INSTITUTO ENSINAR BRASIL

FACULDADE DOCTUM DE CARATINGA

MIQUÉIAS MATIAS CAETANO

USO DE INTELIGÊNCIA ARTIFICIAL PARA DETECÇÃO DE EVASÃO ESCOLAR EM INSTITUIÇÕES DE ENSINO: UM ESTUDO DE CASO EM UMA INSTITUIÇÃO DE ENSINO SUPERIOR

CARATINGA

2019

MIQUÉIAS MATIAS CAETANO

FACULDADE DOCTUM DE CARATINGA

USO DE INTELIGÊNCIA ARTIFICIAL PARA DETECÇÃO DE EVASÃO ESCOLAR EM INSTITUIÇÕES DE ENSINO: UM ESTUDO DE CASO EM UMA INSTITUIÇÃO DE ENSINO SUPERIOR

Trabalho de Conclusão de Curso apresentado ao Curso de Ciência da Computação da Faculdade Doctum de Caratinga, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Área de Concentração: Uso de inteligência artificial para detecção de evasão escolar.

Orientador: Professor Hudson Silva de Souza.

CARATINGA

2019

rede de ensino DOCTUM	FACULDADES DOCTUM DE CARATINGA	FORMULÁRIO 9
	TRABALHO DE CONCLUSÃO DE CURSO	
	TERMO DE APROVAÇÃO	

TERMO DE APROVAÇÃO

O Trabalho de Conclusão de Curso intitulado: USO DE INTELIGÊNCIA ARTIFICIAL PARA DETECÇÃO DE EVASÃO ESCOLAR EM INSTITUIÇÕES DE ENSINO: UM ESTUDO DE CASO EM UMA INSTITUIÇÃO DE ENSINO SUPERIOR, elaborado pelo(s) aluno(s) MIQUÉIAS MATIAS CAETANO foi aprovado por todos os membros da Banca Examinadora e aceito pelo curso de CIÊNCIA DA COMPUTAÇÃO das FACULDADES DOCTUM DE CARATINGA, como requisito parcial da obtenção do título de

BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Caratinga 10/12/2019

HUDSON SILVA
Prof. Orientador

RICARDO BOTELHO
Prof. Avaliador 1

ELIAS GONÇALVES
Prof. Examinador 2

O Deus que me criou e deu o fôlego de vida, que até aqui me ajudou e me guardou debaixo de sua poderosa mão.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por seu infinito amor, bondade e graça. Por ter me dado forças e condições para chegar até aqui, reconheço que sem Ele, esse momento não seria possível.

Agradeço também a minha família e amigos pelo apoio que sempre me deram durante toda a minha vida, pelas palavras de incentivo e confiança em mim.

Deixo um agradecimento especial ao meu orientador pela dedicação do seu escasso tempo ao meu projeto de pesquisa.

Também quero agradecer a Faculdade Doctum de Caratinga e a todos os professores do meu curso pela elevada qualidade do ensino oferecido.

“Tudo quanto te vier à mão para fazer, faze-o conforme as tuas forças.”

Eclesiastes 9:10ª

LISTA DE SIGLAS

BDOO - Banco de dados orientado a objetos

BDR - Banco de dados relacional

ETL - Extract Transform Load

KDD - Knowledge Discovery in Databases

MLP – Multilayer Perceptron

SGBD – Sistemas Gerenciadores de Bancos de Dados

WEKA - Waikato Environment for Knowledge Analysis

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

RESUMO

Com o avanço tecnológico, empresas e instituições se tornaram dependentes de tecnologias que auxiliem em suas gestões de negócios a fim de reduzirem prejuízos e aumentarem seus lucros. Para uma instituição de ensino uma das formas de obter prejuízo financeiro é a falta de alunos ou uma excessiva evasão escolar. Diante dessas situações, pensou-se no uso de uma inteligência artificial capaz de prever tais evasões e combatê-las. Para tal objetivo, utilizou-se uma base de dados para criação do Data Warehouse, aliado ao Pentaho Data Integration para a mineração dos dados, além da utilização da ferramenta WEKA e algoritmos de rede neural artificial Multilayer Perceptron para prever futuras evasões e dar chances para tomadas de decisões mais rápidas e eficientes.

Palavras-chave: Inteligência Artificial. Evasão Escolar. Data Warehouse. Pentaho Data Integration. Multilayer Perceptron.

ABSTRACT

With technological advancement, companies and institutions have become dependent on technologies that aid their business management in order to reduce losses and increase their profits. For an educational institution one of the ways to make a financial loss is the lack of students or excessive dropout. Faced with these situations, it was considered the use of artificial intelligence capable of predicting such evasions and combating them. For this purpose, we used a database to create the Data Warehouse, combined with Pentaho Data Integration for data mining, as well as the use of WEKA tool and Multilayer Perceptron artificial neural network algorithm to predict future evasions and give chances for faster and more efficient decision making.

Key-words: Artificial Intelligence. School Dropout. Data Warehouse. Pentaho Data Integration. Multilayer Perceptron.

SUMÁRIO

8.1 ANEXO 1: Autorização de divulgação de dados

8.2 ANEXO 2: Questionário socioeconômico

1. INTRODUÇÃO

A evasão escolar é a prática do abandono a instituição educacional por decorrência de algum motivo. Esses motivos podem ser diversos, variando de acordo com a idade, nível escolar e condições socioeconômicas do indivíduo. Por meio de estudos e pesquisas realizadas sobre o assunto, fica evidente que a evasão escolar é um grande problema que atinge o país, provocando impactos negativos a médio e longo prazo e afeta diretamente a economia.

Tais desistências também podem acarretar perdas e prejuízos para a instituição, os estudantes, a sociedade e até o governo. Como as causas que levam a evasão escolar são várias, por exemplo, financeira e social, fica impossível para o gestor da instituição conseguir prever futuras evasões em um ambiente com grande número de alunos e dar uma solução para essas evasões.

Diante disso, foi pensado na possibilidade da utilização de inteligência artificial como uma metodologia capaz de prever padrões e determinar os futuros discentes da instituição que poderão evadir. Assim, dará chances para o gestor daquela instituição agir e buscar junto ao aluno, formas de mantê-lo no curso.

Este trabalho divide-se em quatro seções principais além da introdução, sendo elas: Referencial Teórico, Metodologia, Resultados e Conclusão.

No Referencial Teórico são descritos os conceitos teóricos utilizados para a obtenção do resultado final e a conceituação das ferramentas que serão utilizadas no trabalho. Na Metodologia são apresentados os passos seguidos e a utilização das ferramentas descritas no Referencial Teórico deste trabalho. Nos Resultados são discutidos os dados obtidos pela rede neural artificial sobre a previsão de evasão dos discentes, além de exibir gráficos e tabelas descrevendo esses dados e por fim a Conclusão com uma análise final dos resultados e a utilização de inteligência artificial em prever futuras evasões com base nas respostas do questionário socioeconômico dos alunos de graduação.

2. REFERENCIAL TEÓRICO

Para uma melhor compreensão da aplicabilidade da Inteligência Artificial no ramo educacional, foi desenvolvido um estudo pesquisando trabalhos já realizados na área, suas ferramentas e os impactos para a instituição. O objetivo deste estudo é o desenvolvimento de uma ferramenta eficiente e confiável, capaz de detectar futuras evasões de alunos na instituição de ensino por meio de algoritmos de rede neural.

2.1. EVASÃO ESCOLAR

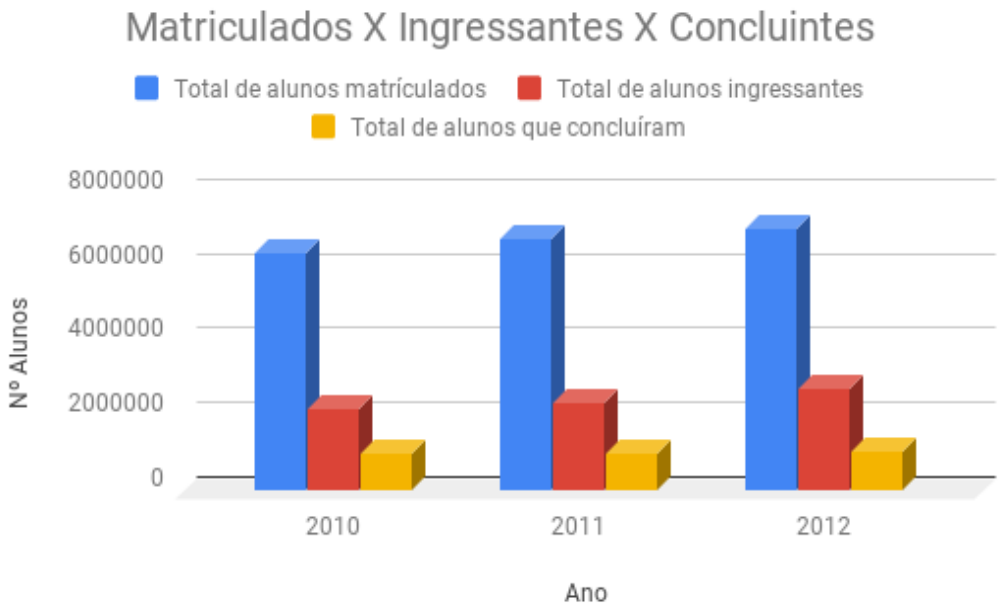
Palharini (2010, p.13) compreende a evasão como sendo "a saída definitiva do aluno do curso de origem sem concluí-lo". Além dessa definição, Palharini (2010, p.13) acrescenta:

> São consideradas as seguintes formas de saída: o aluno não se matricula e abandona o curso; o aluno comunica oficialmente a desistência; o aluno opta pela transferência para outro curso da mesma instituição; o aluno é excluído por norma institucional, o aluno opta por transferir-se para o mesmo curso em outra instituição.

Silva Filho, Motejunas, Hipólito e Lobo (2007) reforçam que essas saídas podem causar prejuízos, provocando graves consequências sociais, acadêmicas e econômicas.

No Brasil, a evasão escolar é um problema de longa data, sem uma solução definitiva, ela ocupa até os dias atuais grande relevância e importância. Sendo assim, o Governo Federal juntamente com os estados e municípios buscam desenvolver políticas públicas voltadas para a educação. Porém, se for pego os dados dos censos de 2010, 2011 e 2012, o total de matrículas em cursos superiores foram de 6.407.733, 6.765.540 e 7.058.084, respectivamente. Por sua vez, o número de alunos que concluíram foi de 980.662, 1.022.711 e 1.056.069 e o número de ingressantes foi de 2.196.822, 2.359.409 e 2.756.773 (FRITSCH; ROCHA; VITELLI, 2015). Sendo possível notar um elevado índice de evasão.

Figura 1 - Número de alunos matriculados, ingressantes e concluintes ao longo de 3 anos de acordo com os dados dos censos de 2010, 2011 e 2012



Fonte: Próprio autor (2019)

É possível notar um alto índice de evasão ocorrido nos 3 anos. Pode-se perceber também que a quantidade de alunos que concluem os cursos correspondem a menos da metade daqueles que ingressam. Essa discrepância entre concluintes e ingressantes gera preocupação para a instituição de ensino.

Dutra (2015) atribui um dos fatores da evasão escolar, o fato do aluno ingressar no curso com expectativas irreais a realidade ou até mesmo frustrações com a estrutura curricular limitada, com isso levando o discente a optar por mudança ou abandono do curso. De acordo com levantamentos de dados realizados por alguns estudos, uma fragilidade na escolha inicial e expectativas que não condizem com a carreira são uma das principais causas de evasão. (BARDAGI, 2007).

