INSTITUTO DOCTUM DE EDUCAÇÃO E TECNOLOGIA

FACULDADES INTEGRADAS DE CARATINGA – FIC

CIÊNCIA DA COMPUTAÇÃO

O USO DE INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DE EVASÃO NA REDE DOCTUM DE ENSINO

RENAN MARTINS DUTRA

Caratinga

2015

Renan Martins Dutra

O USO DE INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DE EVASÃO NA REDE DOCTUM DE ENSINO

Monografia apresentada ao Curso de Ciência da Computação das Faculdades Integradas de Caratinga como requisito parcial para obtenção do título de bacharel em Ciência da Computação orientada pelo Prof. Maicon Vinícius Ribeiro

Caratinga – MG

2015

FACULDADES INTEGRADAS DE CARATINGA

TRABALHO DE CONCLUSÃO DE CURSO

TERMO DE APROVAÇÃO

TÍTULO DO TRABALHO

O USO DE INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DE EVASÃO NA REDE DOCTUM DE ENSINO

por

Renan Martins Dutra

Este Trabalho de Conclusão de Curso foi apresentado perante a Banca de Avaliação composta pelos professores Maicon Vinícius Ribeiro, Bruno Vieira Becattini e Glauber Luiz da Silva Costa, às 19 horas e 40 minutos do dia 14 de dezembro de 2015 como requisito parcial para a obtenção do título de bacharel. Após a avaliação de cada professor e discussão, a Banca Avaliadora considerou o trabalho aprovado, com a qualificação: Ótimo.

Trabalho indicado para publicação: (X)SIM ( )NÃO

Caratinga, 14 de Dezembro de 2015

Maicon Ribeiro
Professor Orientador e Presidente da Banca

Bruno Vieira Becattini
Professor Avaliador 1

Professor Avaliador 2

Renan Martins Dutra
Aluno(a)

Coordenador (a) do Curso

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus, pela força dada até o momento e por todas as pessoas importantes para mim que foram colocadas em meu caminho.

Agradeço principalmente a minha mãe por todo o apoio, minha irmã por fazer meus dias mais felizes, a Tais por todo o apoio moral e a paciência neste ano, a Camila pela paciência e ao Eduardo pela ajuda.

Agradeço também aos professores, por todos os ensinamentos e pela amizade. Um agradecimento especial ao professor Maicon, por suas orientações acadêmicas e de vida, você é um exemplo a ser seguido. Agradeço também aos professores Glauber Costa e Fabrícia Pires e a equipe da Flux Softwares, pela paciência, ensinamentos e contribuições no presente trabalho.

“Aqueles que passam por nós, não vão sós, não nos deixam sós. Deixam um pouco de si, levam um pouco de nós.”
Antoine de Saint-Exupéry

RESUMO

A evasão escolar é considerada um fenômeno complexo que alcança diversos níveis educacionais trazendo consequências negativas para a instituição, os estudantes, a sociedade e até o governo. O objetivo deste trabalho é criar uma metodologia capaz de prever de forma precoce a evasão de um aluno, utilizando tal informação para a criação de uma metodologia de retenção do aluno, minimizando a taxa de evasão. Criou-se um Data Warehouse com a finalidade de realizar mineração de dados na instituição de ensino Rede Doctum, utilizando a ferramenta denominada Weka e um algoritmo de redes neurais artificiais Multilayer Perceptron, classificando os alunos em evadidos e não evadidos, e informando o percentual de evasão da instituição, promovendo a criação de grupos de risco de evasão. O percentual de acerto foi de aproximadamente 95% durante a fase de treinamento e 93% durante a confirmação dos dados, sendo que o percentual geral não é suficiente para determinar a viabilidade dos resultados para uso prático. Os valores mais importantes para determinação da viabilidade são os percentuais de erros de cada classe. Analisando apenas o grupo de evadidos, durante a fase de treinamento obteve-se um acerto de 67% nas quais o discente realmente evadiu, mas durante a etapa de confirmação esse acerto foi de apenas 43%. Apesar de um bom desempenho geral, quando analisado de forma mais detalhada a metodologia se mostra falha, classificando de forma erronia mais da metade dos alunos que realmente evadiram. Ainda que os resultados não tenham suprido as expectativas, tal estudo abre caminho para o desenvolvimento de metodologias de análise mais apuradas que minimizariam possíveis erros, tornando viável sua aplicação institucional.

Palavras-chave: Evasão. Mineração de Dados. Inteligência Artificial. Redes Neurais Artificiais. Multilayer Perceptron. Data Warehouse. Weka.

ABSTRACT

The evasion is considered a complex phenomenon that reaches various educational levels bringing negative consequences for the institution, students, society and even the government. The objective of this work is to create a methodology able to predict early on the escape of a student using such information for the creation of a student's retention methodology, minimizing the dropout rate. It created a Data Warehouse in order to perform data mining on Rede Doctum de Ensino, using the so-called Weka tool and an algorithm of artificial neural networks Multilayer Perceptron, sorting students into dropouts and no dropouts, and stating the percentage of avoidance of the institution, promoting the creation of dropout risk groups. The hit percentage was approximately 95% during the training phase and 93% during confirmation of the data, and the overall percentage is not sufficient to determine the viability of the results for practical use. The most important values for determining the viability percentages are the errors for each class. By analyzing only the group escaped during the training phase afforded a 67% setting in which the student actually bolted, but during the step of confirming this arrangement was only 43%. Despite a good overall performance, when analyzed in more detail the methodology shown failure, sorting of wrong form more than half of the students who actually escaped. Although the results have not supplied expectations, this study paves the way for the development of more accurate methods of analysis which would minimize potential errors, making feasible its institutional application.

Keywords: Evasion. Data mining. Artificial intelligence. Artificial Neural Networks. Multilayer Perceptron. Data Warehouse. Weka.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE SIGLAS

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

MEC – Ministério da Educação

SGBD – Sistemas Gerenciadores de Bancos de Dados

MLP – *Multilayer Perceptron*

WEKA - *Waikato Environment for Knowledge Analysis*

EM - *Expectation Maximization*

FIES - Fundo de financiamento estudantil

PROUNI – Programa Universidade para Todos

IBM - *International Business Machines*

SUMÁRIO

1 INTRODUÇÃO

A evasão escolar é considerada um fenômeno complexo que alcança diversos níveis educacionais, trazendo consequências negativas para todos envolvidos no processo de ensino-aprendizagem. Os prejuízos provocados pela evasão atingem a instituição, os estudantes, a sociedade e até o governo. Esses fatores alteram os elementos emocionais, psicológicos e humanos refletindo em prejuízos sociais e econômicos aos sujeitos envolvidos (FIALHO e PRESTES, 2014).

