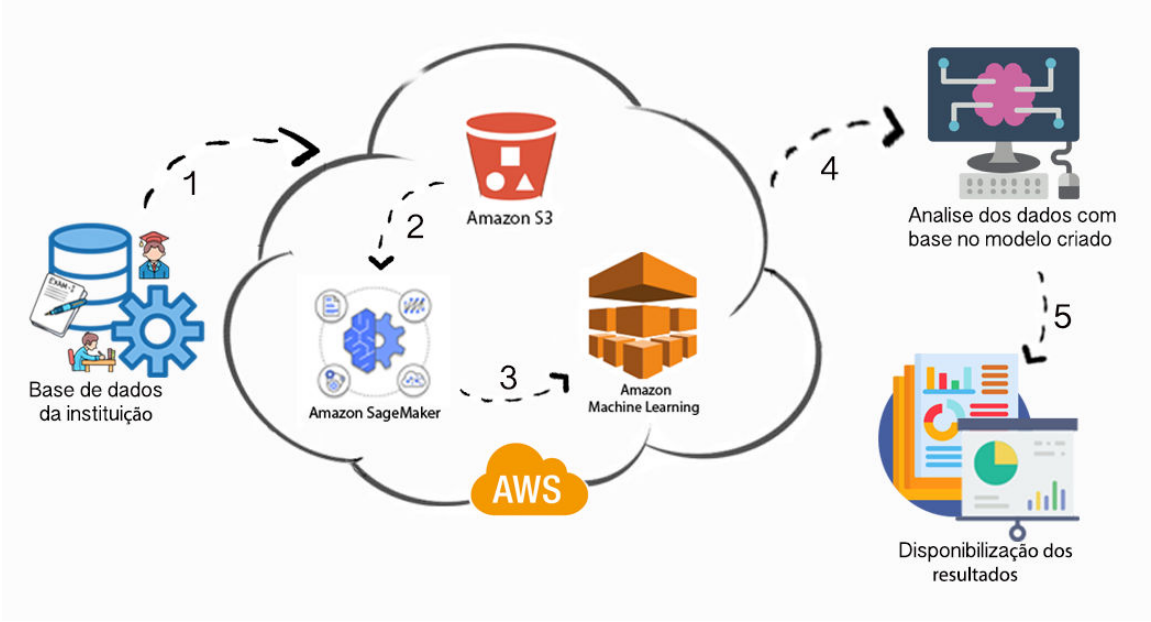


Figura 2: Arquitetura proposta utilizando o modelo *IaaS* com serviços *AWS*.

A fase 2 compreende a etapa de preparação dos dados onde a base de dados armazenada na fase anterior é enviada a uma ferramenta de análise e manipulada sempre garantindo que seu significado não seja alterado, para que se obtenha um conjunto final de dados livre de atributos incorretos, campos vazios ou dados inconsistentes que podem prejudicar a geração dos modelos de aprendizado.

A fase 3 apresenta a fase central da mineração, onde são criados os modelos de *machine learning* a partir do conjunto final de dados utilizando as técnicas de mineração de dados escolhidas. A validação deste modelo gerado consiste na revisão dos passos anteriores verificando se os resultados encontrados estão de acordo com os objetivos propostos e só então são realizadas as predições e disponibilização dos resultados a partir dos dados dos estudantes que se quer verificar a possibilidade de evadir como mostrado nas fases 4 e 5 da arquitetura proposta.

6. Estudo de caso

Para validação da arquitetura proposta neste trabalho, foram gerados 3 modelos de *machine learning* e realizadas predições de evasão com base nos dados de 3 alunos fictícios de bases de dados geradas para este estudo de maneira a simular diversos tipos de casos, para entender possíveis variações nos resultados encontrados.

Como destacado na Figura 3, as 3 bases de dados utilizadas, foram geradas com quantidades distintas de atributos a fim de verificar a variações geradas por esse fator nas métricas de qualidade entre os modelos gerados.