No entanto, com a crescente evolução dos sistemas de informação e o aumento da capacidade de armazenamento de dados, é possível desenvolver ferramentas que auxiliem na redução desses índices por meio dos dados coletados pelas instituições. A seguir serão mostrados métodos e sistemas que

em conjunto poderão ser aplicados com o objetivo de prever essas evasões.

2.2. ARMAZENAMENTO DE DADOS

Com o surgimento dos sistemas computacionais, houve uma necessidade pelo armazenamento de dados gerados por esses sistemas. No início do século 21 era algo muito custoso desenvolver um hardware que fosse capaz de guardar tanta informação de forma rápida, segura, eficiente e ocupando pouco espaço físico. Porém, com o crescimento tecnológico, os custos foram reduzidos. Camilo e Silva (2009) afirmam que no decorrer das décadas ocorreram quedas nos custos dos hardwares, tornam assim possível o avanço e desenvolvimento de novas formas de armazenamentos com maior capacidade.

Na subseção seguinte será falado sobre o banco de dados, um sistema de armazenamento de dados muito utilizado atualmente.

2.2.1. Banco de dados

Entre os sistemas de armazenamento de dados, o mais popular e usual é o banco de dados, Heuser (1998, p. 14) define banco de dados com sendo “o conjunto de arquivos integrados que atendem a um conjunto de sistemas”. Pode-se exemplificar um banco de dados como sendo uma espécie de lista telefônica, onde em seu interior existem contatos organizados de pessoas e empresas. Dentre os bancos de dados temos o banco de dados relacional (BDR) e o banco de dados orientado a objetos (BDOO). Boscarioli, Bezerra, Benedicto, Delmiro (2006, p.2) definem que:

> BDRs e BDOOs possuem características distintas mas basicamente servem ao mesmo propósito: persistir dados necessários para a manutenção do negócio para o qual são aplicados, possibilitando a recuperação, comparação e tratamento desses dados a fim de produzir resultados tangíveis.

Esses bancos têm um sistema de gerenciador de banco de dados (SGBD). Para Heuser (1998, p. 15) SGBD é um “software que incorpora as funções de definição, recuperação e alteração de dados em um banco de dados”. Esse sistema surgiu na década de 1970, porém os primeiros SGBD demandavam de especialistas treinados para por usá-los. Normalmente o projeto de um banco de

dados pode ocorrer em três etapas. Sendo a primeira etapa a modelagem conceitual, a segunda é o projeto lógico e a terceira etapa o projeto físico.

Quando utilizados corretamente, esses bancos de dados facilitam a obtenção e a análise das informações para a instituição. Com o objetivo de ajudar na busca dessas informações, foram desenvolvidas metodologias e ferramentas como o Data Warehouse e a mineração de dados.

A seguir será descrito sobre uma linguagem muito utilizada com a finalidade de manipular os dados desses bancos.

2.2.2. MySQL e Linguagem SQL

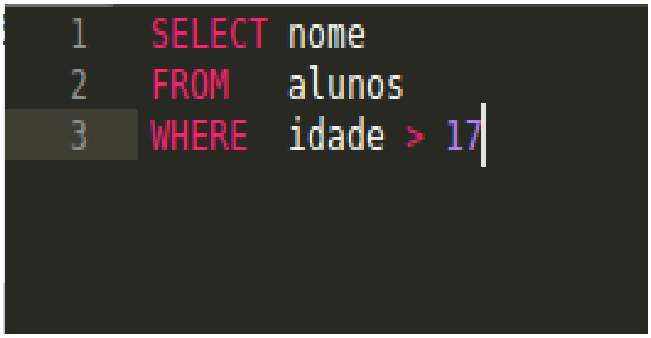
O MySQL é um SGBD de código aberto, bastante usado nas maiorias das aplicações gratuitas para gerenciar as bases de dados. Esse serviço utiliza a linguagem SQL (Structure Query Language), uma das linguagens mais populares para inserir, buscar, excluir e atualizar o conteúdo armazenado num banco de dados. No decorrer desta subseção será demonstrado como realizar uma simples consulta a partir dessa linguagem.

Niederauer (2008) reforça que o uso do MySQL é uma alternativa atrativa devido ao fato do seu baixo custo, mesma possuindo uma tecnologia complexa de banco de dados. Além dessas vantagens, também se pode destacar:

- Gratuidade do MySQL;
- Código fonte aberto;
- Fácil aprendizado e programação;
- Empregabilidade versátil, podendo ser utilizado em qualquer tipo de aplicação;
- Multi-plataforma;
- Permite que sejam implementados regras de segurança no servidor.

Como informado anteriormente, o MySQL utiliza a linguagem SQL. Ela surgiu no final dos anos 70 e foi desenvolvida pela International Business Machines (IBM). Pelo fato de ser uma linguagem universal e padronizada, ela se tornou vantajosa entre os profissionais de banco de dados. A instrução abaixo refere-se a uma simples consulta sobre o banco de dados da Figura 2:

Figura 2 - Consulta utilizando a linguagem SQL



Fonte: Próprio autor (2019)

A consulta acima busca os alunos que tenham a idade superior a 17 e retorna o nome desses alunos. Vale a pena diferenciar dados de informação, assunto que será abordado na próxima subseção.

2.2.3. Distinção entre dado e informação

Dependendo do autor e do ramo de estudo, o dado e a informação terá a mesma equivalência, no entanto, se tratando desse projeto em questão há uma distinção entre esses termos, todavia o dado e informação estarão estreitamente relacionados e formam as bases para se obter o conhecimento.

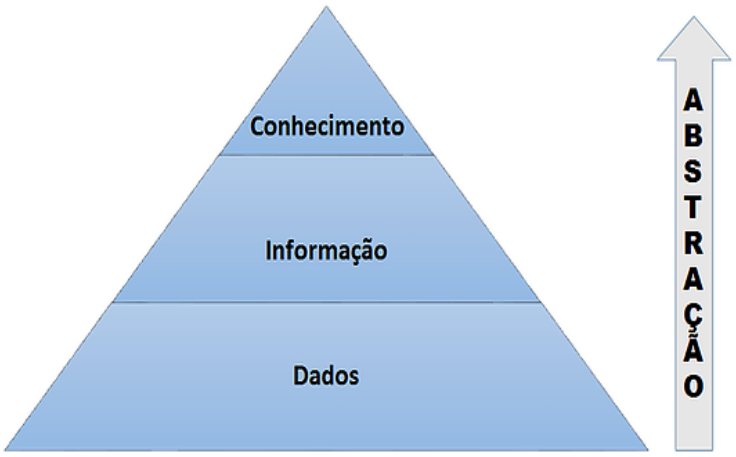
Setzer (2014, p.1) tem a seguinte definição para dado: “Defino dado como uma sequência de símbolos quantificados ou quantificáveis”. Exemplificado, os dados podem ser números, medidas ou valores e quando são transmitidos sozinhos não têm um significado e nem tão pouco levam a alguma compreensão.

A informação estará dentro de um determinado contexto, de modo organizado e ordenado, de tal modo que esse conjunto transmite um conhecimento (ELIAS, 2019). Para Setzer (2014) a informação representa algo que tenha algum significado para uma pessoa.

Como demonstra a Figura 3, pode-se perceber que na medida que distancia-se dos dados, o nível de abstração tende a aumentar, levando a uma melhor compreensão e a construção do conhecimento. Outra coisa perceptível na Figura

3 é o fato do conhecimento precisar da informação e está ser gerada com base nos dados obtidos.

Figura 3 - Pirâmide demonstrando o nível de abstração



Fonte: Elias (2019)

2.3. DATA WAREHOUSE

O Data Warehouse busca a extração e padronização de dados de uma forma automatizada, provendo assim para o usuário maneiras mais eficientes de visualizar os dados. Hokama, Camargo, Fujita e Fogliene (2004, p. 9) definem de uma forma geral o Data Warehouse como:

> Sistemas de Data Warehouse compreendem um conjunto de programas que extraem e tratam dados do ambiente operacional da empresa, um banco de dados que os mantém, e sistemas que fornecem estes dados aos seus usuários, dando suporte a consultas ad-hoc (consultas com acesso casual único e tratamento dos dados segundo parâmetros nunca antes utilizados), relatórios analíticos e à tomada de decisão.

O Data Warehouse armazena os dados de forma consolidadas, com o objetivo de apresentar as informações para os níveis gerenciais e estratégicos da empresa. Nele devem ser armazenados dados históricos da empresa, viabilizando as tomadas de decisões, descobertas de tendências e análises estratégicas.

Os elementos de estruturação do Data Warehouse é composto por Data Mart. Esses Data Marts são subconjuntos de dados do Data Warehouse que

trazem inúmeras vantagens na sua aplicabilidade, dentre elas, a possibilidade de retorno rápido (MACHADO, 2007). Em um mercado altamente competitivo, se faz necessário respostas rápidas e objetivas com baixo custo. Evitando assim perdas e prejuízos.

O Data Warehouse é imprescindível para as instituições, auxiliando nas estratégias e decisões. Fica evidente esse benefício quando Amaral (2003, p. 18) diz que “o usuário final do Data Warehouse (tomador de decisão) contará com um instrumento a mais para orientar suas análises, aumentando a confiabilidade do processo de tomada de decisão como um todo”.

Na próxima subseção será trazida uma ferramenta utilizada para analisar os dados e ajudar nas deliberações das mesmas.

2.3.1. Ferramentas OLAP

Uma ferramenta de Business Intelligence utilizada para apoiar as análises das informações das empresas e visando obter novos conhecimentos que serão empregados em tomadas de decisão, essa é uma definição para o OLAP (On-line Analytical Processing), além de ser voltado para acesso e análise ad-hoc de dados, tendo por finalidade o objetivo de transformar dados em informações e a partir disso ser capaz de dar suporte nas decisões gerenciais de forma amigável e flexível ao usuário e em tempo hábil. (ARAÚJO, BATISTA E MAGALHÃES, 2007).

Segundo Scheps (2008, p.68) “os bancos de dados OLAP são otimizados para armazenar dados de maneira a acelerar as tarefas analíticas”. Scheps atribui essa agilidade ao fato dos dados serem pré-agregados, ele acrescenta:

> Os dados são pré-agregados (somados e armazenados no banco de dados quando o banco de dados é processado em vez de quando os dados somados são solicitados), de modo que, quando você faz uma pesquisa detalhada, o computador não precisa resumir números de componentes; o computador simplesmente consulta o valor calculado anteriormente e o apresenta.

O OLAP pode ser aplicado em diferentes áreas e setores, tais como marketing, vendas, finanças, manufaturas e outros. Essencial para consultas constantes em banco de dados, além de tomar decisões baseadas nas informações contidas nele.(SISNEMA, 2019).

2.3.2. Ferramentas de extração de dados

As ferramentas de extração,transformação e carregamento de dados (ETL) são softwares muito utilizados em projetos de Data Warehouse e Business Intelligence. Ferreira, Miranda, Abelha e Machado (2010, p.2) define ETL como sendo “um processo para extrair dados de um sistema de Bases de Dados (BD), sendo esses dados processados, modificados, e posteriormente inseridos numa outra BD”.

Cruz, Miranda, Turchette (2014) acrescenta que toda parte de extração de informação de fontes externas é responsabilidade do ETL e essas fontes vão além de um banco de dados. Podendo ter dados extraídos de planilhas, documentos de texto, programas de ERP, programas de CRM e diversas outras fontes. A seguir será detalhado sobre como essa ferramenta auxilia na mineração de dados.

2.4. MINERAÇÃO DE DADOS

A Mineração de Dados ou Data Mining pertence a um ramo da computação que teve início nos anos 80. Ela surgiu devido a uma preocupação dos profissionais e empresas com o grande volume de dados que eram estocados e inutilizado na empresa. Basicamente nesta época a mineração de dados consistia em extrair informações de grandes bases de dados de forma mais automatizada. Atualmente, Data mining também analisa os dados após a extração (AMO, 2003).