O acesso ao ensino superior é um fato cada vez mais comum, que pode ser comprovado pelo crescimento do número de alunos matriculados a cada ano. No último senso realizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) foi registrado o recorde no número de alunos matriculados, chegando a aproximadamente 7,3 milhões de discentes, alcançando um percentual de aumento de 3,8% comparado ao ano anterior (INEP, 2014).

Apesar do crescimento no número de matrículas impulsionar a expansão das instituições de ensino superior, um fator alarmante para tais instituições é o controle da taxa de evasão, aliado a isso, o controle de evasão se torna mais complexo, em virtude do grande número de novos estudantes, sendo que em muitos casos nota-se a evasão de forma mais evidente quando não há meios para elaboração de medidas de retensão.

O controle da evasão nunca foi simples, mas devido ao amplo desenvolvimento e evolução da tecnologia, as instituições têm investido cada vez mais em estudos que utilizam essas novas tecnologias com a finalidade de criar métodos capazes de prever a evasão de um discente, com o objetivo de diminuir as taxas de evasão utilizando métodos computacionais.

Os bancos de dados das instituições tem crescido consideravelmente com dados relacionados aos discentes que ingressam. Esses dados estão sendo cada vez mais detalhados, mas geralmente são utilizados apenas para fins administrativos e acadêmicos. Diante disso, diversos estudos são concebidos dentro instituições de ensino, utilizando técnicas de mineração de dados com o objetivo de se criar uma metodologia capaz de auxiliar na retenção discente, criando grupos de risco ou prevendo possíveis alunos evasores, com o objetivo de melhorar o controle da evasão. Com o foco no setor gerencial, *Data Mining* ou mineração de dados, é

um novo processo que vem ganhando mais mercado nos últimos anos, e tem como objetivo realizar uma análise de dados em busca de informações que possam ser utilizadas pelos setores estratégicos da instituição.

Considerando os fatores expostos, pretende-se explorar os bancos de dados da Rede Doctum, em busca de padrões de evasão discente utilizando redes neurais artificiais, com a finalidade de usar estes padrões para criar grupos de risco de evasão e uma possível predição da evasão de determinado grupo de alunos, proporcionando a possibilidade de a instituição criar medidas preventivas com foco na redução da taxa de evasão.

Primeiramente é necessário compreender todo o contexto sobre a atual situação do ensino superior e as técnicas de mineração de dados, identificando quais os fatores que podem dificultar o uso destas metodologias para determinar sua viabilidade. As seções a seguir contextualizam os temas citados, demonstrando toda a metodologia utilizada para se alcançar os resultados.

2 REFERENCIAL TEÓRICO

Com intuito compreender os métodos de mineração e de redes neurais artificiais, foi realizado um estudo bibliográfico pesquisando trabalhos já feitos na área, ferramentas e o impacto que a utilização destas ferramentas tem no contexto educacional. Tal análise tem como intuito a criação de uma ferramenta que auxilie na elaboração de metodologias que visam minimizar a evasão discente.

2.1 MATRÍCULAS NO ENSINO SUPERIOR

Em 24 de abril de 2007 o Governo criou o decreto nº 6.096 que tem “o objetivo de criar condições para a ampliação do acesso e permanência na educação superior”, o que abriu as portas para novas políticas de educação e políticas de financiamento estudantil, facilitando dessa forma o ingresso de novos estudantes nas instituições de ensino superior. Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) no ano de 2013 o número de estudantes matriculados nos cursos presenciais e à distância de graduação chegaram a 7,3 milhões, divididos em instituições públicas e privadas, sendo que aproximadamente 6,15 milhões estão matriculados em cursos presenciais. Além disso, também vale ressaltar que:

> Nos últimos vinte anos, o Brasil assistiu a um notável processo de crescimento de seu ensino superior. No começo dos anos noventa, somavam-se 1.540.080 estudantes matriculados no ensino superior no Brasil. Esse número saltou para 2.694.245 de estudantes em 2000 e para 6.379.299 em 2011 (INEP, 2013).

Analisando os dados de matrículas disponibilizados pelo INEP a cada ano, é possível notar uma curva de crescimento do número de matriculados no país nos últimos anos. Considerando apenas os dados de matrículas em cursos presenciais, pode-se notar um crescimento médio de 4,1%, chegando a 6 milhões de alunos matriculados no ano de 2013.

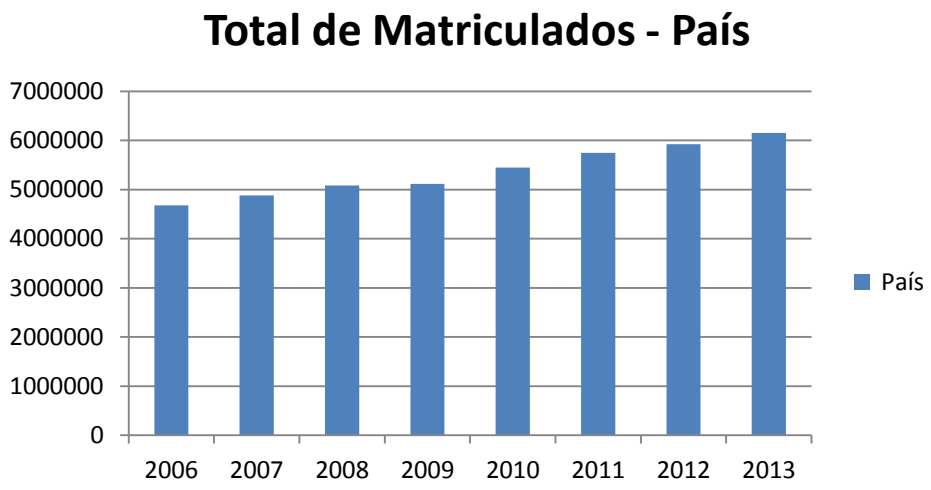

Gráfico 1: Total de matriculados – País – número de alunos matriculados nas instituições de ensino superior do país nos últimos 8 anos.
Fonte: Próprio autor

Analisando separadamente a região sudeste e o estado de Minas Gerais, é possível notar que os mesmos crescem em proporções semelhantes a do território brasileiro. Além disso, observamos que apenas na região sudeste há quase 50% do total de alunos matriculados em cursos de ensino superior do país.