Base de dados 1	Base de dados 2	Base de dados 3
age	age	age
gender	gender	gender
maritalStatus	admission	maritalStatus
skinColor	period	skinColor
admission	wasChangeCourse	admission
typeEducation	typeEducation	typeEducation
areaCourse	-	areaCourse
period	-	period
pendingDisciplines	-	pendingDisciplines
wasChangeCourse	-	wasChangeCourse
hasChildren	-	hasChildren
working	-	working
-	-	liveOtherCity
-	-	entranceType
-	-	aprovedDisciplineRate
-	-	noAprovedDisciplineRate

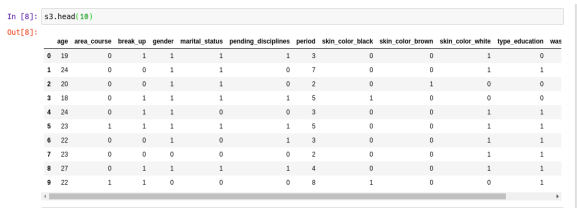
Figura 3: Atributos das bases de dados utilizadas na criação dos modelos.

Inicialmente foi realizada a triagem dos dados gerados. As bases de entrada contendo os dados de alunos consistem cada uma em um arquivo no formato *.csv* que possui 2000 registros com diversos tipos de dados dentre eles, idade do aluno, período cursado, gênero etc.

Foi criada uma instância de armazenamento dentro do *Amazon S3* para guardar dos dados coletados como descrito na etapa 1 da arquitetura proposta.

Na etapa de preparação dos dados foi utilizado o *Amazon SageMaker* para a criação de uma instância *Jupyter notebook que é* uma aplicação web *openSource* em que se pode utilizar várias linguagens de programação e bibliotecas para a visualização e entendimento de dados e resultados de análise, neste caso, foi utilizada a linguagem *Python*.

A figura 4 mostra a tabela com alguns dos atributos de uma das bases de dados utilizada para a análise dentro do *Amazon SageMaker*. Nota-se que, exceto pelas colunas de idade e período, os atributos são de tipo binário, considerando verdadeiro ou falso para cada coluna. Essa abordagem torna mais fácil o processamento dos dados dentro do *Sagemaker* e a criação de padrões e médias de valores necessários na análise dos dados.



```
In [8]: s3.head(10)
Out[8]:
```

	age	area_course	break_up	gender	marital_status	pending_disciplines	period	skin_color_black	skin_color_brown	skin_color_white	type_education	was
0	19	0	1	1	1	1	3	0	0	1	0	
1	24	0	0	1	1	0	7	0	0	1	1	
2	20	0	0	1	1	0	2	0	1	0	0	
3	18	0	1	1	1	1	5	1	0	0	0	
4	24	0	1	1	0	0	3	0	0	1	1	
5	23	1	1	1	1	1	5	0	0	1	1	
6	22	0	0	1	0	1	3	0	0	1	1	
7	23	0	0	0	0	0	2	0	0	1	1	
8	27	0	1	1	1	1	4	0	0	1	1	
9	22	1	1	0	0	0	8	1	0	0	1	

Figura 4: Exibição dos dados carregados no *SageMaker* a partir do *S3*

A partir dos dados carregados, foram realizadas no *SageMaker* as validações que identificaram dados ausentes ou inconsistentes, a delimitação de colunas importantes ou não para a criação dos modelos.

A figura 5 exemplifica um gráfico de dispersão gerado na análise feita dentro do *SageMaker* dos dados presentes nos campos de idade e período cursado pelo aluno. Nota-se a presença de dados inconsistentes, como idades menores que zero ou maiores que 200, períodos negativos ou maiores que a quantidade existente nas universidades, etc. Esses dados são os chamados *outliers e* podem ocorrer devido a erros no cadastro dos dados e até mesmo de digitação durante a inserção dos dados na base.

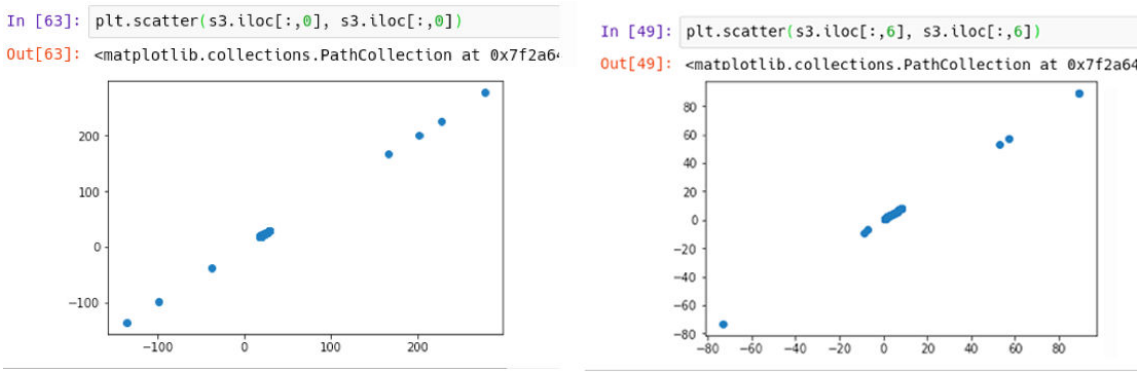


Figura 5: Gráficos de dispersão, demonstrando a presença de *outliers* nos campos de idade e período, respectivamente.

