Caio Ferreira Bernardo – 9276936
Caique Honório Cardoso – 8910222
Matheus Aparecido do Carmo Alves – 9791114
Rafael Rodrigues Santana – 7594375
William Luis Alves Ferreira – 9847599

Projeto 2 da Disciplina Inteligência Artificial: Predição de evasão de cursos

São Carlos, SP

11 de junho de 2019

Sumário

1 Introdução

Ao se explorar a área de *Inteligência Artificial (IA)* em Ciência da Computação, diversos tipos de problemas são encontrados e diferentes soluções são propostas. Dentro deste contexto, a *Mineração de dados* surge como importante ferramenta para estudo e apresentação de soluções em cenários que podem ser estudados sobre uma análise orientados a dado. Sobre esta abordagem e com o suporte da utilização de técnicas de mineração de dados, torna-se possível extrair informações úteis a partir de uma base de dados, a qual se é inviável ou impossível de se analisar de forma manual. Através de um correto tratamento dos dados, é possível se viabilizar a criação ou aplicação de um modelo computacional que realizará está tarefa, extraindo informação de maneira automatizada. Este processo composto pelo tratamento de dados, processamento de informação e extração de informação via dados é o que denomina-se como um processo de aprendizado de máquina via mineração de dados.

Esse aprendizado pode ocorrer de forma *supervisionada* (quando o processo de aprendizado de maquina é guiado por um "professor" que possui conhecimento sobre o resultado esperado em cada etapa de treinamento), *não-supervisionada* (quando a máquina não possui conhecimento prévio de resultados esperados e não possui um "guia" de aprendizado) ou *semi-supervisionada* (possui "professor"externo para apenas parte dos exemplos de treinamento). Independente da forma de aprendizado escolhido, o objetivo é tornar viável e com resultados satisfatórios a resolução de um problema, como por exemplo identificação de doenças a partir de imagens ou dados do paciente, identificação de nichos de mercado para marketing direcionado, reconhecimento de faces, voz, assinaturas, e assim por diante.

Uma possível definição para este processo é:

> "Um programa de computador *"aprende"* a partir de uma *experiência* E com respeito a alguma *classe de tarefas* T e *medida de desempenho* P, se seu desempenho em tarefas de T, medido por P, melhora com a experiência E." (Mitchell, 1997)

Tomando esta definição, para que seja possível utilizar um algoritmo de aprendizado sobre um problema escolhido, é necessário se encontrar: **(i)** a classe de tarefas alvo; **(ii)** uma medida de desempenho que se deseja melhorar, e; **(iii)** experiências para se analisar ou gerar aprendizado.

Dado esta breve introdução, este projeto possui como principal propósito aplicar um algoritmo de mineração de dados utilizando um modelo de aprendizado de máquina *supervisionado* para avaliar o desempenho de alunos em um curso a partir de suas notas nas atividades e predizer quais deles irão sair do curso antes de sua conclusão.

Neste trabalho serão apresentados a base de dados escolhida para extração de conhecimento assim como a motivação do problema de estudo (Seção 2), uma discutição aprofundada sobre a metodologia de desenvolvimento utilizada (Seção 3), os resultados obtidos no desenvolvimento da proposta (Seção 4), e uma análise geral sobre o desenvolvimento e os resultados obtidos durante a execução do trabalho (Seção 5).

2 Motivação da análise e Descrição da base de dados

Para estudo e análise dentro deste documento, escolheu-se implementar um método para Predição de Evasão de Cursos Online.

A análise do desempenho de alunos é de grande importância, tanto para melhorar a tomada de decisão da instituição para melhorar seu curso, como para auxiliar o processo de ensino, melhorando a avaliação do professor e, consequentemente, o aprendizado do aluno.

O curso *Big Data Analytics da Future Learn* [1] disponibiliza um banco de dados sobre as submissões de atividades de alunos cadastrados em seus cursos. Mais especificamente, o curso de análise compreende quatro módulos online, cada um apresentando e analisando um elemento diferente de aplicação e teoria de *big data*, sendo ministrado pela *Universidade de Tecnologia de Queensland*. Os criadores pensando na mesma análise exposta anteriormente, geram *logs* de acessos dos alunos matriculas ao longo das 12 semanas do programa. Este banco de *logs* é de domínio publico desde que não proporcione artigos acadêmicos ou de caráter lucrativo no formato CSV.