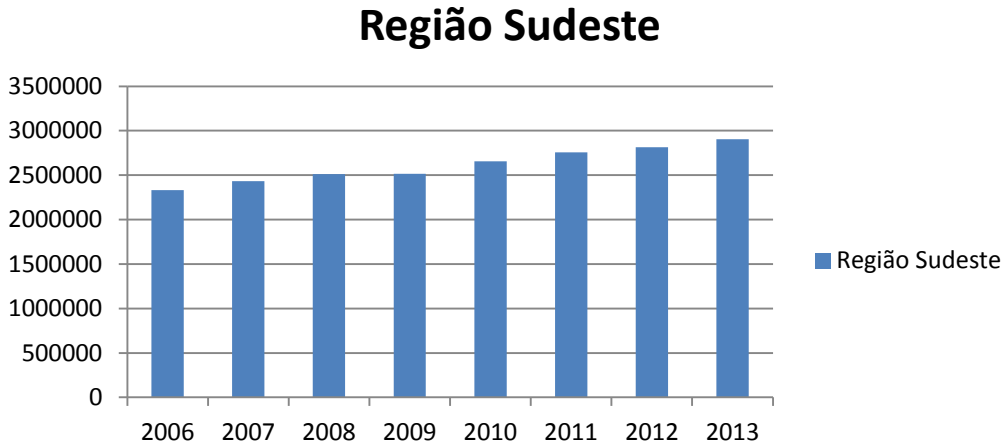

Gráfico 2: Região Sudeste – O número de alunos matriculados nas instituições de ensino superior da região sudeste nos últimos 8 anos.
Fonte: Próprio autor

Minas Gerais, apesar de ter alcançado um percentual de 6,05% de crescimento entre os anos de 2009/2010 obteve uma média de crescimento abaixo

da média nacional, chegando a aproximadamente 3,76%, mas ultrapassou a marca dos 600 mil alunos matriculados nos cursos de ensino superior no ano de 2013. Ou seja, apenas Minas Gerais possuiu no ano de 2013 aproximadamente 10% do total de alunos do território nacional.

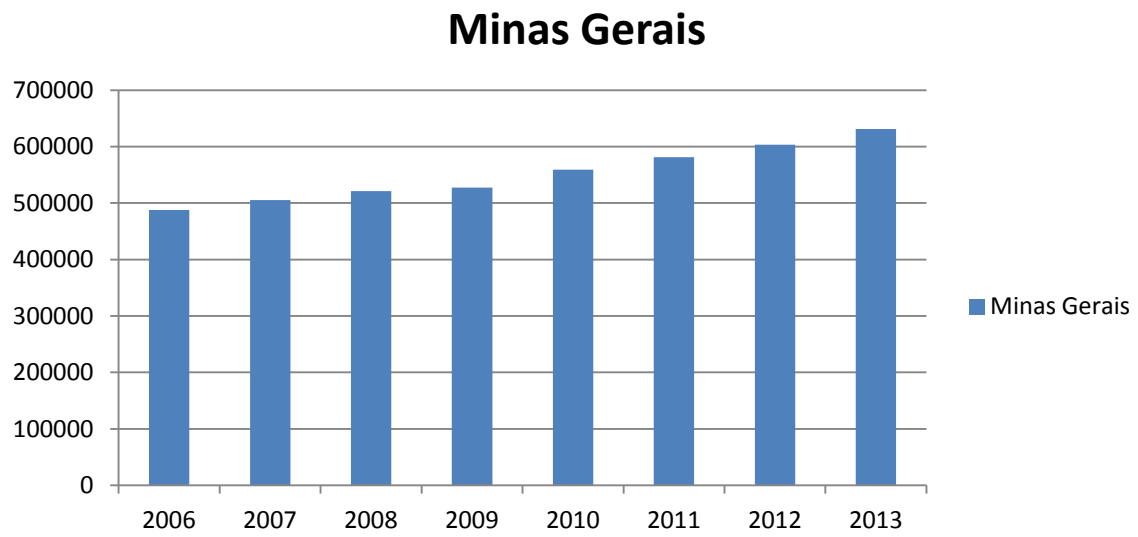

Gráfico 3: Minas Gerais – O número de alunos matriculados nas instituições de ensino superior do estado de Minas gerais nos últimos anos.
Fonte: Próprio autor

Atualmente a Rede de ensino Doctum possui 14 filiais espalhadas por 14 cidades em dois estados, sendo eles Minas gerais e Espírito Santo. Algumas destas sedes disponibilizam além do ensino superior, outras modalidades de ensino, como educação básica e cursos de ensino técnico de várias modalidades diferentes.

O próximo tópico explicará a história da instituição e seus atuais dados, com a finalidade de contextualizar a realidade dessa rede de ensino, que vem crescendo a cada ano.

2.1.1 Rede Doctum

A Rede de Ensino Doctum foi criada há mais de 70 anos e atualmente tem campus de ensino distribuídos em cidades de dois estados brasileiros, sendo eles Minas Gerais e Espírito Santo. Além da modalidade de ensino superior, ela

disponibiliza em suas filiais outras modalidades de ensino, como educação básica, pós-graduação e ensino técnico. (Doctum, 2015)

No estado de Minas Gerais a Rede Doctum possui unidades implantadas em várias cidades, sendo elas: Caratinga, Carangola, Ipatinga, Cataguases, Juiz de Fora, Leopoldina, Manhuaçu, Teófilo Otone, João Monlevade e Piau. As unidades que estão espalhadas por Espírito Santo foram integradas a Rede Doctum há aproximadamente 4 anos, sendo que se localizam nas cidades de Guarapari, Vila Velha, Vitória, Serra e Iúna. Logo após a incorporação destas instituições a rede, pôde-se notar um crescimento considerável no número de matriculados da rede, como mostra o gráfico 4 (Doctum, 2014).

Da mesma forma que os demais, a Rede de Ensino Doctum, experimentou de um aumento significativo no número de alunos matriculados nos últimos anos chegando um crescimento médio de 12% ao ano, indo de pouco mais de 5 mil alunos para mais de 10 mil alunos em 6 anos.

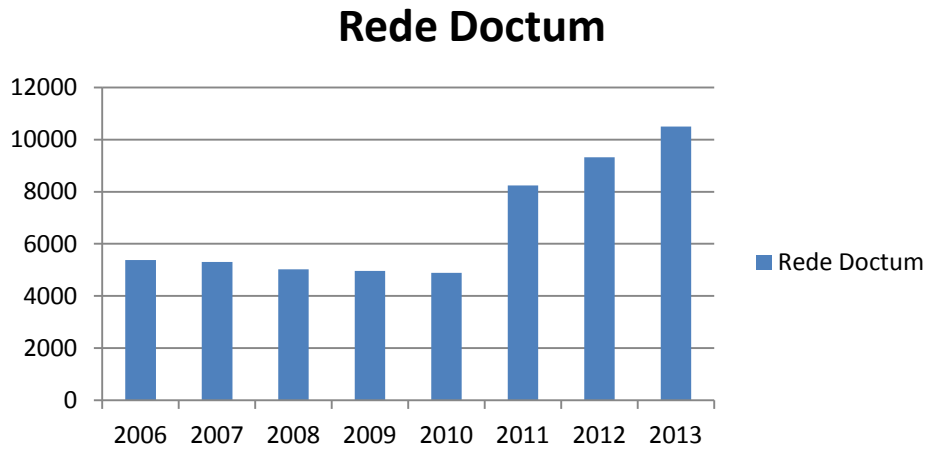

Gráfico 4: Rede Doctum – O número de alunos matriculados nas instituições de ensino superior do país nos últimos 8 anos.
Fonte: Próprio autor

Apesar do crescimento notável no número de alunos matriculados nos cursos de ensino superior, um fator que gera preocupação dentro de qualquer instituição de ensino é a evasão. Nos últimos a interrupção dos estudos nas diversas modalidades de ensino tem se tornado foco de estudos mais frequentemente, em busca do perfil de evasão, problemas causados pela mesma e até ferramentas capazes de auxiliar na retenção do discente, buscando diminuir este índice alarmante.