Após a realização da análise e preparação dos dados dentro do *SageMaker*, foi gerado o modelo de *machine learning* utilizando o *AWS Machine Learning* para realizar as predições. Para a criação do modelo, foram passados para o *Amazon Machine Learning* os dados já tratados anteriormente pelo *SageMaker*.

Antes de realizar a criação do modelo, o *AWS Machine Learning* realiza uma verificação na base de dados fornecida, para determinar se ela é válida ou não, onde são avaliados alguns dos aspectos tratados na etapa anterior como por exemplo, se a base possui dados faltantes ou inconsistentes.

Durante a criação do modelo de *machine learning*, o *Amazon Machine Learning* divide a base de dados carregada em dois grupos com o objetivo de realizar testes de qualidade no modelo que será gerado, o primeiro grupo contendo 70% dos dados é utilizado para a criação e treinamento do modelo, é onde o serviço cria padrões baseados nos dados lidos, para poder identificar o perfil dos alunos que evadem.

Para determinar a eficácia de um modelo e constatar se ele é bom o suficiente para ser colocado em uso, o *Amazon Machine Learning* utiliza o segundo grupo dos dados, com os 30% restantes, para validar o modelo e gerar métricas de qualidade que determinam confiabilidade preditiva do modelo criado.

A figura 6 exibe o log da criação de um dos modelos gerados neste estudo de caso, onde são exibidas informações referentes a criação do modelo, como quantidade de testes realizados, tempo gasto para criação e o resultado das métricas de qualidade de predição obtidas em sua criação.

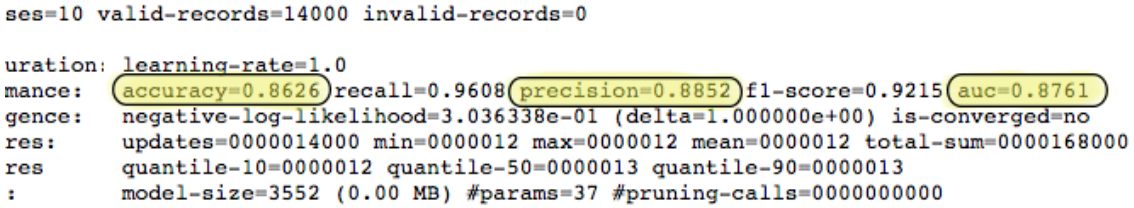


Figura 6: Exibição do Log da criação de um modelo.

A primeira delas, destacada na figura 6, é a acurácia, que mede a fração de resultados positivos reais entre os dados previstos como positivos, em outras palavras ela determina a proporção de predições corretas obtidas na validação realizada com os 30% dos dados sobre o modelo criado.

Segundo a AWS (2018), para determinar a taxa de acurácia, são considerados os casos de predições corretas e incorretas encontradas pelo modelo, no caso do estudo realizado, uma predição correta pode ser classificada como:

- Verdadeiro positivo – O modelo previu que o aluno apresentava grandes chances de evadir e o aluno realmente evadiu.
- Verdadeiro negativo – O modelo previu que o aluno não apresentava chances de evadir e o aluno realmente não evadiu.

Já uma predição incorreta pode ser classificada como:

- Falso positivo – O modelo previu que que o aluno apresentava grandes chances de evadir porém o aluno não evadiu.
- Falso negativo – O modelo previu que o aluno não apresentava chances de evadir porém o aluno evadiu.

A taxa de acurácia do modelo exibido na Figura 6 foi de 0.86, é importante destacar que quanto mais alta a taxa de acurácia do modelo, mais confiáveis são as predições realizadas a partir dele.

A segunda métrica de qualidade destacada na figura é a precisão, que segundo Tan et al. (2005) é considerada a mais intuitiva delas, ela mede a porcentagem de positivos reais entre as instâncias do modelo criado, ou seja, a porcentagem dos casos em que o modelo previu que o aluno iria evadir e ele realmente evadiu, como visto na figura 6 a taxa de precisão encontrada para o modelo de exemplo foi de 0.88, onde quanto maior o valor encontrado melhor será a precisão preditiva.