Resumidamente, os logs registram a permanência dos alunos, histórico de acertos, erros e questões não realizadas ao longo do tempo durante as semanas que compreendem o curso, além de dados dos alunos como: sexo, faixa etária, país, momento de contratação do curso. Estas bases de dados são indexadas pelo uso de chaves para identificação de alunos, orientadores e desenvolvedores do curso. Para o estudo proposto neste documento, os dados dos orientadores e desenvolvedores não compreendem a analise a ser realizada. Desta forma, o foco de análise será definido apenas 2 dos 6 arquivos disponibilizados, *bigdata2-enrolments e bigdata2-question-response*, cujo os campos são apresentados na Tabela 1. Essa decisão será melhor explicada na Seção 3.2, onde se explora e descreve pré-processamento para este banco de dado.

[1] Big Data Analytics Course: www.futurelearn.com

Arquivo de Log	Campos Disponíveis e Descrição
bigdata2-enrolments.csv	(0) learner_id: chave identificadora do cadastro; (1) enrolled_at: data e horário de cadastro; (2) unenrolled_at: data e horário de cancelamento de cadastro; (3) role: categoria de cadastro (aluno, professor desenvolvedor); (4) fully_participated_at: data e horário de finalização do curso; (5) purchased_statement_at: data e horáriro da compra de um curso; (6) gender: sexo informado no cadastro; (7) country: país de origem; (8) age_range: intervalo de idade; (9) highest_education_level: nível de escolaridade informado no cadastro; (10) employment_status: vínculo empregatício informado no cadastro; (11) employment_area: área profissional de atuação (12) detected_country: país detectado de acesso ao curso;
bigdata2-question-response.csv	(0) learner_id: chave identificadora do cadastro; (1) quiz_question: número do conjunto de questões; (2) week_number: semana proposta para questão; (3) step_number: número do passo da questão; (4) question_number: número da questão; (5) response: resposta submetida; (6) submitted_at: horário de submissão; (7) correct: identificador se a questão foi respondida corretamente;

Tabela 1 – Apresentação dos arquivos *log* utilizados para desenvolvimento do método proposto, seus campos e descrição.

3 Metodologia

Nesta seção será apresentada a metodologia de desenvolvimento do projeto, que seguiu a metodologia de um processo de mineração de dados apresentada em sala de aula e ilustrada pela Figura 1.

Esta Seção está organizado da seguinte forma: na Seção 3.1, será discutido o problema identificado na base de dados utilizada, cuja solução é o objetivo do projeto. Na Seção 3.2, serão discutidas as transformações realizadas na base de dados original e a base de dados resultante a ser utilizada no processo de aprendizado. Na Seção 3.3, será discutida a criação da *rede neural MLP* (multi-layer-perceptron) utilizada no trabalho, justificando sua escolha e abordando suas vantagens e desvantagens. Na Seção 3.4, uma análise dos resultados iniciais do projeto será realizada e possíveis alterações serão propostas. Por fim, na Seção 3.5, será discutida a utilização do conhecimento extraído no projeto e como isso pode impactar a forma como a evasão de alunos é tratada atualmente.

Figura 1 – Ilustração da definição de um processo de mineração de dados.

3.1 Identificação do Problema

Um importante ponto para análise de plataformas de ensino a distância (EAD) é o método de avaliação possível para auxiliar os orientadores. A plataforma de atividades EAD precisa ser capaz de oferecer uma análise robusta do desempenho de cada aluno, coerente com o método de ensino aplicado.

As plataformas como *E-disciplinas* e o portal *Tidia-ae* (utilizada pela Universidade de São Paulo), apresentam diversas funcionalidades, e.g., envio de atividade, aplicação de exercícios múltipla escolha online, fórum da turma, chat e agenda de atividades. O problema e a vantagem associados a esses tipos de abordagem, é a diminuição do contato aluno professor e a otimização da proposta de atividades, assim como sua rápida revisão e correção pelo corpo discente e docente, respectivamente. Contudo, entende-se que nestas plataformas EAD, existe uma carência de ferramentas que auxiliam na análise da curva evolutiva no curso e de aprendizado, tornando-se difícil lidar com certos casos onde extrair esta informação é complexa e, por muitas vezes, inviável dado as condições de ensino.