2.2 EVASÃO

Estima-se que a cada ano, aproximadamente 30% dos alunos que ingressam no ensino superior acabam evadindo em instituições privadas e cerca de 40% em públicas (BUARQUE, 2014). De acordo com os dados levantados por alguns estudos, uma fragilidade na escolha inicial e expectativas irreais sobre a carreira são uma das principais causas de evasão. (BARDAGI, 2007).

Fialho e Prestes (2014, p. 5) na tentativa de definir evasão, concluíram que “a evasão escolar é uma interrupção no processo de escolarização do estudante impossibilitando a conclusão do curso”. Gaioso (2005) define a evasão como a interrupção do ciclo de estudos. Com isso conclui-se que a evasão é a saída do aluno do curso à qual ele havia ingressado antes de concluí-lo.

Os altos índices de evasão acabam gerando pontos negativos para diversas áreas, desde à instituição da qual o discente evade, ao individuo e até à sociedade. Um dos principais pontos negativos gerados à instituição de ensino são os gastos extras com manutenção e investimentos feitos para manter um local de ensino adequado para seus alunos.

Segundo Gaioso (2005) e Saliba (2006) a evasão causa uma perda pessoal ao indivíduo e desperdício de investimento da instituição pois já que o aluno evadido ocupou uma vaga e não concluiu o curso.

Fialho e Prestes enfatizam ainda que:

> Os prejuízos provocados pela evasão escolar atingem os estudantes, a instituição, a sociedade e o governo. Assim como altera os aspectos emocionais, psicológicos, humanos e financeiros, refletindo direta e indiretamente na sociedade, provocando a ausência de capital humano qualificado para o mercado de trabalho e contribuindo para a elevação das taxas de desemprego e ampliando as desigualdades sociais. (FIALHO e PRESTES, 2014, p 1).

Sobre as responsabilidades da instituição no âmbito da evasão escolar Fialho e Prestes afirmam que:

> É preciso que a instituição de ensino esteja atenta às necessidades do aluno de modo a identificar e minimizar os motivos que podem levá-lo ao abandono, sejam problemas de ordem pessoal, curricular, econômica, institucional ou profissional (FIALHO e PRESTES, 2014, p 5).

Alguns estudos feitos buscando encontrar quais os principais motivos que levaram a evasão do aluno, mostram que além do fator financeiro, um dos principais motivos que o incentivaram está relacionado a sua insatisfação com o curso escolhido. Segundo Gois:

> Acreditava-se muito que a questão financeira era a vilã da história, mas percebemos em vários estudos que há várias outras razões. A principal delas talvez seja o desestímulo com o curso ou a falta de conhecimento prévio sobre a carreira escolhida no vestibular. Se o ensino for de qualidade e houver bons professores, no entanto, ele fará de tudo para continuar estudando. (GOIS, 2006).

Estudos realizados pelo MEC (Ministério da Educação) (1996) mostram diversas características que podem levar um aluno a evadir, estas características podem ser divididas em três categorias:

a) Fatores referentes às características individuais:

• Relativos às habilidades de estudo;

• Relacionados à personalidade;

• Decorrentes da formação escolar anterior;

• Vinculados à escolha precoce da profissão;

•Relacionados a dificuldades pessoais de adaptação à vida universitária;

• Decorrentes da incompatibilidade entre a vida acadêmica e as exigências do mundo do trabalho;

• Decorrentes do desencanto ou da desmotivação dos alunos com cursos escolhidos em segunda ou terceira opção;

• Decorrentes de dificuldades na relação ensino-aprendizagem, traduzidas em reprovações constantes ou na baixa frequência às aulas;

• Decorrentes da desinformação a respeito da natureza dos cursos;

• Decorrente da descoberta de novos interesses que levam à realização de novo vestibular.

Segundo o MEC a maioria dos estudantes que se encaixam nesta primeira categoria acabaram de sair do ensino fundamental e estão iniciando os estudos no ensino superior, geralmente com idade entre 16 e 18 anos de idade. Pode-se concluir com isso que além da situação socioeconômica do aluno, a falta de preparo e orientações sobre os estudos referentes ao curso que o discente pretende cursar,

podem levá-lo a uma escolha precipitada, com expectativas falhas relacionadas ao curso, ocasionando a evasão nos primeiros períodos (MEC, 1996). O Segundo fator apontado pelo MEC está ligado à instituição e aponta diversas características que serão abordadas a seguir.

b) Fatores Internos da instituição:

• Peculiares a questões acadêmicas; currículos desatualizados, alongados; rígida cadeia de pré-requisitos, além da falta de clareza sobre o próprio projeto pedagógico do curso;

• Relacionados a questões didático-pedagógicas: por exemplo, critério impróprio de avaliação do desempenho discente;

• Relacionados à falta de formação pedagógica ou ao desinteresse do docente;

• Vinculados à ausência ou ao pequeno número de programa institucional para o estudante, como Iniciação Científica, Monitoria, programas PET (Programa Especial de Treinamento), etc.;

• Decorrentes da cultura institucional de desvalorização da docência na graduação;

• Decorrentes de insuficiente estrutura de apoio ao ensino de graduação: laboratórios de ensino, equipamentos de informática, etc.;

• Inexistência de um sistema público nacional que viabilize a racionalização da utilização das vagas, afastando a possibilidade da matrícula em duas universidades.

Ao matricular-se em um curso, o aluno geralmente vem com expectativas voltadas à sociedade e seu futuro profissional, mas frustrações com a estrutura curricular limitada ou será adequada para o seu futuro, podem levar o discente a optar por mudança ou abandono do curso. Além disso, a falta de docentes, somado ao despreparo destes com os procedimentos didáticos, agrava o desenvolvimento das suas práticas, tornando comum a ocorrência de baixo rendimento dos alunos, aumentando a possibilidade de evasão discente cada vez mais (MEC, 1996). O Terceiro fator apontado pelo MEC aponta demonstra características que estão diretamente ligadas a carreira profissional, social e financeira da profissão que o discente exerceria no mercado de trabalho, como pode-se ver a seguir.

c) Fatores externos a instituição:

• Relativos ao mercado de trabalho;

• Relacionados ao reconhecimento social da carreira escolhida;

• Afetos à qualidade da escola de primeiro e no segundo grau;

• Vinculados a conjunturas e econômicas específicas;

• Relacionados à desvalorização da profissão.