A terceira métrica destacada na figura 6 é a AUC (área sob uma curva), ela expressa o desempenho estimado do modelo. A AUC do modelo exibido na imagem foi de 0.87, que indica a capacidade do modelo criado de prever uma pontuação maior de resultados assertivos.

Assim como as métricas destacas anteriormente, a métrica AUC retorna um valor decimal de 0 a 1, onde no seu caso os valores mais próximos de 1 indicam um modelo de predição altamente preciso enquanto valores mais próximos a 0 indicam problemas com dados, no geral, bons modelos e classificadores devem estar entre 0.5 e 1 como destacado por autores como Hart et al. (2001) e Bradley, (1997).

O gráfico exibido na Figura 7 mostra as taxas de acurácia, precisão e AUC encontradas na geração dos 3 modelos não possuem uma grande variação entre si, visto que se baseiam de cálculos bastante parecidos como destacado anteriormente, porém observa-se uma variação considerável entre os 3 modelos gerados.

Como mostrado na figura 7 o modelo 3 obteve os maiores índices de métricas de qualidade, o que lhe fornece uma maior confiabilidade em suas predições, esse resultado está diretamente ligado a quantidade de atributos existentes na base de dados utilizada em sua criação, quanto maior a pluralidade da base de dados, mais preciso será o modelo. No entanto, a inclusão de muitas variáveis com pouco poder preditivo, que não estão relacionadas ao contexto do restante dos dados, pode interferir de maneira negativa na qualidade da precisão do modelo.

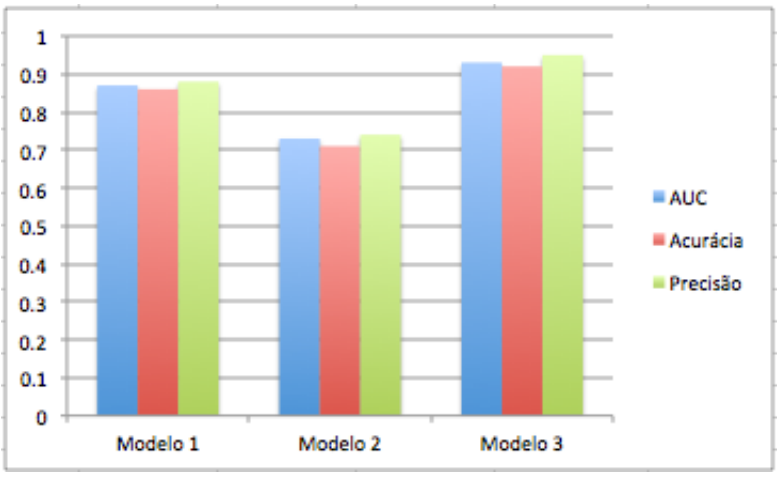


Figura 7: Comparativo das métricas de qualidade dos 3 modelos gerados

Concluída a etapa de geração dos modelos de *machine learning*, foram testados três cenários de predição para cada modelo gerado, onde foram inseridos dados de 3 alunos diferentes, aqui representados como alunos X, Y e Z.

A figura 8 mostra alguns dos atributos considerados nas predições, as colunas como *typeEducation, pendingDisciplines e wasChangeCourse*, foram cadastrados de maneira binaria, considerando 1 para verdadeiro e 0 para falso em cada coluna. As

colunas como *areaCourse* e *maritalStatus* foram cadastrados como o tipo numérico, onde cada valor representa um determinado status, como por exemplo no caso de *typeEducation*, diz respeito ao tipo de ensino de origem do aluno onde o valor 1 representa ensino público e 2 ensino privado.

i	studentId	age	gender	maritalStatus	admission	typeEducation	areaCourse	period	pendingDisci	wasChangeCourse
	x	25	0	2	2017	2	1	4	0	1
	y	34	1	2	2016	1	2	5	0	0
	z	21	0	1	2018	1	1	2	1	0

Figura 8: Alguns dos atributos dos 3 alunos utilizados para previsão

Nas previsões realizadas, os dados foram passados para o *Amazon ML* através de uma requisição ao *endPoint* gerado na criação de cada um do modelos de *machine learning*. O formato de resposta para as requisições feitas depende do modelo de *machine learning* que está sendo consultado na predição, o estudo de caso apresentado neste trabalho foi realizado com base em um modelo de predição binário que visa prever se um aluno irá evadir ou não, existem segundo a AWS (2018) dentro do *Amazon Machine Learning* outros dois modelos de predição, são eles o modelo multiclasse geralmente aplicado em casos onde o resultado da predição pode se dar em várias classes, como um modelo de vá identificar gêneros de filmes por exemplo, e também o modelo de regressão que pode retornar um valor de predição numérico como por exemplo um modelo que prevê a temperatura média de um determinado dia.