Sobre nosso caso de estudo, no qual o ensino é feito de forma totalmente online, causa a carência da análise e extração da curva evolutiva dos alunos, pois torna-se um fator crítico ao permitir maximizar o acompanhamento do aluno por parte do corpo docente e, até mesmo, possibilitando a automatização de ferramentas didáticas que faça com que o aluno volte a curva esperada de aprendizado; com isso, existe a necessidade de ferramentas que analisam e engaje os alunos, desta forma, visando reduzir a evasão associada aos cursos e, portanto, aumentando a qualidade do aprendizado.

Para o estudo neste documento, vale ressaltar, que a partir de varias turmas formadas temos o acumulo de eventos sendo proporcional a redução de situações não prevista em uma quantidade inferior de dados, ou seja, aumentando o número de eventos diminuímos a oportunidade de ocorrer falsos negativos ou falso positivos em nossa predição, essa mecânica de aprender com turmas passadas pode ser facilmente quantificada em cursos totalmente EAD, assim inaugurando um ensino escalonável e adaptativo de forma padronizada e confiável, nunca antes possibilitado pelo ensino presencial; com tudo, não restringe que em uma mesma iteração do curso (turma) não seja possível realizar o treinamento da Rede neural descrita no Secção 3.2.

3.2 Pré-Processamento

A etapa de pré-processamento definida pelo grupo pode ser descrita em 6 passos:

1. **Identificar os atributos de importância para o estudo proposto:** este passo busca identificar todos os arquivos de dados relevantes na base de dados fornecida visando reduzir a quantidades de parâmetros necessários para análise e otimizar o

aprendizado que se deseja obter no processo de mineração. Em linhas gerais, foram selecionados 2 de 6 arquivos CSVs e sobre estes mantivemos o foco de mineração. Essa decisão foi tomada pois o grupo pretende somente realizar uma previsão de evasão como uma ferramenta de suporte para o tutor do curso, cabendo a ele identificar deficiências juntamente ao aluno (uma vez que, para ensino, os obstáculos vão além da avaliação em um teste online e é necessário um acompanhamento efetivo e um acompanhamento próximo ao aluno para um bom diagnostico de aprendizado).

2. **Recolher as informações de cadastro:** definidos o conjunto de arquivos úteis pertencente a base de dados principal, foram recolhidos todos os dados de cadastro existente. Esta informação será utilizada para identificação de todas as entradas usadas no aprendizado.

3. **Marcar alunos desistentes e não desistentes:** com os dados recolhidos e indexados pelos identificadores de cadastro, realiza-se a extração dos alunos e sua classificação de alunos desistentes e não desistentes. Essa classificação é feita com base no dado de cancelamento de cadastro do aluno e pre-processamento do arquivo *bigdata2-enrolments.csv*. As informações recolhidas e de interesse no programa para este processo de extração são os campos (0) learner_id, (1) enrolled_at e (2) unenrolled_at, descritos na Tabela 1. Para a classificação, criou-se um sinalizador *True* ou *False* e todos os alunos foram marcados baseado na informação dos campos (1) enrolled_at e (2) unenrolled_at – caso existisse cadastro e não existisse cancelamento do cadastro, marque *True*; caso contrário, marque *False*.

4. **Criar relacionamento cadastro-submissões:** com a identificação de todos os cadastros, inicia-se a se criar a relação cadastro-submissões que o banco oferece. A partir do pre-processamento do arquivo *bigdata2-question-response.csv*, cria-se a associação de cadastro e de questões para todas as submissões e para cada identificador (chave) de cadastro. Esta associação possibilita que nosso método possa aprender, a cada exercício novo lançado para avaliação, quais configurações dessa matriz podem predizer qual aluno e com qual probabilidade ele pode desistir ou não do curso. Desta forma, o pré-processamento sumariza e disponibiliza uma experiência otimizada para abordagem do problema.

5. **Separação dos conjunto de dados desistentes e não desistentes:** após a realização da associação cadastro-submissões, é feita uma divisão em dois sub conjuntos de dados de alunos desistentes e não desistentes. Essa divisão é feita, pois procura-se realizar o treinamento com porcentagens iguais de alunos desistentes e não desistentes – otimizando o futuro processo de aprendizado.

6. **Transformação de dado em experiências (matrizes de entrada):** dado todos os passos anteriores de pre-processamento, este passo visa transformar todos os

relacionamentos e conjuntos extraídos em entradas válidas para o método de aprendizado escolhido – no caso, uma rede neural Multi Layer Perceptron (MLP). Desta forma, todos os dados são transformados em dados numéricos e formatados como matrizes. São criadas n matrizes de ordem n, com n sendo o número de questões aplicadas durante o curso, que disponibilizam de forma incremental todos os exercícios não submetidos, submetidos com a resposta certa e submetidos com a resposta errada.