A esse respeito Camilo e Silva (2009, p.2) declaram que:

> A Mineração de Dados é uma das tecnologias mais promissoras da atualidade. Um dos fatores deste sucesso é o fato de dezenas, e muitas vezes centenas de milhões de reais serem gastos pelas companhias na coleta dos dados e, no entanto, nenhuma informação útil é identificada.

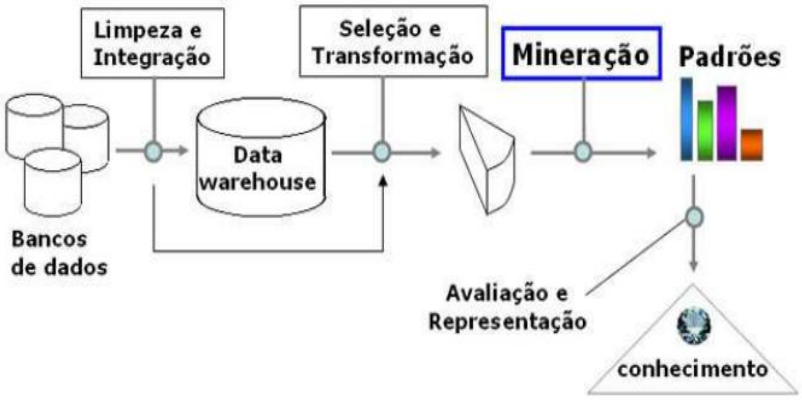
Nas subseções seguintes serão detalhado as 7 etapas para o processo de mineração de dados e a ferramenta ETL utilizada neste trabalho.

2.4.1. Etapas para a mineração de dados

Para Amo (2003) grande parte das pessoas considera o termo Mineração de Dados como sinônimo para Knowledge Discovery in Databases (KDD). Para ele, KDD é um processo mais amplo que apresenta 7 etapas:

1. Limpeza dos dados: etapa onde são eliminados ruídos e dados inconsistentes.
2. Integração dos dados: etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.
3. Seleção: etapa onde são selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir que informações como endereço e telefone não são relevantes para decidir se um cliente é um bom comprador ou não.
4. Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (por exemplo, através de operações de agregação).
5. Mineração: etapa essencial do processo consistindo na aplicação de técnicas inteligentes de se extrair os padrões de interesse.
6. Avaliação ou Pós-processamento: etapa onde são identificados os padrões interessantes de acordo com algum critério do usuário.
7. Visualização dos Resultados: etapa onde são utilizadas técnicas de representação de conhecimento a além de apresentar ao usuário o conhecimento minerado.

Figura 4 - As etapas do processo de KDD



Fonte: Amo (2003)

Na etapa do processo de KDD, inicia-se, com seleção das bases de dados que deseja trabalhar. Em seguida, é efetuada a limpeza e integração, pois geralmente os dados são encontrados com inúmeras inconsistências. Logo após, é realizado a seleção e transformação que consiste em reduzir ou projetar esses dados, nessa etapa utilizou-se o Pentaho Data Integration, na próxima subseção será detalhada sobre a ferramenta. Subsequente a essas três etapas, inicia-se a mineração de dados, onde serão escolhidas as técnicas e algoritmos que possibilitem a extração de padrões. Finalmente, efetua-se a etapa de avaliação e representação que compreende na interpretação dos padrões minerados. Após análises, usa-se o conhecimento, juntamente com sistemas de apoio a decisões, ou apenas documenta esse conhecimento.

2.4.2. Pentaho Data Integration

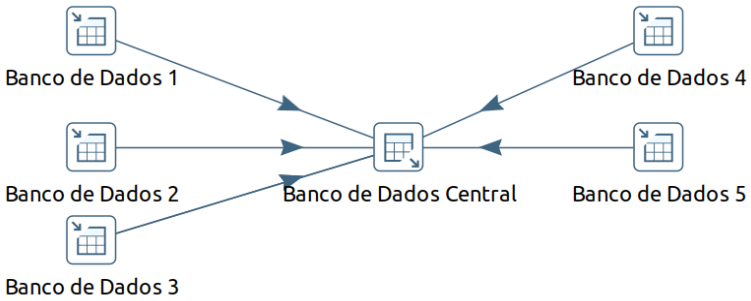
A ferramenta Pentaho, também conhecida como KETTLE é uma ferramenta ETL. Foi desenvolvida na linguagem JAVA, de modo que é possível utilizá-lo tanto em sistemas operacionais Windows quanto em Linux. A ferramenta Pentaho também oferece recursos avançados de extração, transformação e carregamento, em inglês, extraction, transformation, Loading (ETL). A plataforma conta com um ambiente de design intuitivo, gráfico, de arrastar e soltar. Dessa forma, basta o usuário selecionar o componente e arrastá-lo para onde deseja utilizá-lo.(HITACHI, 2019).

Cruz, Miranda, Turchette (2014, p.52) complementam também que:

> Além de trazer, organizar e apresentar os dados de acordo com as necessidades da empresa, o Pentaho é focado nos processos de trabalho da organização e traz soluções para a empresa. Ele implementa esses processos na resolução de problemas detectados, além de apresentar uma visão geral do funcionamento da empresa.

A escolha do Pentaho para o desenvolvimento desse projeto deve-se ao fato da ferramenta ser Open Source, implementável em qualquer tipo de empresa, fácil integração a outras infraestruturas de TI. Além da facilidade de unir diferentes bancos de dados simultâneos, podendo desse modo fazer o cruzamento desses dados e centralizá-los em um único local.

Figura 5 - Diversos bancos de dados sendo centralizados em uma única base.



Fonte: Próprio autor (2019)

Como é possível visualizar na figura 5 a ferramenta é utilizada para realizar as leituras das bases de dados 1 até 5, e posteriormente inserir no Banco de Dados Central, dessa forma enviando apenas os dados necessários e assim criando um Data Warehouse ou Data Mart.

2.5. INTELIGÊNCIA ARTIFICIAL

O uso da Inteligência Artificial (IA) está presente em diversas áreas, como por exemplo, nas áreas militar, industrial, de segurança e em algumas agências de inteligência. A IA busca difundir toda a capacidade de processamento de dados das máquinas, a capacidade dos seres humanos em aprender, comunicar e se interagir (FERNANDES, SILVA, BROCK, QUEIROGA, RODRIGUES, 2018).

De acordo com o Dicionário Michaelis a palavra inteligência originou-se do latim intelligentia, e significa a faculdade de entender, pensar, raciocinar e interpretar. A palavra artificial etimologicamente também teve origem no latim artificialis, produzido por arte ou indústria do homem e não por causas naturais, de acordo com o mesmo dicionário. Portanto, inteligência artificial é um produto produzido pelo homem com a intenção de criar e capacitar máquinas a pensarem de maneira similar ao ser humana. Ramos, Silva, Prata (2018, p. 4) conceituam IA da seguinte maneira:

> Inteligência artificial é a capacidade do computador em predizer o futuro e, com isso, tomar decisões entre possíveis opções. Ou seja, a Inteligência Artificial toma ações baseadas no que acredita que vai acontecer (de acordo com seu raciocínio inteligente) de melhor para

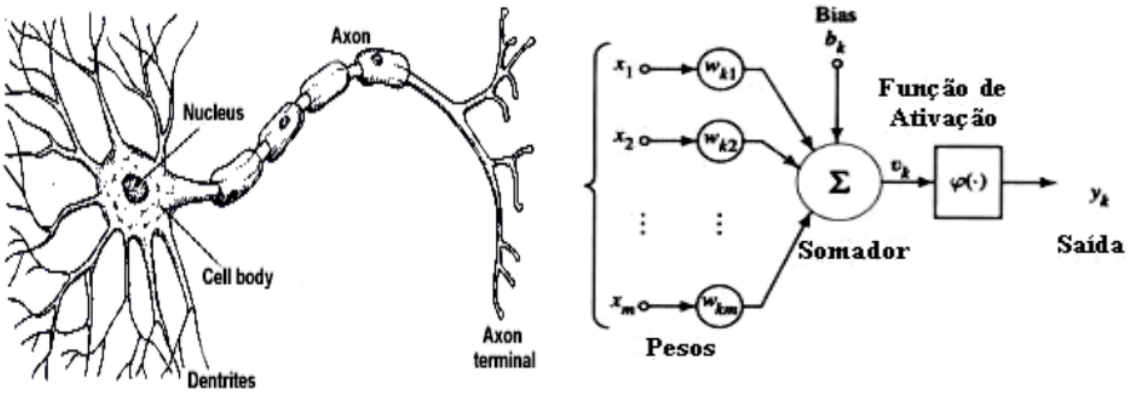
resolver uma tarefa ou um problema.

Geralmente o termo IA vem associado ao desenvolvimento de sistemas especialistas. Esses sistemas têm a capacidade de solucionar um problema de forma muito similar ao homem.

2.5.1. Rede Neural Artificial

A IA utiliza diversas ferramentas, entre elas pode se destacar a Rede Neural Artificial (RNA). Essa é uma ferramenta que possui a capacidade de se adaptar e a aprender a realizar algumas tarefas. Basicamente uma RNA é um circuito composto por grandes quantidades de unidades de processamento que foram inspiradas no sistema neural humano (FERNANDES, SILVA, BROCK, QUEIROGA, RODRIGUES, 2018).

Figura 6 - Representação artificial de um neurônio natural



Fonte: Moreira (2003)

De acordo com a Figura 6, pode-se ver a comparação entre um neurônio natural e a representação de um neurônio artificial. Na primeira imagem, o neurônio natural, os dentrites (dendritos) são responsáveis por receber as informações externas. No cell body (Corpo celular) essas informações são processadas e emitidas pelos axon (Axônios) (ANDRADE, SILVA, MOREIRA, SANTOS, DANTAS, ALMEIDA, LOBO, NASCIMENTO, 2003). Em relação a segunda imagem, o neurônio artificial, pode-se separá-lo em 3 partes. As entradas (x1, x2, ... xn), onde recebem os sinais de outros neurônios. No somador é onde

de fato ocorre o processamento das entradas e por fim a saída dessas informações.

Na RNA os sinais de comunicações entre esses neurônios são simulados por pesos e podendo ser ajustados, enquanto que num neurônio humano esses sinais são transmitidos por sinapses. Alvarenga, Correa, Osório (2011, p.6) afirmam que:

> O algoritmo de aprendizado irá otimizar e adaptar estes pesos que são denominados de “pesos sinápticos”, e através da adaptação do valor destes, é que fazemos com que a rede vá ajustando suas respostas de modo a responder da melhor maneira possível as entradas fornecidas (aproximando de modo iterativo as respostas obtidas na saída da rede dos valores da saída desejada indicada no arquivo de treinamento).

Alvarenga, Correa, Osório (2011, p.6) resumem que o objetivo do processo de aprendizado é melhorar as respostas da rede neural, até alcançar um determinado critério de parada. Geralmente esse critério de parada será quando a rede acerta grande parte ou todas as respostas da base de aprendizado.