A figura 9 exibe os dados retornados na requisição feita para um dos alunos de exemplo, onde o valor *predictedScores* também chamado de pontuação de predição, indica a certeza do sistema de que o resultado da análise dos dados de acordo com o modelo pertence a classe positiva representada pelo valor 1, ou seja, a certeza do sistema de que o aluno irá evadir.

O exemplo da imagem resultou em 0.62, número esse que representa as chances de tal aluno evadir considerando que esse atributo varia de maneira decimal de 0 a 1 onde resultados mais próximos a 1 representam maiores chances do aluno evadir.

```
https://realtime.machinelearning.us-east-1.amazonaws.com
{
    "Prediction": {
        "predictedLabel: "1",
        "predictedScores": {
            "1": 0.621117353439331
        },
        "details": {
            "Algorithm": "SGD",
            "PredictiveModelType": "BINARY"
        }
    }
}
```

Figura 9: Retorno da predição realizada para um dos alunos em um dos modelos criados.

A figura 10 ilustra os resultados das predições obtidas para os 3 alunos utilizados no exemplo. A predição apontou estimativas que variam de 0.62 a 0.85 para o aluno X, 0.36 a 0.41 para o aluno Y, e 0.73 a 0.76 para o aluno Z nos 3 modelos gerados. Em cada um dos 3 modelos gerados para comparação, nota-se que para um mesmo aluno há uma variação considerável nos resultados da predição realizada, tal diferença se dá pela diversidade dos dados utilizados na criação dos modelos como destacado anteriormente.

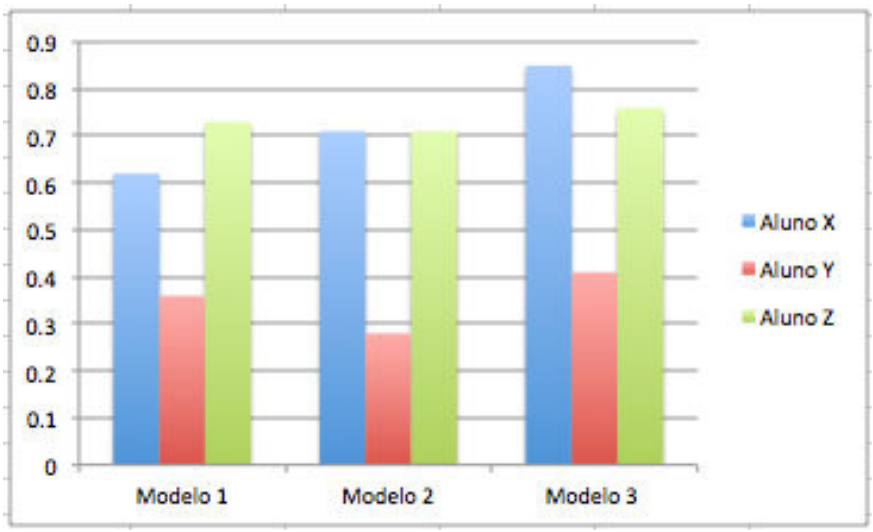


Figura 10: Comparação da predição realizada em cada um dos modelos gerados.

As taxas de predição podem variar muito de aluno para aluno, levando em consideração a base de dados utilizada para a geração do modelo de *machine learning,* por isso a importância de alimentar a base de dados com a maior e mais diversa quantidade de dados possível.

7. Conclusão

Das diversas propostas que visam resolver ou reduzir o problema da evasão nas universidades, a identificação precoce das chances de um aluno evadir ainda se mostra a mais viável, visto que a partir da identificação dos padrões dos perfis dos alunos evasores é possível prever quando um aluno irá ou não evadir.