Estes passos definem integralmente o pré-processamento realizado, que também estão descritos e comentados no código. A decisão sobre a abordagem de aprendizado utilizada (na aplicação de uma MLP) será melhor justificada adiante na Seção 3.3.

3.3 Processo de Aprendizado de Máquina

O aprendizado é realizado pelo método de Redes Neurais Artificiais. A arquitetura utilizada foi definida de maneira simples: quatro camadas completamente interligadas. A saída da rede é binaria, representando se o aluno desistiu ou não do curso.

Para o processo de aprendizado, o conjunto de alunos foi embaralhado e separado em conjunto de testes e conjunto de treinamento, com proporções de 30% e 70% respectivamente. Após separados os conjuntos, as n matrizes de um aluno por vez são inseridas ao mesmo tempo na rede neural, de modo a otimizar o processo de aprendizagem, acelerando a quantidade de dados processada em cada iteração.

Após passar todos os alunos do conjunto de treinamento pela rede neural, é realizada a fase de experimentos. Nesta fase, cada uma das n matrizes de cada aluno é entregue individualmente para a rede neural, e é registrado o resultado obtido.

3.4 Pós-Processamento

Após a implementação da rede neural, aplicou-se os dados de teste a mesma e verificou-se os resultados. Apesar dela apresentar uma boa taxa de acertos, um problema se destacou durante os testes: A rede inicial só era capaz de fazer predições com base na entrada de todas as atividades do curso. Ou seja, não havia como fazer uma predição em tempo real, já que para entrar com os dados de todas as atividades dos alunos, o curso já teria chegado ao fim.

Por isso o processo, que é cíclico, voltou à etapa de pré-processamento, onde foi gerada uma matriz de entrada para cada semana do curso, passando à etapa de desenvolvimento, onde a rede foi adaptada para suportar essa mudança (as alterações estão

descritas nas respectivas seções), e então chegou-se à etapa de pós-processamento pela segunda vez.

Dessa vez, os testes tiveram bastante sucesso, inclusive aumentando a velocidade de treinamento da rede, além de resolver o problema principal, o de predição em tempo real, antes do término do curso. Dado o sucesso dessa iteração, não foi preciso fazer mais mudanças no trabalho, passando-se assim para a ultima fase do processo: A utilização do conhecimento.

3.5 Utilização do Conhecimento

Após a conclusão da extração de padrões, temos como resultado o conhecimento do comportamento dos alunos que tem maior probabilidade de sair do curso antes do fim. A partir desse conhecimento, as entidades que aplicam tais cursos podem criar diversas abordagens para estimular esses alunos a continuarem no curso até sua finalização.

Por exemplo, inicialmente, pode-se entrar em contato com os alunos por e-mail ou por telefone e perguntar a eles se eles pensam em sair do curso antes de sua conclusão e qual seria o principal motivo, caso isso viesse a se tornar realidade.

Caso o motivo da saída do curso seja a dificuldade do aluno de acompanhar a matéria, elas podem direcionar ofertas de aulas complementares para os alunos que tenham dificuldades em acompanhar o curso normal.

Caso seja um problema de demanda muito grande de tempo, elas podem direcionar ofertas de cursos menos densos e que permitam um aprendizado mais lento.

Caso o problema seja desmotivação ou desinteresse pela matéria que está sendo ensinada, lembrar esses alunos de todas as vantagens que eles irão ter caso persistam no curso até o seu fim pode ser um bom motivador para que eles não saiam antes do tempo.

E caso o problema seja que a expectativa do curso não coincidiu com a realidade, pode-se perguntar a eles o que eles esperavam do curso e que não foi alcançado, para que assim essas características possam ser incorporadas em cursos futuros, tornando-os assim mais interessantes.

Certamente existem diversas outras possibilidades de iniciativas para se reter os alunos, mas conhecer quais alunos tendem a sair é essencial para se direcionar essas iniciativas aos alunos certos.

4 Resultados

Nesta seção serão apresentados os resultados coletados para a rede MLP desenvolvida.