2.5.2. Aprendizado de máquina

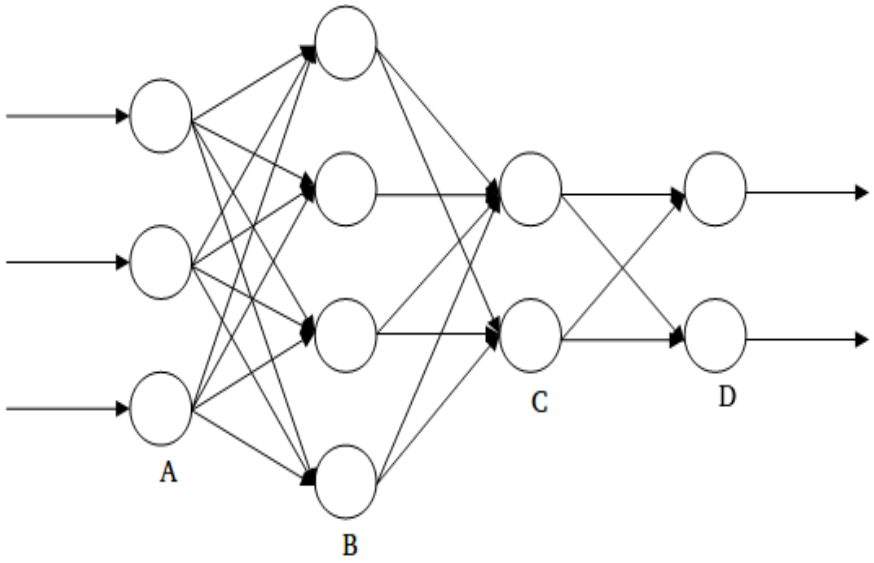
O aprendizado de máquina ou machine learning é um ramo da IA que consiste na ideia de que sistemas possam identificar padrões e tomar decisões com pouca ou nenhuma intervenção humana, utilizando apenas análises de dados.

O machine learning pode ser subdividido em várias categorias diferentes, entre elas temos a aprendizagem supervisionada que necessita de uma pessoa para oferecer exemplos de quais entradas ela precisa a fim de obter os resultados e validá-los e a aprendizagem não-supervisionada onde o sistema desenvolve suas próprias conclusões a partir de um determinado conjunto de dados.

2.5.3. Multilayer Perceptron

A Multilayer Perceptron (MLP) é um rede neural unidirecional e distribuída em camadas construída de um conjunto de nós, onde cada nó forma a camada de entrada da rede, contendo uma ou mais camadas ocultas de nós e uma camada de saída. Todas as camadas realizam processamento, exceto a camada de entrada (SOUZA, 2012).

Figura 7 - Arquitetura de uma MLP



Fonte: Salatas (2011)

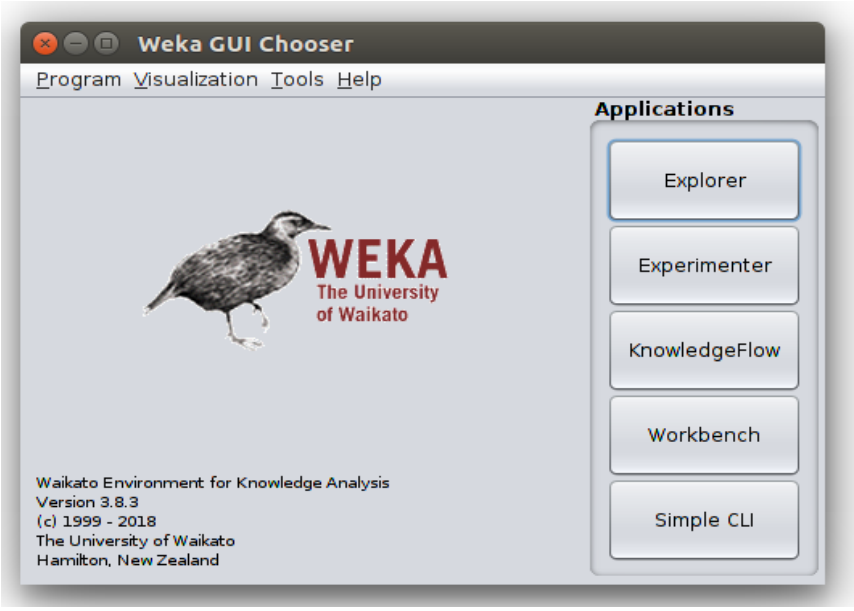
Na Figura 7, temos a camada de entrada (A), as camadas ocultas (B), (C) e a camada de saída (D).

A MLP tem o processo de aprendizado que consiste na apresentação do conjunto de dados de treinamento. No decorrer do treinamento, poderão existir erros de classificação. Esses erros são corrigidos na medida em que os pesos são ajustados, de modo a minimizá-los ou corrigi-los por definitivo nas iterações seguintes (SOUZA, 2013).

2.5.4. Weka

A ferramenta Weka contém uma coleção de algoritmos de aprendizado de máquinas para mineração de dados, além de integrar ferramentas para preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização. Pelo fato de ter sido desenvolvido usando uma abordagem de framework, o WEKA se torna extensível, permitindo que novos algoritmos ou funcionalidades sejam adicionadas (UNIVERSITY OF WAIKATO, 2019).

Figura 8 - Interface inicial do Weka com os 5 botões para iniciar os aplicativos



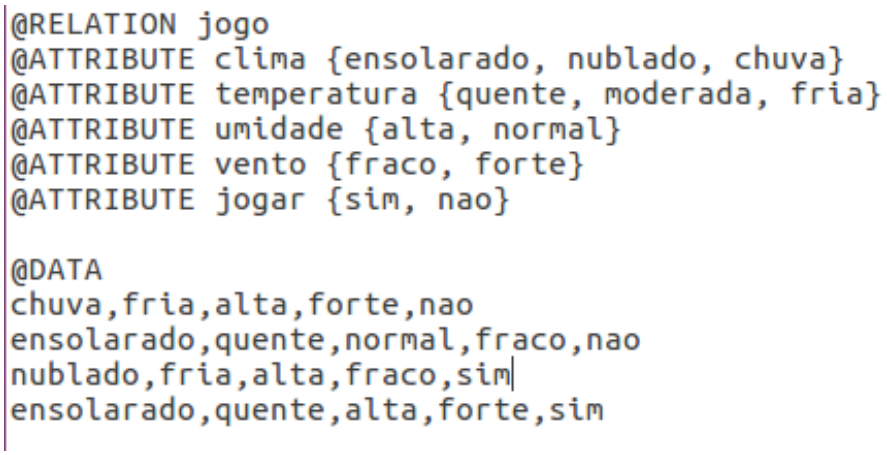
Fonte: Próprio autor (2019)

A ferramenta também permite escolher entre 5 aplicações diferentes, sendo elas Explorer, Experimenter, KnowledgeFlow, Workbench e Simple CLI.

- Explorer: Nesta aplicação é possível explorar dados com o WEKA. O ambiente disponibiliza ao usuário metodologias de mineração de dados, como classificação, agrupamento, associação e seleção de atributos.
- Experimentador: Um ambiente para a realização de experimentos e a realização de testes estatísticos entre esquemas de aprendizado.
- KnowledgeFlow: Este ambiente suporta essencialmente as mesmas funções que o Explorer, mas com uma interface de arrastar e soltar. Uma vantagem é que ele suporta aprendizado incremental.
- Workbench: Um aplicativo multifuncional que combina todos os outros dentro “perspectivas” selecionáveis pelo usuário.
- SimpleCLI: Fornece uma interface simples de linha de comando que permite execução de comandos WEKA para sistemas operacionais que não fornecem sua própria interface de linha de comando (UNIVERSITY OF WAIKATO, 2019).

A ferramenta WEKA utiliza um formato próprio de arquivo (.arff). Esse arquivo se divide em duas partes principais, uma delas é o cabeçalho, onde possui os atributos que serão utilizados, e a outra contendo os dados. A seguir temos uma imagem de um desses arquivos:

Figura 9 - Imagem de um arquivo .arff



Fonte: Próprio autor (2019)

Na figura 9 é possível ver 5 atributos que constituem cada instância no Dataset “jogo”. Por exemplo, o atributo “clima” é um dentre vários valores possíveis. Por padrão, o último atributo definido é considerado o atributo class e a partir dele que será gerado o modelo, bem como o que o modelo gerado irá calcular, neste caso, se será possível jogar ou não.

A próxima seção descreve a metodologia utilizada neste estudo para alcançar os objetivos citados neste referencial.

3. METODOLOGIA

Com a intenção de detectar futuras evasões na Rede de Ensino Doctum por meio do uso da IA, foram utilizadas técnicas de mineração de dados, uso do algoritmo de redes neurais artificiais MLP e ferramentas que auxiliaram nesse estudo, o Pentaho Data Integration e o Weka. Eles serão detalhados nas próximas subseções.

Também para o desenvolvimento deste projeto, foi de suma importância conhecer conceitos fundamentais teóricos de inteligência artificial e redes neurais artificiais que foram abordados nas seções anteriores, além de seguir determinadas etapas antes do resultado final. Pode-se destacar 6 etapas que foram seguidas para a obtenção do resultado:

1. Pré-processamento;
2. Criação do Data Warehouse;
3. Mineração dos dados;
4. Entrada de dados;
5. Treinamento da rede neural;
6. Análise dos dados.

Nas subseções seguintes serão detalhadas as etapas citadas.

3.1. PRÉ-PROCESSAMENTO

Essa é uma etapa muito importante, pois é onde ocorre a escolha, preparação, organização e estruturação dos dados para armazenamento no Data Warehouse. Pelo fato dos dados usados serem da Rede de Ensino Doctum, foi necessário solicitar uma autorização junto ao coordenador de tecnologia da rede de ensino (Anexo I) para que pudesse obter acesso aos dados dos alunos que estão armazenados nos bancos de dados da instituição.

Devido a grande massa de dados que seria processado, optou-se por utilizar o Pentaho Data Integration para migrar os dados para o Data Warehouse, esse processo será descrito em seções posteriores.

Dos muitos dados disponíveis dos discentes da instituição escolheu-se por dar ênfase nas respostas do questionário socioeconômico. Além do questionário, utilizou-se o nome do curso o valor da mensalidade do curso para integrar o treinamento da rede neural. Para esse projeto, Informações pessoais como nome do aluno, e-mail, CPF, RG, data de nascimento, nome dos responsáveis pelo aluno ou endereço foram irrelevantes.

A seguir será detalhado sobre o questionário socioeconômico, sua funcionalidade para a instituição de ensino e as perguntas utilizadas.

3.1.1. Questionário socioeconômico

A Rede de Ensino Doctum oferece bolsas de estudos de até 100% (cem por cento) para alunos que responderam o questionário socioeconômico e cumpriram com os requisitos do edital de Filantropia da Instituição. Essas bolsas são ofertadas pelo fato da condição de entidade filantrópica da Instituição (DOCTUM, 2019).

Para a análise das respostas do questionário pela rede neural, optou-se pelas perguntas de respostas fechadas, ou seja, onde é listado as opções de escolha para o discente. Dessa maneira será possível obter respostas padronizadas e ter um maior controle nelas. No anexo II temos o formulário usado com as perguntas disponíveis do questionário socioeconômico para os alunos e suas possíveis respostas.

De acordo com as respostas escolhidas pelo aluno, o algoritmo de rede neural MLP deverá prever se o aluno continuou os estudos após responder o questionário. Após o pré-processamento, foi criado um Data Warehouse para receber os dados filtrados.

A seguir serão descritas as tabelas criadas no data warehouse para cada pergunta do questionário.

3.2. CRIAÇÃO DO DATA WAREHOUSE

Na etapa anterior os dados foram captados, tratados, transformados e após isso foram armazenadas em um novo banco, dando origem ao Data Warehouse que será utilizado para a mineração de dados. Como escolheu-se trabalhar

apenas com o questionário socioeconômico da instituição, não foi preciso criar um data mart.