Nos 3 modelos gerados para análise neste estudo, percebe-se uma variação considerável nas métricas de qualidade de predição obtidas em cada um, isso ocorreu devido a variação da quantidade de colunas da base de dados utilizada para a criação de cada modelo de *machine learning,* com isso foi visto que o modelo com a menor quantidade de colunas utilizadas para sua geração, obteve os menores índices de precisão, o que comprova uma realidade esperada: quanto maior a quantidade de dados e maior sua precisão, maiores serão os níveis de acurácia do modelo, que consequentemente fornecerá estatísticas de predição mais acuradas.

Os experimentos realizados demonstram que a abordagem proposta neste trabalho é apropriada para prever a evasão quando utilizadas bases reais de universidades, o que fornecerá uma pluralidade maior nos dados e consequentemente uma maior qualidade nos índices de predição encontrados, visto que dos 3 modelos gerados, o modelo 3 demonstrou o melhor índice de acurácia de 0.93, o que se deu devido a sua maior quantidade de colunas utilizadas na geração do modelo.

8. Considerações finais e trabalhos futuros

A sistemática utilizada neste trabalho é considerada um processo incremental constituída por uma série de experimentos com diversos tipos de dados. Neste artigo, foram apresentados os resultados obtidos em um estudo de caso onde foram empregadas as etapas da metodologia CRISP-DM para mineração de dados com o objetivo de identificar o percentual de risco de evasão de alunos.

Durante as pesquisas para a realização deste estudo, foi possível constatar a importância de análises e estudos em torno das causas da evasão escolar, visto que a evasão é um problema enfrentado na maioria das instituições de ensino, sejam elas públicas ou privadas e que acarretam diversos problemas tanto para as instituições quanto para os alunos que evadem. A utilização das técnicas e ferramentas de mineração de dados utilizadas possibilitou uma identificação prévia do percentual de chance que um aluno tem de evadir da instituição.

O foco da pesquisa, foi identificar padrões e dados correlatos entre grupos de alunos já evadidos para geração de modelos de *machine learning*, e assim realizar a análise de dados de atuais alunos para prever a chance deste determinado aluno evadir, de acordo com seus dados. A aplicação desta proposta em bases de dados reais de instituições será de grande relevância e de fácil aplicação, visto que a análise realizada leva em consideração dados gerais de fácil controle como, período cursado, gênero, idade, cor da pele, se o aluno veio de escola pública ou não etc.

Com o objetivo de continuidade deste trabalho, é considerada a utilização de um banco de dados não relacional como por exemplo o *DynamoDB,* onde os dados das previsões serão salvos e posteriormente disponibilizados para o sistema da universidade com o uso da arquitetura *rest*, o que facilitará a integração com qualquer tipo de sistema utilizado nas universidades, podendo assim serem analisados de maneira contínua e ações preventivas poderão ser tomadas.

Os resultados obtidos foram satisfatórios, sendo considerados um ponto de partida para a construção de uma ferramenta que além de analisar de maneira mais ampla os dados de alunos, consiga criar padrões qualitativos de análise, levando em consideração não apenas dados relacionados ao histórico escolar do aluno, mas também dados socioeconômicos, o que forneceria uma precisão maior quanto as chances de evasão visto que um aluno pode evadir por fatores além de desempenho ou histórico acadêmico.

Das principais vantagens da utilização da abordagem proposta, está na combinação de várias tecnologias e métodos de mineração de dados em uma mesma predição, o que otimiza o resultado do processo. Trabalhar com tecnologias em nuvem também foi de grande ganho para este estudo, considerando os benefícios oferecidos por essa tecnologia como segurança, qualidade do serviço de dados, escalabilidade, dentre outros.

Com bases de dados reais da instituições de ensino, acredita-se que os resultados destas previsões sejam cada vez mais aprimorados, visto que a variação dos dados analisados será maior abrangendo mais situações que acontecem na vida acadêmica da população em geral.

Referências

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2009) “Investimentos Públicos em Educação”, http://portal.inep.gov.br/estatisticas-gastoseducacao e “Censo da Educação Superior’, http://portal.inep.gov.br.

Amazon Web Services. Disponível em: <https://aws.amazon.com/pt/>. Acesso em 03/07/2018.