Inicialmente, foi realizada a divisão da base de dados final em dois grupos, o de treinamento (com 70% da base total original) e o de teste (com 30% da base total original). Note que foi utilizando uma abordagem de teste e treinamento por *divisão estratificada* (i.e., mantendo a porcentagem de grupos de treinamento e teste com porcentagens fixa) sobre os dois grupos obtidos no pré-processamento (alunos desistentes e não desistentes).

A fase de treinamento (utilizando a base de dados disponível) apresentou rápida convergência de aprendizado, executando essa etapa em um tempo médio de 2 minutos. Este resultado é um resultado interessante uma vez que, como será apresentado a seguir, alcançou-se uma classificação interessante para os casos de testes aleatórios definidos.

Ao final do treinamento, aplicou-se o modelo desenvolvido ao grupo de testes, coletou-se os resultados de taxa de acerto e taxa de erro, e construiu-se uma matriz de confusão - apresentada na figura 2 - demonstrando a efetividade do algoritmo. Foram coletados resultados de 20 execuções (i.e., reiniciando o estado estado da rede e reexecutando todo o processo definido neste documento) e a análise foi feita sobre a média das execuções.

Este resultado é interessante, uma vez que porcentagens significativas de classificações corretas foram atingidas a um baixo custo de processamento e sem a necessidade de um banco de dados *Big Data*.

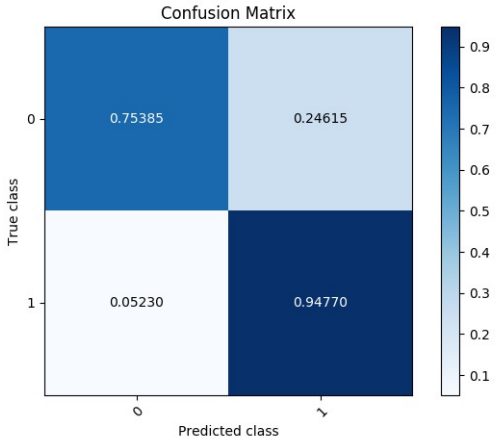

Figura 2 – Matriz de confusão produzida pelos resultados da classificação.

5 Conclusão

Neste documento, desenvolveu-se uma proposta de um método de mineração de dados para estudo e abordagem sobre o problema de taxa de evasão em cursos EAD. Aplicando-se a metodologia estudada durante o curso de Inteligência Artificial e seguindo o processo formal definido por esta, foi possível se apresentar resultados que justificam a eficiência esperada na aplicação de um processo de Mineração de Dados para resoluções de problemas sobre a ótica orientada a dados.

Através da modelagem proposta, foi realizada uma analise da eficiência associada ao processo de *mineração de dados* para habilitar a extração de informações e um *aprendizado de máquina*. Especificamente, implementou-se um algoritmo de aprendizado *supervisionado* para construção de conhecimento modelado para uma *rede neural MLP*.

Para resolução do problema e aplicação do algoritmo proposto, foi realizada uma análise e tratamento da base de dados original, buscando-se identificar os dados essenciais para o modelo, eliminando os dados ruídos e desnecessários para a análise que pudessem diminuir o desempenho da rede (otimizando o tempo necessário para treinamento e resposta da rede). Frente a isso, os dados foram pré-processados e transformados para aplicação no modelo, treinado para identificar tendencias de evasões, e então testados em simulações de um cenário com *fluxo de dados* na etapa de pós-processamento.

Pela análise dos resultados, a *rede neural MLP* apresentou resultados satisfatórios, sendo capaz de predizer e classificar número significativo de casos de desistência de alunos em tempo real, considerando somente o perfil de atividades entregues do aluno. Devido as propriedades associadas as redes neurais MLP, não se pode dizer a causa da desistência prevista, apenas se o aluno possui alguma probabilidade de evasão do curso. Entretanto, reitera-se que o grupo compreende que este modelo apresenta uma solução válida e suficiente para que exista uma tomada de decisão ativa por parte dos educadores para evitar desistências ou, ao menos, melhorar o processo de ensino como um todo pela identificação e análise de insatisfações ou dificuldade de alunos desistentes.

Por fim, entende-se a importância deste projeto realizado e, o grupo, julga que a oportunidade de aplicação e desenvolvimento de um processo de mineração de dados sobre um cenário real (como proposto pela atividade) é capaz de proporcionar uma significativa evolução de entendimento e maturação do conhecimento atrelado a disciplina.