• Vinculados a dificuldades financeiras do estudante;

• Relacionados às dificuldades de atualizar-se a universidade frente aos avanços tecnológicos, econômicos e sociais da contemporaneidade;

• Relacionados à ausência de políticas governamentais consistentes e continuadas, voltadas ao ensino de graduação.

Diversas vezes o aluno acaba por evadir ao deparar-se com um mercado de trabalho que pode lhe trazer eventuais dificuldades profissionais futuras, como baixa remuneração e dificuldades financeiras. Devido a isso, ainda que tenha vocação para a profissão, tais fatores acabam impelindo a opção por uma nova carreira ou simplesmente por uma desistência (MEC, 1996).

A evasão é certamente um dos maiores problemas que afligem as instituições de ensino em geral e a busca de suas causas tem sido um objeto de estudo em muitas pesquisas educacionais. Além disso, no Brasil há poucas instituições que possuem um programa de combate a evasão com planejamento de ações e acompanhamento de resultados (FIALHO E PRESTES, 2007).

A evolução dos sistemas de informação na última década tem facilitado a captação de dados de diversas formas. Devido a isto, as instituições de ensino desenvolveram métodos para captar melhor os dados dos alunos matriculados. O crescimento significativo dos dados armazenados nos bancos de dados proporcionou o estudo e criação de modelos de predição de evasão, tornando cada vez mais comum a realização de pesquisas feitas nesta área, como os trabalhos acadêmicos feitos por Raposa (2009), Rigo et al (2012) e Martinho(2014), que utilizam estes dados e ferramentas computacionais capazes de analisá-los, em busca de padrões de evasão, a fim de elaborar uma metodologia capaz de proporcionar às instituições de ensino uma forma de diminuir o índice de evasão.

2.3 ARMAZENAMENTO DE DADOS

A partir da evolução dos sistemas de informação e dispositivos de armazenamento, os processos de captação dos dados foram otimizados com a utilização de ferramentas computacionais, devido a isso os bancos de dados tornaram-se uma ótima opção para armazenamento de dados e passaram a ser considerados essenciais na sociedade moderna, além de representarem um papel crítico em quase todas as áreas da computação (ELMASRI, 2005)

Atualmente a utilização dos sistemas de bancos de dados, tem facilitado o controle das principais atividades das instituições onde os mesmos são utilizados. Uma instituição de ensino, por exemplo, tem processos como lançamento de notas, emissão de boletos e matrículas de alunos melhorados, tornando mais fácil o seu acompanhamento (RAPOSA, 2009).

Moraes e Souza definem banco de dados como:

> De uma forma conceitual podemos dizer que os bancos de dados são formados por modelos de informações originadas do mundo real. Os modelos mostram a estrutura da informação desde o seu processo de modelagem, onde são captadas do mundo real o que se quer do banco de dados, até a descrição de sua estrutura física de armazenamento, onde e como estarão localizados no banco de dados. (MORAES e SOUZA, 2000, p. 6).

Segundo Takai et al “Uma base de dados é uma coleção de dados logicamente relacionados, com algum significado. Associações aleatórias de dados não podem ser chamadas de base de dados;” (TAKAI et al, 2005, p. 14).

Elmasri e Navathe (2005), em seu livro de sistemas de bancos de dados, conceitua bancos de dados como uma “coleção de dados relacionados que podem ser gravados e possuem algum significado implícito” (ELMASRI e NAVATHE, 2005).

Os primeiros bancos de dados surgiram em 1950 com a criação dos discos rígidos, mas eram limitados e seus problemas eram claros, como repetição de dados. Nos anos 70, com a necessidade de aumentar a independência dos dados nos Sistemas Gerenciadores de Bancos de Dados (SGBD), foi criado o modelo relacional, que se mostrou ser o mais flexível e adequado para solucionar diversos problemas dos seus antecessores (TAKAI et al, 2005), e hoje são predominantes no mercado. (MORAES e SOUZA, 2000).

Praciano (2013) diz que, “Em um banco de dados relacional, todos os dados são guardados em tabelas. Estas têm uma estrutura que se repete a cada linha, como você pode observar em uma planilha. São os relacionamentos entre as tabelas que as tornam ‘relacionais’.”

Além dos bancos de dados relacionais existem outros tipos de bancos de dados, com características e funcionamento distintos, que serão brevemente explicados a seguir.

2.3.1 Tipos de bancos de dados

O primeiro modelo de banco de dados criado era chamado de modelo hierárquico, e surgiu no início da década de 60, a partir deste surgiram outros, que foram criados com a principal finalidade de resolver problemas que o primeiro possuía.

Segundo Takai et al:

> O modelo hierárquico foi o primeiro a ser reconhecido como um modelo de dados. Seu desenvolvimento somente foi possível devido à consolidação dos discos de armazenamento endereçáveis, pois esses discos possibilitaram a exploração de sua estrutura de endereçamento físico para viabilizar a representação hierárquica das informações. Nesse modelo de dados, os dados são estruturados em hierarquias ou árvores. Os nós das hierarquias contêm ocorrências de registros, onde cada registro é uma coleção de campos (atributos), cada um contendo apenas uma informação. O registro da hierarquia que precede a outros é o registro-pai, os outros são chamados de registros-filhos. (TAKAI, et al, 2005, p.6).

Este modelo apesar de funcional apresentava alguns graves problemas, como repetição de dados citado anteriormente, flexibilidade insuficiente e ineficiente quando ao acessar os registros utilizando novas consultas e transações (ELMASRI e NAVATHE, 2005).

Nos anos 70 houve uma necessidade de melhoria dos bancos de dados, devido a isso surgiu um novo conceito em banco de dados que foi chamado de modelo relacional. Esse novo modelo trazia algumas melhorias, como o aumento da independência de dados e novas funções que melhoravam o armazenamento e a recuperação de dados. Desta forma o modelo relacional mostrou-se mais flexível e

adequado para solucionar diversos problemas relacionados a bancos de dados (TAKAI, et al, 2005).

Nos anos 80, surgiram as linguagens orientadas a objetos, e com elas a necessidade de armazenar e compartilhar objetos e dados mais complexos, nesse momento um novo modelo de bancos de dados foi criado, o orientado ao objeto. Eles foram rapidamente considerados concorrentes dos modelos relacionais, pois possuíam estruturas de dados mais genéricas, mas a complexidade e a falta de padrões no início contribuíram para a limitação do seu uso (ELMASRI e NAVATHE, 2005).

Apesar de serem amplamente utilizados e cada vez mais dados serem captados diariamente, os bancos de dados possuem um problema, uma grande massa de dados por si só não possui significado expressivo, tornando necessário o uso de metodologias e ferramentas computacionais para o tratamento, e então transformá-los em informação útil, que será usada para fins específicos (GOLDSCHMDIT e PASSOS, 2005).

A seguir será explicado mais detalhadamente a diferença entre dado, informação e conhecimento, para uma melhor contextualização deste assunto sobre banco de dados.