No novo banco de dados foi organizado cada pergunta como sendo uma tabela, de modo a facilitar as análises e caso alguma pergunta não tivesse peso considerável para o projeto poderia ser descartada facilmente. Também foi criado uma tabela dataset_resposta contendo o conjunto de todas as respostas de forma que cada linha desse dataset_resposta representa-se um aluno. A seguir as 13 tabelas do novo banco de dados e suas respectivas descrições:

- dados_curso: Armazena os dados do curso do aluno no momento em que respondeu o questionário socioeconômico, nessa tabela foram salvas informações importantes como o nome do curso, valor da mensalidade, total de períodos do curso e período atual do aluno. O objetivo dessas informações é saber se o curso, mensalidades e períodos longos influenciam na evasão do aluno.
- pergunta_aluno_familia_possui_problemas_saude: Nesta tabela armazenou-se a resposta do aluno para a pergunta “Existem problemas de saúde entre as pessoas que moram com sua família, inclusive você:” , onde era possível obter duas respostas, “não” ou “sim”. Esta pergunta ajuda a definir se a saúde da família ou até mesmo do próprio aluno pode ser motivo de evasão.
- pergunta_aluno_mora_longe_faculdade: Nesta tabela armazenou-se também a resposta “não” ou “sim” para a pergunta “Mora fora da cidade onde está localizada a Faculdade e/ou Colégio?”. Essa pergunta foi útil para definir se alunos que moram longe da instituição de ensino são mais propensos a evadir pelo fato de terem que ter algum meio de transporte e gastar mais tempo para chegar até a faculdade .
- pergunta_aluno_mora_sozinho: Nesta tabela foi armazenado as respostas referentes à pergunta “Você mora:”. Pelo fato de existir 6 respostas diferentes para essa pergunta, optou-se por dividir as respostas em dois grupos, “não” ou “sim”. Para as respostas “com cônjuge / companheiro (a)”, “com os pais (ou somente com um dos pais)”, “em casa de familiares / casa de amigos”, “em república / quarto / pensão / pensionato” e “outros”, associou-se ao grupo de resposta “não” e para a resposta “sozinho”

associou-se ao grupo “sim”. O objetivo da escolha desta pergunta foi detectar se uma pessoa morando sozinha tem uma maior chance de evasão, visto que as responsabilidades da casa ficam por conta dela.

- pergunta_aluno_responsavel_financeiro_familia: Nesta tabela armazenou-se as respostas referentes a pergunta “Quem é (são) o (os) responsável (is) pela manutenção financeira do grupo familiar:”. Como existem 5 opções de respostas para essa pergunta, também optou-se por criar dois grupos de resposta, “não” e “sim”. O grupo do “não” faz referência às respostas “outros”, “outros membros do grupo familiar”, “pai / mãe” e “somente um dos pais”. E o grupo do “sim” a resposta “próprio estudante”. Aqui é possível detectar se as responsabilidades financeiras da família podem influenciar na motivação do aluno em se manter na instituição.
- pergunta_aluno_responsavel_pela_mensalidade: Nesta tabela tem-se as repostas para a pergunta “De que forma são mantidas as despesas financeiras com a instituição?”. As respostas foi divido em dois grupos, “não” ou “sim”. As respostas “outros”, “recebo ajuda de parentes”, “sustentado pelos meus pais (ou por somente um dos pais)” e “tenho bolsa de estudo” ficou associado ao grupo “não” e “sim” para a resposta “sou responsável pelo meu próprio sustento”. Um pouco semelhante à pergunta anterior, aqui o foco está voltado para as mensalidades do curso do aluno, acredita-se que alunos com bolsas de estudos tem uma baixa evasão.
- pergunta_aluno_sempre_cursou_ensino_publico: Nesta tabela foi armazenado as respostas referentes a pergunta “A Instituição de ensino na qual cursou as séries ou anos anteriores é:”. Foi criado dois grupos, “não” e “sim”, onde que para as respostas “parte em particular depois em pública”, “parte em pública e depois em particular”, “particular” e “particular com bolsa integral” ficou no grupo do “não” e a resposta “pública” no grupo “sim”. Espera-se que alunos de ensinos privados tenham uma taxa de evasão menor pois geralmente tem uma melhor estrutura educacional em comparação com as redes públicas se tratando de ensino médio.
- pergunta_aluno_vai_a_pe_faculdade: Nesta tabela armazenou-se as respostas para a pergunta “Qual o principal meio de transporte utilizado para chegar à Faculdade e/ou Colégio?”. Também foi necessário a criação de dois grupos, “não” e “sim”, para as respostas. Para o grupo “não” temos as

respostas “de carona”, “oferecido gratuitamente por prefeitura e/ou escola”, “outros”, “transporte coletivo pago diariamente com recursos próprios”, “transporte locado, gasto mensal” e “transporte próprio”, e no grupo sim, “a pé / de bicicleta”. Esta pergunta é muito importante, por ela pode-se perceber se a locomoção poderá influenciar na evasão, visto que a utilização de um transporte se faz necessário caso o aluno more longe da faculdade.

- pergunta_familia_mora_aluguel: Aqui temos armazenadas as respostas para a pergunta “Sua família reside em imóvel:”. Para essa pergunta, era disponível 6 respostas, sendo preciso criar dois grupos de resposta, “não” e “sim”. O grupo do “não” refere-se às respostas “emprestado ou cedido”, “outros”, “próprio, em pagamento / financiamento”, “próprio, já quitado” e “próprio, por herança”. No grupo do “sim” ficou a resposta “alugado”. Com esta pergunta será possível detectar os gastos financeiros do aluno, uma vez que ele ou a família precisará gastar com moradia, influenciando nos gastos com as mensalidades caso o aluno não tenha bolsa de estudo. Podendo ser um fator que leve a evasão.
- pergunta_familia_mora_zona_rural: Nesta tabela armazenou-se repostas referentes a pergunta “Sua família reside em:”. Como é possível ter mais de duas opções de respostas, optou-se por criar dois grupos de respostas, “não” e “sim”. No grupo “não” temos as respostas “bairro padrão médio”, “bairro padrão popular”, “outros” e “vila ou aglomerado”, já no grupo “sim”, “fora do perímetro urbano”. Como em perguntas anteriores, o objetivo era verificar se a distância poderia ser influência na evasão, pois um aluno que more fora do perímetro urbano precisará de mais tempo para chegar até a faculdade. Podendo assim levar ao desânimo de continuar os estudos.
- pergunta_familia_possui_automoveis: As respostas para a pergunta “A família possui Automóveis? Qual a faixa de valor (da soma dos automóveis do aluno e do grupo familiar):” ficaram armazenadas nesta tabela. Também criou-se dois grupos de respostas, “não” e “sim”. A resposta “sim” faz referência as perguntas “acima de R$ 30.000,00”, “até R$ 10.000,00”, “R$ 10.000,01 a R$ 20.000,00” e “R$ 20.000,01 a R$ 30.000,00”. Para o grupo “não”, ficou a resposta “não possui automóvel”. Uma pergunta indispensável, pois pode-se concluir que famílias que possuem um automóvel tem condições financeiras para mantê-lo, podendo concluir que a família do aluno

tem uma renda financeira capaz de manter seus estudos.

- pergunta_familia_possui_outros_imoveis: Por fim temos a tabela para a última pergunta que foi escolhida do questionário socioeconômico, “A família possui outros imóveis além do que habita? (Lotes, sítios, fazendas, casas na praia, aptos, barracões ou outros)”, pelo fato de ter apenas duas opções de escolha, “não” ou “sim”, não foi preciso fazer tratamentos para receber essa resposta. Como na pergunta referente ao automóvel, caso a família do aluno tenha mais de um imóvel, significa uma renda financeira capaz de mantê-lo, não precisando evadir por questões financeira.
- dataset_respostas: Foi criado uma tabela para receber todas as respostas anteriores já tratadas.

A extração, transformação e carregamento desses dados deu-se por meio da ferramenta Pentaho Data Integration, a seguir serão abordados mais detalhes a respeito da ferramenta e os recursos utilizados.

3.2.1. Pentaho Data Integration

Antes de iniciar a configuração da conexão da ferramenta com o banco de dados foi preciso baixar a biblioteca Java MySql JDBC, essa biblioteca permitiu criar uma conexão com o MySql. Após baixar e configurar a biblioteca, iniciou-se a etapa de pré-processamento.

Foram utilizados 3 componentes do PDI, sendo eles as conexões, Tables Input e Tables Output. A seguir serão falado da etapa de conexão.

3.2.2. Unidades da Rede de Ensino Doctum

Para este projeto optou-se apenas pela extração das unidades de ensino da Rede Doctum que tenham curso de graduação. Nessa etapa foi necessário criar 15 conexões com os bancos de dados de onde será feita a extração. O motivo desse número enorme de conexões deu-se pelo fato da base de dados da instituição ser separado por filiais. A seguir tem-se as cidades que tiveram suas bases de dados extraídas:

1. Cataguases, MG;

2. Carangola, MG;
3. Caratinga, MG;
4. Guarapari, ES;
5. Ipatinga, MG;
6. Iúna, ES;
7. Juiz de Fora, MG;
8. Juiz de Fora - Zona Norte, MG;
9. João Monlevade, MG;
10. Leopoldina, MG;
11. Manhuaçu, MG;
12. Serra, MG;
13. Teófilo Otoni, MG;
14. Vila Velha, ES;
15. Vitória, ES.

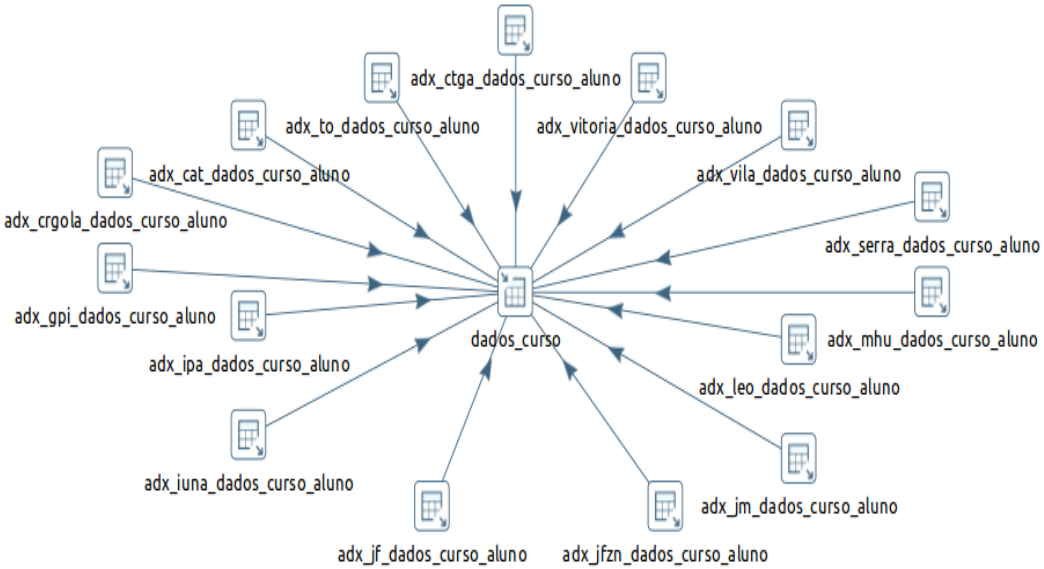
Dentre essas unidades, um destaque para Juiz de Fora e Juiz de Fora - Zona Norte. Essas unidades estão na mesma cidade, porém em prédios diferentes, por esse motivo as suas bases de dados estão separadas.

3.2.3. Entrada de tabelas

Para cada uma das 15 filiais utilizou-se uma entrada de tabela diferente, com suas respectivas consultas ao banco de dados da instituição. O retorno dessas consultas era centralizada em uma única saída de tabela responsável pela inserção no Data Warehouse.