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

SCHREIBER , J. N. C.; BESKOW, A. L. B.; NARA , E. O. B.; DA SILVA , J. I.; DE MORAES, J.; NAJDZION, V. M. (2017) “SOFTWARE SDBAYES: UM AUXÍLIO PARA A PREDIÇÃO DE EVASÃO DISCENTE”

Kantorski, G. Z.; Hoffmann, I. L.; Limberger, S. J. And Muller, F. M.. Uma Visão Do Futuro: Previsão De Evasão Em Cursos De Graduação Presenciais De Universidades Públicas: O Caso Do Curso De Zootecnia. XV Colóquio Internacional De Gestão Universitária. ISBN: 978-85-

68618-01-1. Mar Del Plata – Argentina. 2015.

LYKOURENTZOU, I.; GIANNOUKOS, I.; NIKOLOPOULOS, V.; MPARDIS, G.; LOUMOS, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. Journal Computers & Education, New york, v. 53, n. 3, p. 950-965, 2009.

CARVALHO, Cristina Helena Almeida. Estudo comparado sobre a expansão do ensino superior: Brasil e Estados Unidos. In: ESCENARIOS MUNDIALES DE La EDUCACIÓN SUPERIOR. Análisis global y estudios de casos. Buenos Aires: CLACSO (Consejo Latinoamericano de Ciencias Sociales), 2007.

NUNES, Edson. Desafio estratégico da política pública: o ensino superior brasileiro. Revista Administração Pública, Rio de Janeiro, n. 41, p. 103-147, 2007.

KIRA, Luci Frare. A evasão no ensino superior: o caso do curso de pedagogia da Universidade Estadual de Maringá (1992 – 1996). 106p. Dissertação (Mestrado em Educação). Universidade Metodista de Piracicaba, Piracicaba-São Paulo, 2002.

Barroso, M. F. e Falcão, E. B. M. (2004) "Evasão Universitária: O Caso do Instituto de Física da UFRJ", IX Encontro Nacional de Pesquisa em Ensino de Física.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, England, 1996, p.1-37.

Ng, Raymond & Han, Jiawei. (2002). CLARANS: A method for clustering objects for spatial data mining.

Tan, P., Steinbach, M., and Kumar, V. (2007). Introduction to Data Mining. Pearson international Edition. Pearson Addison Wesley.

Costa E., Baker R. S. J., Amorim L., Magalhães J. e Marinho T. (2012) "Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações." Anais Jornada de Atualização em Informática na Educação.

Santos, H. L. dos; Camargo, F. N. P.; Camargo, S. da S. (2012). "Minerando Dados de Ambientes Virtuais de Aprendizagem para Predição de Desempenho de Estudantes". In: Seventh Latin American Conference on Learning Objects and Technologies, Guayaquil. Proceedings of the 7th LACLO.

Martins, L. C., Lopes, D. A., & Raabe, A. (2012). Um Assistente de Predição de Evasão aplicado a uma disciplina Introdutória do curso de Ciência da Computação. In: Anais do XXIII Simpósio Brasileiro de Informática na Educação.

Gotardo, R., Cereda, P. R. M., & Junior, E. R. H. (2013). Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores. In: Anais do XXIV Simpósio Brasileiro de Informática na Educação

Digiampietri, L. A., Nakano, F., & Lauretto, M. (2016). Mineração de Dados para Identificação de Alunos com Alto Risco de Evasão: Um Estudo de Caso. *Revista De Graduação USP*, *1*(1), 17-23.

dos Santos, R. N., de Alburqueque Siebra, C., & Oliveira, E. S. (2014). Uma Abordagem Temporal para Identificação Precoce de Estudantes de Graduação a Distância com Risco de Evasão em um AVA utilizando Árvores de Decisão. In: Anais dos Workshops do III Congresso Brasileiro de Informática na Educação.

Adeodato, P. J., Santos Filho, M. M., & Rodrigues, R. L. (2014). Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar. In: Anais do XXV Simpósio Brasileiro de Informática na Educação.

Costa, F., dos Santos Silva, A. R., de Brito, D. M., & do Rêgo, T. G. (2015). Predição de sucesso de estudantes cotistas utilizando algoritmos de classificação. In: Anais do XXVI Simpósio Brasileiro de Informática na Educação.

GAIOSO, N. P. L. Evasão discente na educação superior: a perspectiva dos dirigentes e dos alunos. Brasília: UCB, 2005, 99 P.

HART, P. E.; STORK, D. G.; DUDA, R. O. Pattern classification. John Willey & Sons, 2001.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7):1145–1159, 1997.

Tan, P.-N.; Steinbach, M. & Kumar, V. (2005). Introduction to Data Mining. Addison-Wesley.