2.3.2 Dados, informação e conhecimento

Os dados por si só não carregam significado expressivo, devido a isso é possível traçar uma diferença notável entre dados, informações e conhecimento. Segundo Goldschmidt e Passos (2005) dados são a base de uma pirâmide de hierarquia e são os principais itens, que são captados e armazenados por recursos da tecnologia de informação. As informações são os dados já processados com algum significado. O topo da pirâmide é representado pelo conhecimento, que é a aplicação prática das informações e os dados com um conjunto de padrões.

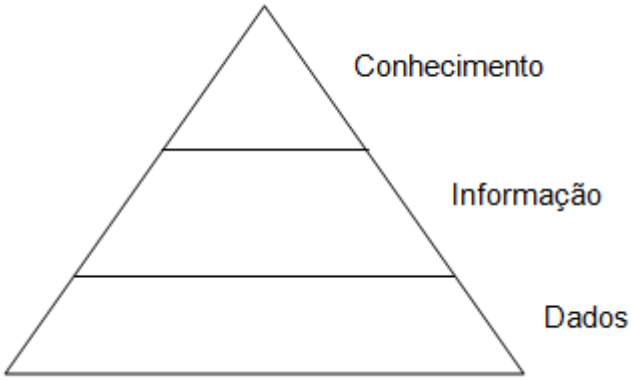

Figura 1: Hierarquia existente entre dados, informação e conhecimento.
Fonte: Próprio autor

Desta forma, entende-se que, dados são fragmentos de informação e não carregam nenhum significado expressivo quando isolados, por exemplo, a idade dos alunos de um curso. Já a informação é uma representação expressiva dos dados, contendo algum significado, que varia de acordo com a necessidade de quem precisa de uma determinada informação, por exemplo, saber a idade média das pessoas que estudam neste curso. O conhecimento é o uso prático desta informação, como por exemplo, sabendo qual a idade média dos alunos que cursam um determinado curso, criar estratégias direcionadas a um público específico para captação de alunos.

Segundo Goldschmdit, e Passos:

> A análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Portanto, torna-se imprescindível o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada contexto de aplicação. (GOLDSCHMDIT e PASSOS, 2005).

Com a finalidade de ajudar na busca de informações foram criadas algumas metodologias e ferramentas como o *Data Warehouse* e a mineração de dados, esses tópicos serão explicados a seguir para se ter um melhor entendimento do trabalho em questão.

2.4 DATA WAREHOUSE

Em uma tradução mais literal da palavra, *Data Warehouse* significa “Armazém de dados”, ou “deposito de dados”. Geralmente são utilizados em empresas para manter as grandes massas de dados consolidadas, facilitando a visualização das informações. Segundo Elmasri e Navathe “os *Data Warehouse* têm características distintas de que são direcionados principalmente para aplicações de apoio a decisões. Eles são otimizados para recuperação de dados.” (ELMASRI e NAVATHE, 2005).

Goldschmdit e Passos (2015) afirmam que deve-se reforçar a diferença entre um banco de dados tradicional e o *data Warehouse*, realçam ainda que nas bases de dados tradicionais os dados são normalmente encontrados na forma analítica, ou seja, são representados de forma mais detalhada possível, voltados para o uso nos setores operacionais e corporativos, já no *Data Warehouse* os dados encontram-se consolidados, de forma a apresentar informações para os níveis gerenciais e estratégicos da empresa. Os *Data Warehouses* devem armazenar dados históricos da empresa, viabilizando as consultas, descobertas de tendências e análises estratégicas.

Portanto entende-se que os *Data Warehouses* são criados com um objetivo específico: ajudar as empresas na tomada de decisões, auxiliando a obtenção de informações relevantes as decisões tomadas pelos setores estratégicos e que apesar de serem considerados um banco de dados, eles diferem dos bancos tradicionais.

Oliveira (2009) destaca em sua obra As principais características de um *Data Warehouse,* sendo elas:

> Orientado a Assunto: A primeira característica de um Data Warehouse é que ele está orientado ao redor do principal assunto da organização. O percurso do dado orientado ao assunto está em contraste com a mais clássica das aplicações orientadas por processos/funções ao redor dos quais os sistemas operacionais mais antigos estão organizados.
> Integrado: Facilmente o mais importante aspecto do ambiente de Data Warehouse é que dados criados dentro de um ambiente de Data Warehouse são integrados. SEMPRE. COM NENHUMA EXCEÇÃO. A integração mostra-se em muitas diferentes maneiras: na convenção consistente de nomes, na forma consistente das variáveis, na estrutura consistente de códigos, nos atributos físicos consistentes dos dados, e assim por diante.
> Não Volátil: sempre inserido, nunca excluído.

> Variante no Tempo: posições históricas das atividades no tempo (OLIVEIRA, 2009, p. 2,3).

No mercado atual a competitividade entre as empresas cresce a cada dia, fazendo necessário que a tomada de decisões de uma empresa seja rápida e dinâmica. Devido a isso, tornou-se imprescindível a utilização de sistemas de informações que auxiliem os setores estratégicos, por estes e outros motivos o *Data Warehouse* ganhou grande força. Outra função na qual o Data Warehouse é amplamente utilizado é no auxílio da mineração de dados, que será explicada a seguir.

2.5 PASSOS PARA A MINERAÇÃO DE DADOS

A mineração de dados ou *data mining,* é uma metodologia que auxilia a busca por padrões e informações ainda desconhecidas nas bases de dados, é comumente utilizada em conjunto com um *Data Warehouse*, pois o mesmo detêm os dados de uma empresa já consolidados com foco na tomada de decisões (ELMASRI e NAVATHE, 2005).

Segundo Camilo e Silva (2009, p. 3). “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.” (CAMILO e SILVA, 2009).

A mineração de dados começou a ser utilizada nos anos 80, quando as empresas começaram a se preocupar com os grandes volumes de dados armazenados e sem uso. Inicialmente a mineração de dados consistia apenas em retirar informações de grandes bases de dados, mas atualmente ela consegue extrair informações voltadas ao levantamento de necessidades reais, como por exemplo, a realização de marketing para um cliente, buscando informações sobre seus comportamentos, preferências e hábitos de compras (AMO, 2004).

Com a finalidade de realizar a mineração de dados deve-se passar por algumas etapas essenciais, sendo elas a etapa de seleção dos dados, pré-processamento, transformação e finalmente a mineração. A seguir será explicado no que consiste cada etapa.

2.5.1 Pré-processamento

A etapa de pré-processamento de dados é subdividida em outras etapas, responsáveis por captar, tratar, transformar e armazenar os dados em um novo banco de dados, dando origem a um Data Warehouse (GOLDSCHMDIT e PASSOS, 2005).