No caso desse projeto era necessário conectar-se em diferentes bases de dados, uma das opções era utilizando a entrada de tabela devido ao fato dele poder ler as informações de um banco de dados, usar instruções SQL, ser replicável e após isso, enviar para um ou vários outros banco de dados seu retorno. A seguir temos ligações de diversas entrada de tabela.

Figura 10 - Diversas entradas de tabelas sendo centralizada em uma saída de tabela



Fonte: Próprio autor (2019)

Pode-se perceber as 15 entradas de tabelas, onde cada um representando uma filial da Rede de Ensino Doctum. O resultado dessas entradas de tabelas está sendo centralizado em uma saída de tabela. A seguir será falado sobre a funcionalidade de saída de tabela.

3.2.4. Saída de tabelas

A saída de tabela diferencia-se da entrada de tabela no momento da conexão com o banco de dados. Enquanto a entrada de tabela faz uma conexão com o banco de dados de onde os dados serão extraídos, na saída de tabela é necessário estabelecer uma conexão com o banco de dados onde os dados serão armazenados. Após a conexão com os bancos de dados, basta realizar a conexão entre a entrada de tabela e a saída de tabela.

Após a extração dos dados para o Data Warehouse, deu-se início a etapa de mineração de dados.

3.3. MINERAÇÃO DE DADOS

Na etapa de mineração de dados é onde ocorre de fato o uso de

inteligência artificial e redes neurais artificiais para a predição de evasão escolar dos alunos da instituição de ensino. Foi utilizado um algoritmo de redes neurais chamado Multilayer Perceptron.

A escolha por esse algoritmo deu-se pelo fato do usuário poder usar uma grande quantidade de entradas de dados e o Multilayer Perceptron poder tratá-los mais facilmente do que outros. Juntamente com o Multilayer Perceptron, a ferramenta Open Source Weka também foi utilizada. Essa ferramenta disponibiliza ao usuário variados métodos de análise de dados, além de diversos algoritmos.

Foi utilizada a metodologia de classificação, que consiste em classificar objetos a determinadas classes, buscando prever uma classe de um novo dado automaticamente.

3.4. ENTRADA DE DADOS

A ferramenta Weka utiliza como padrão apenas arquivos de texto na extensão .arff como fonte dos dados para mineração ou uma conexão direta com o Data Warehouse. Para o treinamento da rede optou-se pela conexão com o banco de dados.

Para o treinamento foram escolhidos os questionários respondidos entre 2017 e 2018 de alunos regulares na época, onde as respostas foram verificadas pela secretaria. Escolheu-se duas bases de dados apenas, uma de Caratinga e outra de Teófilo Otoni. Dentro desse período de tempo, 2593 alunos responderam o questionário. Utilizou-se 80% dos dados para treinar a rede neural e 20% para testar a rede. A base de dados, continha quais alunos evadiram e quais não evadiram.

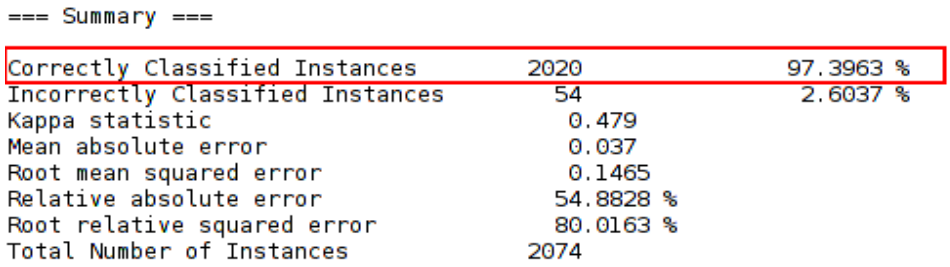
Quadro 1 - Quantidade de alunos não evadidos e evadidos

Evadidos	108
Não Evadidos	2485
TOTAL DE ANÁLISES	2593

Fonte: Próprio autor (2019)

Pode-se perceber que juntando as duas filiais, o índice de evasão é relativamente baixo. A seguir o resultado do treinamento da rede neural:

Figura 11 - Percentual de acerto da rede neural



Fonte: Próprio autor (2019)

A rede neural mostrou-se bastante eficiente na classificação, das 2074 respostas, ela conseguiu um acerto de mais de 97%. Porém foi preciso fazer um teste definitivo para comprovar seu sucesso e aplicabilidade. Para isso pegou-se os outros 20% das respostas dos alunos que ainda não foram analisados pela rede neural. A seguir temos a análise detalhada.

3.5. TREINAMENTO E ANÁLISE DOS DADOS

Para realizar as análises dos dados da rede neural e ter confiabilidade nos resultados utilizou-se o segundo arquivo contendo 519 respostas, de modo que foi possível testar a eficiência da rede neural. Dessa vez foi criado um arquivo .arff e ocultado o atributo que informa se o aluno evadiu-se ou não.

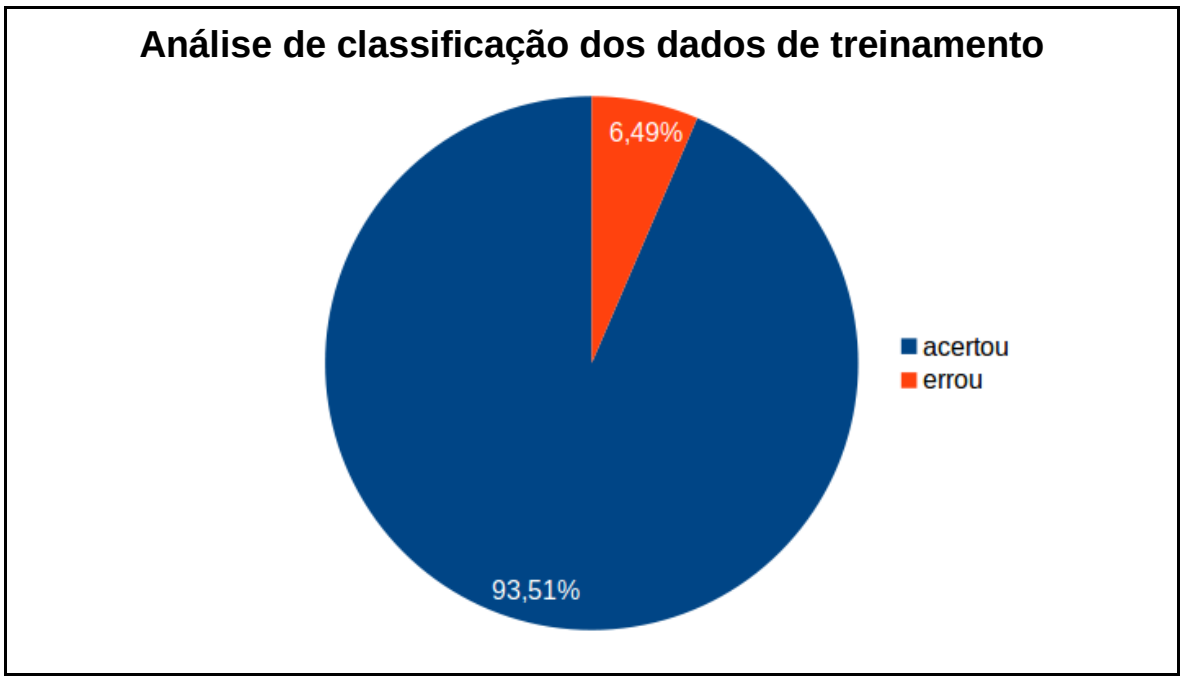
Quadro 2 - Quantidade de acertos e erros da rede neural

Total de acertos	475
Total de erros	44
TOTAL DE ANÁLISES	519

Fonte: Próprio autor (2019)

No quadro 2 pode-se notar que das 519 análises da rede neural, ela conseguiu prever 475 e errou apenas 44.

Gráfico 1 - Percentual da classificação dos dados de treinamento



Fonte: Próprio autor (2019)

Isso significa uma assertividade de 91,40%, um bom resultado, mas ainda não é satisfatório se levar em conta que foi utilizado apenas duas filias da Rede de Ensino Doctum. Para certificar-se do efetivo sucesso da rede ou não, foi utilizada as respostas das 15 filiais da instituição dos questionários respondidos entre 2017 e 2018, o resultado será descrito na seção seguinte.

4. RESULTADOS

Para se obter o resultado final utilizou-se 15 filiais da Rede de Ensino Doctum, somando um total de 8479 respostas de discentes. A análise de evasão levou em consideração os alunos de graduação e que responderam o questionário socioeconômico entre o ano de 2017 e 2018, os demais foram ignorados. Também foi realizado mais um treinamento com a rede neural, com 80% das respostas para treinar e 20% para confirmação, o objetivo desse treinamento era verificar se o algoritmo teria uma maior assertividade na previsão pelo fato de ter mais dados para treinamento. Das 6783 respostas para treinamento, a rede neural conseguiu prever mais de 96% das respostas.

Figura 12 - Percentual da classificação da rede neural

```
=== Summary ===

Correctly Classified Instances        6518               96.0932 %
Incorrectly Classified Instances       265                3.9068 %
Kappa statistic                          0.2956
Mean absolute error                      0.0517
Root mean squared error                  0.1841
Relative absolute error                 59.504  %
Root relative squared error             88.4071 %
Total Number of Instances             6783
```

Fonte: Próprio autor (2019)

Pode-se concluir que nesse caso o fato de ter mais dados para treinamento não foi relevante para se obter uma maior assertividade da rede neural no momento da classificação, visto que na Figura 12 as classificações corretas foram menor que na Figura 11.

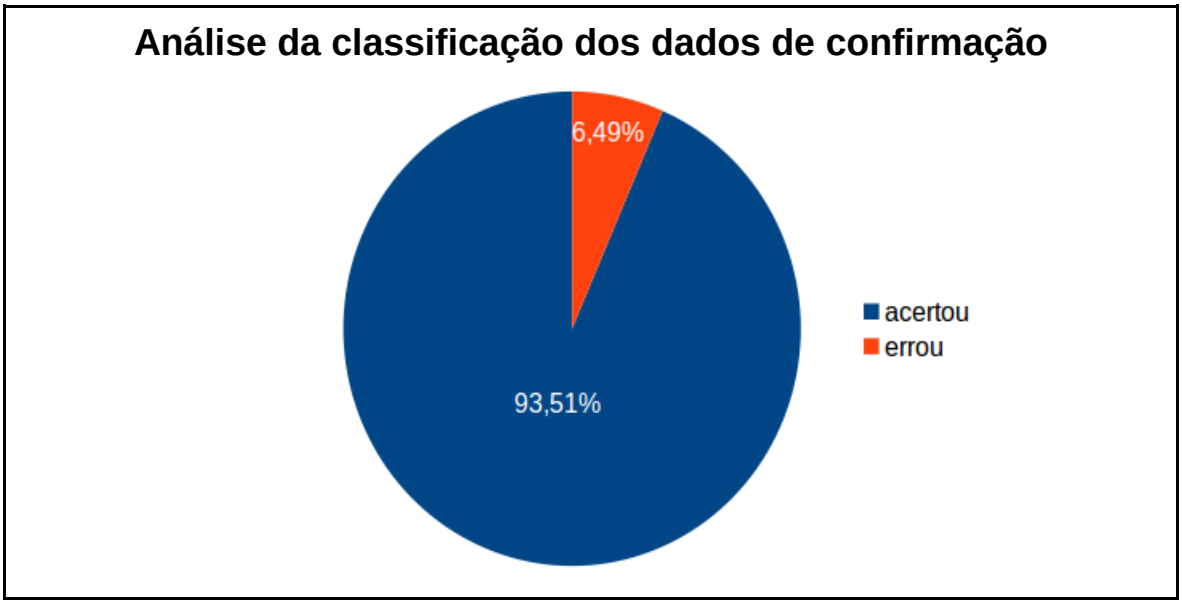
Quadro 3 - Resultado da classificação da rede neural

Total de acertos	1586
Total de erros	110
TOTAL DE ANÁLISES	1696

Fonte: Próprio autor (2019)

Após o treinamento usou-se os 20%, que equivalem a 1696 das respostas restantes. No Quadro 3 tem-se a quantidade de acertos e erros da rede neural. Das 1696 respostas armazenadas no data warehouse, 1586 foram classificadas de forma correta pela rede e apenas 110 classificações incorretas. A seguir um gráfico representando esses números em percentual.