Os bancos de dados tradicionais, possuem uma massa de dados muito grande e detalhada, por isso se torna necessário selecionar apenas os dados necessários para a aplicação, tratá-los e armazená-los. Essa etapa é necessária pois existem dados que comprometem a qualidade dos mesmos, como registros inconsistentes e duplicados que devem ser tratados (SCHMITT, 2005).

2.5.1.1 Seleção de dados

Nesta etapa é feita a seleção de quais dados irão compor o *Data Warehouse* que será utilizado para a busca de informações, com a finalidade de criar um banco de dados mais limpo e específico para a aplicação. Para a seleção dos dados, deve-se tomar os devidos cuidados, pois apenas os dados que serão utilizados na mineração devem ser extraídos, por isso é necessário fazer uma análise detalhada e minuciosa dos dados que compõe o banco. Geralmente os dados são encontrados em bancos de dados tradicionais, organizados em grandes e complexas tabelas (GOLDSCHMDIT, e PASSOS, 2005).

2.5.1.2 Limpeza dos dados

Segundo Schmitt (2015) “As rotinas para limpeza de dados consistem em uma investigação para detectar registros incompletos, duplicados e dados incorretos.” (SCHMITT, 2005, p.25).

Ao término desta etapa, os dados terão uma melhor qualidade, aumentando a confiança nos resultados obtidos. Após a seleção e limpeza dos dados, a próxima etapa é a transformação dos dados, esta etapa é caracterizada pela manipulação dos dados para que estejam no formato adequado para a análise usando as ferramentas computacionais.

2.5.1.3 Transformação

Nesta etapa os dados são extraídos e tratados de forma que estejam no formato adequado para serem armazenados em um novo banco de dados, criando então o *Data Warehouse* que fornecerá os dados utilizados para a análise (GOLDSCHMDIT e PASSOS, 2005)

Os dados captados nesta etapa devem ser tratados com a finalidade de se fornecer melhora na qualidade dos dados, retirando ruídos, inconsistências e dados considerados estranhos, resultando assim no refinamento do resultado final. Com a finalidade de facilitar as etapas de pré-processamento, foram criadas algumas ferramentas de extração e integração de dados, como *Pentaho Data Integration*, *Talend Open Studio e a Oracle Data Integration.*

Segundo Santos, a ferramenta *Pentaho Data Integration* é uma das mais usadas nos últimos anos no mercado brasileiro (SANTOS, 2013). Devido a isso o funcionamento desta ferramenta será detalhado a seguir.

2.5.1.4 Pentaho data integration

A ferramenta disponibilizada pela *Pentaho Corporation,* denominada *Pentaho Data Integration*, também conhecida como KETTLE, possui uma interface gráfica intuitiva com um design de “arraste e solte” (*drag and drop*), onde o usuário deve selecionar o componente desejado e arrastá-lo para onde deseja para utilizá-lo. Essa ferramenta utiliza a linguagem de programação Java, devido a isso é possível utilizá-la nos sistemas operacionais Windows e Linux (PENTAHO, 2015).

A ferramenta disponibiliza ao usuário uma infinidade de componentes distintos, com o intuito de facilitar a extração e transformação de dados, que podem ser provenientes de diversas fontes, como arquivos de texto, planilhas e principalmente de bancos de dados, para armazená-los em outro arquivo ou banco de dados (PENTAHO, 2015).

Uma das principais características dessa ferramenta é a possibilidade de utilização de diversas fontes de dados simultaneamente, com o intuito de fazer o cruzamento das mesmas, ou simplesmente agrupar dados de diversas fontes em apenas um local (PENTAHO, 2015). Por exemplo, uma instituição que possui diversas filiais e cada uma delas possui um banco de dados diferente, tal ferramenta é capaz de unir estas diversas fontes em apenas um local, este passo pode ser observado na Figura 2.

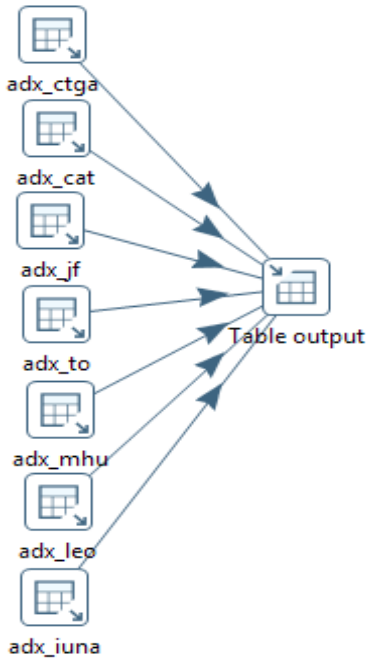

Figura 2: Diversas fontes de dados com o mesmo destino.
Fonte: Próprio autor

Na ferramenta há diversos componentes, denominados *steps*, que são passos, onde cada *step* um é considerado a menor unidade do processo de transformação dos dados e é responsável por uma determinada ação, seja ela buscar os dados, armazenar os dados ou modificá-los de diversas formas. Cada step é ligado a um próximo step por uma linha, chamada ”hop”, criando uma rotina

onde cada passo será executado logo após o término do seu antecessor (SANTOS, 2013).

Figura 3: Sequência de execução do *Pentaho Data Integration*
Fonte: Próprio autor

A etapa de pré-processamento dos dados para a criação do *Data Warehouse* pode ser feita pela utilização dos componentes da *Pentaho Data Integration*, sendo que ao definir de onde se buscará os dados é possível a escolha de quais dados serão utilizados. Com os *steps* é possível alterar os dados da forma mais conveniente ao usuário, para que sejam armazenados em seguida no banco de dados que conterá os dados utilizados na mineração de dados, explicada a seguir.

2.5.2 Mineração de dados

Nesta etapa é realizada a busca efetiva por informações, utilizando ferramentas computacionais que são capazes de analisar grandes massas de dados em busca do tipo de informação que se necessita (ELMASRI e NAVATHE, 2005).

Ao término da mineração de dados, as informações obtidas são coletas e se segue uma análise de quais informações podem ser relevantes para uso prático na tomada de decisões, o uso prático desta informação é denominado conhecimento (GOLDSCHMDIT e PASSOS, 2005)

Para a mineração de dados há diversas técnicas distintas que são usadas de acordo com o tipo de informação que o usuário necessita encontrar. Essas técnicas serão explicadas abaixo mais detalhadamente.

2.6 TÉCNICAS DE MINERAÇÃO DE DADOS

Apesar de existirem diversas técnicas distintas para a mineração de dados, há algumas que tem seu uso mais difundido, como a classificação, agrupamento, associação (CAMILO e SILVA, 2009).

Cada uma destas técnicas possui uma aplicação específica quando se deseja buscar informações em bases de dados, contendo também diversos algoritmos que ajudarão na busca. A seguir será explicado um pouco mais sobre cada técnica.

2.6.1 Classificação

Segundo Amorim (2006), “A classificação é uma das mais utilizadas técnicas de mineração de dados, simplesmente porque é uma das mais realizadas tarefas humanas no auxílio à compreensão do ambiente em que se vive” (AMORIM, 2006, p. 23).