Gráfico 2 - Percentual da classificação dos dados de confirmação



Fonte: Próprio autor (2019)

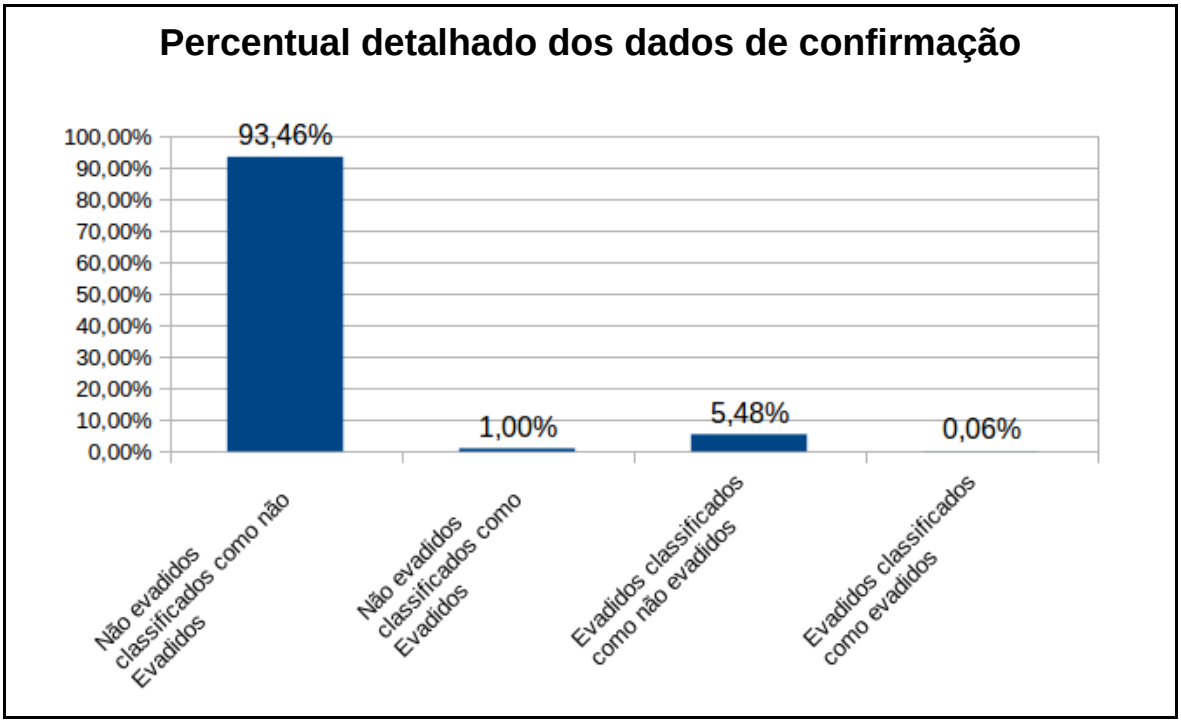
Aparentemente o resultado da rede neural é satisfatório, visto que ele acertou mais de 93% dos casos, porém será preciso fazer outras análises.

Quadro 4 - Resultado detalhado dos dados de confirmação

Não evadidos classificados como não Evadidos	1585
Não evadidos classificados como Evadidos	17
Evadidos classificados como não evadidos	93
Evadidos classificados como evadidos	1
TOTAL DE ANÁLISES	1696

Fonte: Próprio autor (2019)

Gráfico 3 - Percentual detalhado da classificação da ferramenta



Fonte: Próprio autor (2019)

Como visto no Quadro 4 e Gráfico 3, a rede neural obteve um excelente nível de acerto para os alunos que não evadiram, isso aconteceu pelo fato de que das 6783 respostas para treinamento, em 6475 delas os alunos não evadiram. De modo que a rede conseguiu mais dados para reconhecer padrões de alunos não

evadidos. A seguir um quadro contendo a quantidade real de alunos não evadidos e evadidos usados no treinamento da rede neural:

Quadro 5 - Quantidade real de alunos não evadidos e evadidos

Total não evadidos	6475
Total evadidos	308
TOTAL DE ANÁLISES	6783

Fonte: Próprio autor (2019)

Pode-se analisar os resultados de dois modos diferentes. O primeiro como mencionado anteriormente, se considerar os alunos que foram classificados com não evadidos obteve-se um resultado satisfatório, visto que a rede acertou 93,45% dos casos. O segundo modo, e este é o principal objetivo deste trabalho, é levar em consideração os alunos que evadiram, a rede neural mostrou um baixo desempenho, classificando corretamente apenas 0,06%, e errando em 5,48% os alunos que são evadidos e foram classificados como não evadidos.

5. CONCLUSÃO

Devido ao alto índice de evasão escolar e dos prejuízos causados por ela, como mencionado em seções anteriores, este trabalho focou-se na criação de uma rede neural artificial capaz de prever discentes que pudessem evadir das instituições de nível superior. Utilizou-se como base os dados socioeconômicos dos alunos da Rede de Ensino Doctum na tentativa de encontrar padrões de evasão dos mesmos, de modo a criar metodologias capazes de reduzir ou até mesmo extinguir tais índices na instituição.

Os resultados obtidos não foram satisfatórios se levado em consideração que a rede neural não conseguiu prever um número elevado de futuras evasões. Pode-se destacar dois fatores que contribuíram com este resultado.

O primeiro fator a se observar foi o baixo volume de dados obtidos, pois foi ignorado alunos de ensino fundamental, ensino médio e pós graduandos da rede, focando apenas em alunos de cursos superiores de graduação e que tenham respondido em algum momento de sua vida acadêmica o questionário socioeconômico por completo, um questionário que não é obrigatório.

O segundo fator a destacar-se foi o fato de que quase todos os alunos que responderam o questionário não evadiram da instituição, de modo que a rede neural não conseguiu treinar muitos casos de evasão.

Porém com a elaboração deste trabalho foi possível constatar que a inteligência artificial aliada a ferramentas de análise de dados podem ser úteis para tomadas de decisões mais eficientes em instituições de ensino.

6. TRABALHOS FUTUROS

Uma continuação possível deste trabalho, um estudo de caso comparativo entre os classificadores Multilayer Perceptron e o J48 na previsão de evasão de discentes com base no questionário socioeconômico.

Outra possível continuação deste trabalho, um estudo de caso de outras metodologias de mineração de dados.

Mesmo com um baixo índice de acerto, o questionário socioeconômico juntamente com outros dados dos alunos é uma ótima fonte de informação para análise de evasão escolar, devido ao teor das respostas pessoais de cada aluno.

7. REFERÊNCIAS

ALVARENGA, Matheus Lin Truglio. CORREA, Diogo S. Ortiz. OSÓRIO, Fernando Santos. **Redes Neurais Artificiais aplicadas no Reconhecimento de Gestos usando o Kinect.** 2011. USP – Univ. de São Paulo - ICMC – Instit. de Ciências Matemáticas e de Computação. SSC – Depto. de Sistemas de Computação. Disponível em: <https://s3.amazonaws.com/academia.edu.documents/33594111/IHCGestos.pdf?response-contentdisposition=inline%3B%20filename%3DRedes_Neurais_Artificiais_aplicadas_no_R.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20191027%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20191027T154438Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=5fbbefc46d20ba4c29bf3173b1c302a44f14673a55ae7b0c2725aa562fda695b>. Acesso em: 27 out 2019.

AMARAL, Glenda Carla Moura. **AQUAWARE: Um Ambiente de Suporte à Qualidade de Dados em Data Warehouse.** 2003. Universidade Federal do Rio de Janeiro. Instituto de Matemática. Núcleo de Computação Eletrônica. Disponível em:<https://www.researchgate.net/profile/Maria_Campos11/publication/266583807_AQUAWARE_Um_Ambiente_de_Suporte_a_Qualidade_de_Dados_em_Data_Warehouse/links/5567180a08aec2268300f45e/AQUAWARE-Um-Ambiente-de-Suporte-a-Qualidade-de-Dados-em-Data-Warehouse.pdf>. Acesso em: 17 abr 2019.

AMO, Sandra de. **Técnicas de Mineração de Dados.** 2003. Universidade Federal de Uberlândia. Faculdade de Computação. Disponível em: <http://files.sistemas2012.webnode.com.br/200000095-bf367bfb43/Tecnicas%20de%20Minera%C3%A7%C3%A3o%20de%20Dados.pdf>. Acesso em: 17 abr 2019.

ANDRADE, Rosângela Vieira de; SILVA, Aderbal Ferreira da; MOREIRA, Frederico Neiva; SANTOS, Helisbetânia Paulo Souza; DANTAS, Heloiza Ferreira; ALMEIDA, Iramiz Fereira de; LOBO, Leandra de Paula Brito; NASCIMENTO, Mirian Argolo. **Atuação dos Neurotransmissores na Depressão.** 2003. Faculdade de Farmácia do Planalto Central/União Educacional do Planalto Central – UNIPLAC. Disponivel em: <http://aloisiolatge.com.br/arquivos/academicos_2/02-atuacao_dos_neurotransmissores_na_depressao.pdf>. Acesso em: 06 out 2019.

ARAÚJO, Maria Teixeira; BATISTA, Mônica de Lourdes Souza; MAGALHÃES, Teresinha Moreira de. **OLAP: Características, Arquitetura e Ferramentas.** 2007. Instituto Vianna Júnior. Faculdades Integradas Vianna Júnior. Disponível em: <https://www.cin.ufpe.br/~ejvm/OLAP/200725803.pdf>. Acesso em: 03 out 2019.

BARDAGI, Marúcia Patta. **Evasão e comportamento vocacional de universitários: estudo sobre desenvolvimento de carreira na graduação.** 2007. 242 f. Tese (Doutorado em Psicologia). Universidade Federal do Rio Grande do Sul. Instituto de Psicologia. Curso de Pós-Graduação em Psicologia do Desenvolvimento. Rio Grande do Sul, 2007. 25 nov 2018.

BOSCARIOLI, Clodis; BEZERRA, Anderson; BENEDICTO, Marcos de; DELMIRO, Gilliard. **Uma reflexão sobre Banco de Dados Orientados a Objetos**. 2006. UNIOESTE - Universidade Estadual do Oeste do Paraná, FASP - Faculdades Associadas de São Paulo. Disponível em: <https://conged.deinfo.uepg.br/artigo4.pdf>. Acesso em: 03 out 2019.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas.** 2009.Technical Report - RT-INF_001-09 - Relatório Técnico. Instituto de Informática Universidade Federal de Goiás. Disponível em: <https://rozero.webcindario.com/disciplinas/fbmg/dm/RT-INF_001-09.pdf>. Acesso em: 16 abr 2019.

CRUZ, Bruno Campanella Cruz; MIRANDA, Bruno Gabriel Correa; TURCHETTE, Fellipe Barretto. **Conceitos de Business Intelligence por Meio de Estudos de Caso: Ferramentas Pentaho e Qlikview.** 2014. Universidade São Francisco.

Engenharia de Computação. Disponível em: <http://lyceumonline.usf.edu.br/salavirtual/documentos/2704.pdf>. Acesso em: 04 out 2019.

DAMASCENO, Marcelo. **Introdução a Mineração de Dados Utilizando o Weka.** 2010. Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte. Disponível em: <http://connepi.ifal.edu.br/ocs/index.php/connepi/CONNEPI2010/paper/viewFile/258/207>. Acesso em: 03 out 2019.

DOCTUM, REDE DE ENSINO. **Bolsas**. Portal Doctum. 2019. Disponível em: <https://www.doctum.edu.br/bolsas-e-parcelamentos/bolsas/>. Acesso em: 09 out 2019.

DUTRA, Renan Martins. **O uso de inteligência artificial para predição de evasão na Rede Doctum de Ensino**. 2015. 79 f. Trabalho de Conclusão de Curso - Faculdade Doctum de Caratinga. Caratinga-MG, Brasil. 25 nov 2018.

ELIAS, Diego. **Dados VS Informação: Qual a diferença?** 2019. Disponível em: <https://www.binapratica.com.br/dados-x-informacao>. Acesso em: 21 out 2019.

FERNANDES, Janderson Gabriel Limeira; SILVA, Nathália Alves Martimiano da; BROCK, Tarik Robledo; QUEIROGA, Ana Paula Garrido de; RODRIGUES, Luciene Cavalcanti. **Inteligência Artificial: Uma Visão Geral.** 2018. FATEC São José do Rio Preto. Tecnologia em Análise e Desenvolvimento de Sistemas. IFSP Votuporanga. Disponível em: <http://reeed.com.br/index.php/reeed/article/view/25/23>. Acesso em: 18 mar 2019.

FERREIRA, João; MIRANDA, Miguel; ABELHA, António; MACHADO, José. **O Processo ETL em Sistemas Data Warehouse.** 2010. Universidade do Minho. Departamento de Informática. Disponível em: < http://inforum.org.pt/INForum2010/papers/sistemas-inteligentes/Paper080.pdf>. Acesso em: 12 dez 2019.

FRITSCH, Rosangela; ROCHA, Cleonice Silveira da; VITELLI, Ricardo Ferreira. **A evasão nos cursos de graduação em uma instituição de ensino superior privada.** Revista Educação em Questão, Natal, v. 52, n. 38, p. 81-108, maio/ago. 2015. 25 nov 2018. Disponível em: <https://periodicos.ufrn.br/educacaoemquestao/article/view/7963/5724>. Acesso em: 16 abr 2019.

HEUSER, Carlos Alberto. **Projeto de Banco de Dados**. 1998. Séries Livros Didáticos. 4ªe. Instituto de Informática da UFRGS. 25 nov 2018. Disponível em: <http://www.fernandozaidan.com.br/pit-grad/Diversos/Livros_Disciplinas/Projeto_de_Banco_de_Dados_-_Carlos_Alberto_Heuser.pdf>. Acesso em: 16 abr 2019

HITACHI VANTARA. **Pentaho Data Integration**. 2019. Disponível em: <https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-data-integration.html>. Acesso em: 03 de out 2019.

HOED, Raphael Magalhães. **Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação**. 2016. 164 f. Trabalho de Conclusão de Mestrado Profissional em Computação Aplicada - Universidade de Brasília. Instituto de Ciências Exatas. Departamento de Ciência da Computação. Brasília-DF, Brasil. 25 nov 2018.

HOKAMA Daniele Del Bianco; CAMARGO Denis; FUJITA Francine; FOGLIENE João Luiz Valentim. **A Modelagem de Dados no Ambiente Data Warehouse**. 2004. Universidade Presbiteriana Mackenzie Faculdade de Computação e Informática. Disponível em: <http://meusite.mackenzie.com.br/rogerio/tgi/2004ModelagemDW.pdf>. Acesso em: 16 abr 2019.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Indicadores de Fluxo Escolar da Educação Básica.** Disponível em: <http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/inep-

divulga-dados-ineditos-sobre-fluxo-escolar-na-educacao-basica/21206>. Acesso em: 18 mar 2019.

MACHADO, Felipe Nery Rodrigues. **Tecnologia e projeto de DATA WAREHOUSE.** 3. ed. São Paulo: Érica, 2007. 317.

MINISTÉRIO DA EDUCAÇÃO. **Altos índices de desistência na graduação revelam fragilidade do ensino médio, avalia ministro.** Disponível em: <http://portal.mec.gov.br/component/tags/tag/32044-censo-da-educacao-superior>. Acesso em: 18 mar 2019.

NIEDERAUER, Juliano. **Integrando PHP 5 com MySQL**. Guia de Consulta Rápida. São Paulo, 2008. Disponível em: < http://www.martinsfontespaulista.com.br/anexos/produtos/capitulos/526209.pdf>. Acesso em: 03 de out 2019.

PALHARINI, Francisco de Assis. **Evasão, exclusão e gestão acadêmica na UFF: passado, presente e futuro**. Cadernos do ICHF: Série Estudos e Pesquisas. Universidade Federal Fluminense – Instituto de Ciências Humanas e Filosóficas. Niterói, 2010. Disponível em: <http://www.ichf.uff.br/pdf-docs/cadernosichf/CDI95-Palharini-EvasaoExclusaoGestao.pdf>. Acesso em: 16 de abr 2019.

RAPOSA, Erika de Oliveira Barrozo. **A utilização de técnicas de descoberta de conhecimento em ambiente acadêmico, aplicada ao problema de evasão escolar**. 2009. 66 f. Dissertação – Faculdade Doctum de Caratinga. Caratinga-MG, Brasil, 09 dez 2009.

RAMOS, Jefferson David Asevedo; SILVA, Leandro Gomes da; PRATA, David Nadler. **Inteligência Artificial e a Lei de Direitos Autorais**. 2018. Revista Cereus. Disponível em: <http://ojs.unirg.edu.br/index.php/1/article/view/2348/736>. Acesso em: 18 de abr 2019.

SALATAS, J. **Implementation of Elman Recurrent Neural Network in WEKA**.

2011. Disponível em: <jsalatas.ictpro.gr/implementation-of-elman-recurrent-neural-network-in-weka/> Acesso em: 06 out de 2019.

SCHEPS, Swain. **Business Intelligence for Dummies.** 1. ed. Indiana: Wiley Publishing Inc, 2008. 358.

SETZER, Valdemar W. **Dado, Informação, Conhecimento e Competência**. 2014. Depto. de Ciência da Computação, Universidade de São Paulo. Disponível em: <https://www.ime.usp.br/~vwsetzer/dado-info.html>. Acesso em: 01 de out 2019.

SILVA FILHO, Roberto Leal Lobo e; MOTEJUNAS, Paulo Roberto; HIPÓLITO, Oscar; LOBO, Maria Beatriz de Carvalho Melo. **A evasão no ensino superior brasileiro.** 2007. 2f. Cadernos de Pesquisa. Fundação Carlos Chagas. Disponível em: <http://www.institutolobo.org.br/imagens/pdf/artigos/art_045.pdf>. Acesso em: 18 mar 2019.

SISNEMA. **A Tecnologia do OLAP**. 2019. Disponível em: <http://sisnema.com.br/Materias/idmat002228.htm>. Acesso em: 03 de out 2019.

SOUZA, Flávio Clésio Silva de. **Classificação de portfólio de créditos não performados utilizando redes neurais artificiais Multilayer Perceptron.** 2013. Revista Gestão da Produção Operações e Sistemas. Disponível em: <https://revista.feb.unesp.br/index.php/gepros/article/view/1120/566>. Acesso em: 06 out 2019.

SOUZA, Francisco Ary Alves de. **Análise de desempenho da rede neural artificial do tipo multilayer perceptron na era multicore.** 2012. 65 f. Dissertação de mestrado ao Programa de Pós-Graduação em Engenharia Elétrica e de Computação da Universidade Federal do Rio Grande do Norte. Disponível em: <https://repositorio.ufrn.br/jspui/bitstream/123456789/15447/1/FranciscoAAS_DISSERT.pdf>. Acesso em: 06 out 2019.

UNIVERSITY OF WAIKATO. **Weka 3 – Machine Learning Software in Java**.

2019. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 01 de out 2019.

WEISZFLOG, Walter. Dicionário Michaelis Português.São Paulo:Editora Melhoramentos Ltda, 2015. Disponível em: <https://michaelis.uol.com.br/moderno-portugues/>. Acesso em: 04 de out 2019.

8. ANEXOS

8.1. ANEXO 1: AUTORIZAÇÃO DE DIVULGAÇÃO DE DADOS

rede de ensino DOCTUM

AUTORIZAÇÃO DE DIVULGAÇÃO DE DADOS

A Rede de Ensino Doctum, inscrita sob o CNPJ nº 19.322.494/0026-07, vem por meio deste autorizar em caráter temporário o acesso e a divulgação de dados sigilosos a **Miqueias Matias Caetano**, funcionário desta Instituição inscrito sob o CPF 106.828.756-00.

O uso e divulgação desses dados limita-se a exposição em seu Trabalho de Conclusão de Curso, além de limitar-se apenas ao tempo necessário para que este seja finalizado.

As informações por nós fornecidas delimitam-se a idade dos alunos matriculados, bem como os dados dos cursos de ensino superior referentes a matrícula destes alunos e o questionário socioeconômico. excluindo a divulgação desses dados a terceiros e tomando as devidas providências para não expor desnecessariamente as informações da instituição.

Hudson Silva de Souza
Coordenador de Tecnologia
Rede de Ensino Doctum

8.2. ANEXO 2: QUESTIONÁRIO SOCIOECONÔMICO

PERGUNTA	POSSÍVEIS RESPOSTAS
Existem problemas de saúde entre as pessoas que moram com sua família, inclusive você:	Não
	Sim
Mora fora da cidade onde está localizada a Faculdade e/ou Colégio?	Não
	Sim
De que forma são mantidas as despesas financeiras com a instituição?	Outros
	recebo ajuda de parentes
	sou responsável pelo meu próprio sustento
	sustentado pelos meus pais (ou por somente um dos pais)
	tenho bolsa de estudo
Você mora:	com cônjuge / companheiro (a)
	com os pais (ou somente com um dos pais
	em casa de familiares / casa de amigos
	em república / quarto / pensão / pensionato
	Outros
	Sozinho

Quem é (são) o (os) responsável (is) pela manutenção financeira do grupo familiar:	Outros
	outros membros do grupo familiar
	pai / mãe
	próprio estudante
	somente um dos pais
A Instituição de ensino na qual cursou as séries ou anos anteriores é:	parte em particular depois em pública
	parte em pública e depois em particular
	Particular
	particular com bolsa integral
	Pública
Qual o principal meio de transporte utilizado para chegar à Faculdade e/ou Colégio?	a pé / de bicicleta
	de carona
	oferecido gratuitamente por prefeitura e/ou escola
	Outros
	transporte coletivo pago diariamente com recursos próprios
	transporte locado, gasto mensal
	transporte próprio
Sua família reside em imóvel:	Alugado

	emprestado ou cedido
	Outros
	próprio, em pagamento / financiamento
	próprio, já quitado
	próprio, por herança
Sua família reside em:	bairro padrão médio
	bairro padrão popular
	fora do perímetro urbano
	Outros
	vila ou aglomerado
A família possui Automóveis? Qual a faixa de valor (da soma dos automóveis do aluno e do grupo familiar):	acima de R$ 30.000,00
	até R$ 10.000,00
	não possui automóvel
	R$ 10.000,01 a R$ 20.000,00
	R$ 20.000,01 a R$ 30.000,00
A família possui outros imóveis além do que habita? (Lotes, sítios, fazendas, casas na praia, aptos, barracões ou outros).	Não
	Sim