Classificação é o processo que busca encontrar um conjunto de funções, com a finalidade de usá-las para distinguir classes ou conceitos, com o propósito de utilizar estas funções posteriormente na tentativa de predizer classes de objetos que ainda não foram classificados (AMO, 2004)

Goldschmdit e Passos reforçam:

> A classificação consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos predefinidos, denominados classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classes em que tais registros se enquadram (GOLDSCHMDIT e PASSOS, 2005).

Conclui-se então que a classificação procura por padrões de ocorrências, classificando os registros em classes já predefinidas, que posteriormente utiliza estes padrões com a finalidade de prever a que classe cada novo registro se encaixa.

Para a tarefa de classificação, geralmente são utilizados algoritmos de redes neurais artificiais, algoritmos genéticos, lógica indutiva e árvores, como alguns exemplos.

2.6.2 Agrupamento

Diferente da classificação, o agrupamento não precisa das classes predefinidas pelo usuário, pois quem definirá os grupos de registros é o próprio algoritmo. Ele identificará similaridades entre um conjunto de registros, agrupando aqueles que forem mais idênticos. (GOLDSCHMDIT e PASSOS, 2005)

Segundo Amorim (2006):

> Na análise de agrupamentos, os grupos ou classes são construídos com base na semelhança entre os elementos, cabendo ao analisador das classes resultantes avaliar se estas significam algo útil. Por exemplo, agrupar sintomas pode gerar classes que não representem nenhuma doença explicitamente, uma vez que doenças diferentes podem possuir os mesmos sintomas (AMORI, 2006, p.26).

Para a técnica de clusterização pode-se utilizar diversos algoritmos, tais como: K-Means, K-Modes, K-prototypes, K-Medoids, Kohonen, entre outros.

2.6.3 Associação

As regras de associação, diferente das demais já apresentadas, buscam por registros que ocorram frequentemente de forma simultânea. Um exemplo conhecido, é o do mercado americano que, utilizando essa metodologia descobriu que homens que compravam fraldas nas vésperas de finais de semana, também compravam cerveja (GOLDSCHMDIT e PASSOS, 2005).

A técnica de associação conta com vários algoritmos para a busca das ocorrências, sendo eles: algoritmo Apriori, GSP, DHP entre outros.

A finalidade deste trabalho é criação de uma metodologia capaz de predizer a evasão de alunos de uma instituição de ensino. Pelo fato de já existirem as classes

predefinidas, a técnica de mineração de dados que mais se encaixa para uso neste trabalhos são as de classificação, devido a isso optou-se pela utilização de redes neurais artificiais, que é uma área que vem se desenvolvendo cada vez mais nos últimos anos. A seguir temos a contextualização sobre redes neurais e suas aplicações.

2.7 REDES NEURAIS ARTIFICIAIS

As redes neurais artificiais são modelos matemáticos que buscam imitar o funcionamento de um cérebro humano e seus neurônios biológicos. Esses modelos matemáticos têm a capacidade de adquirir, armazenar e utilizar conhecimento experimental, buscando simular habilidades humanas como aprendizado, generalização, associação e abstração de forma computacional (GOLDSCHMDIT e PASSOS, 2005).

Varella (2014) define neurônios como células com capacidade de estabelecer uma conexão e receber estímulos do ambiente que vão emitir impulsos nervosos ao cérebro. Os neurônios são compostos por 3 partes principais, sendo elas: o corpo celular, que acomoda o núcleo e é responsável por coletar informações provenientes de outros neurônios, o axônio que é responsável por conduzir os impulsos e os dendritos, que são ramificações do corpo celular e do axônio e realizam a comunicação entre os neurônios por meio das sinapses (VARELLA, 2014).

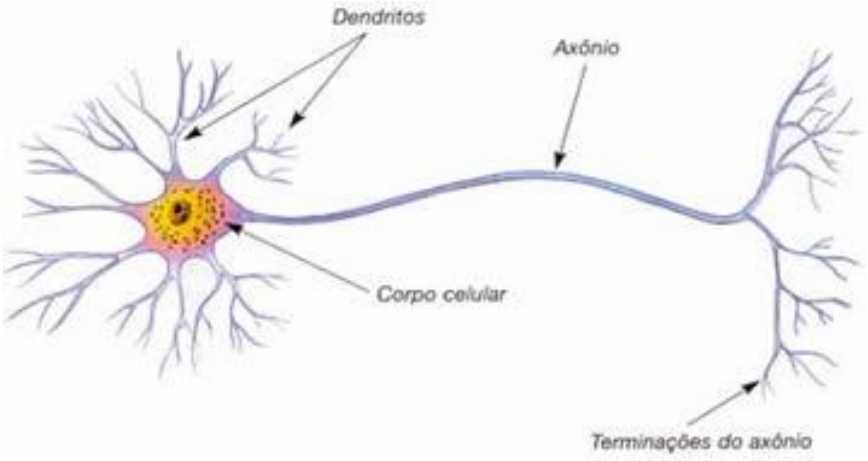

Figura 4: Estrutura que compõe o neurônio.
Fonte: SOBIOLOGIA.

O neurônio utilizado nas redes neurais artificiais possui uma estrutura semelhante a do neurônio biológico, onde as diversas entradas do neurônio são representadas pelos impulsos elétricos captados pelos dendritos. Cada entrada é multiplicada por um valor, denominado peso, gerando entradas ponderadas. Essas entradas ponderadas são somadas, obtendo um valor denominado Net. Tal valor é o potencial de ativação do neurônio, caso o resultado seja suficiente para a ativação do neurônio, ele será ativado, caso contrário, permanecerá inativo (BARRA, 2013 e GOLDSCHMDIT e PASSOS, 2005).

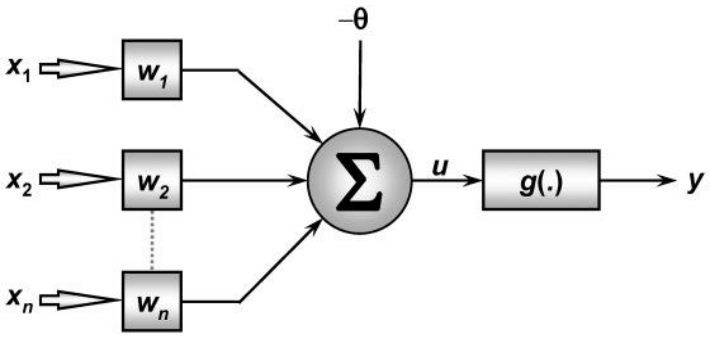

Figura 5: Estrutura que compõe o neurônio artificial. A primeira parte representa as entradas, no centro são feitos os cálculos, caso o resultado atinja o valor de ativação, o neurônio será ativado.
Fonte: BARRA, 2013.

Segundo Brumatti:

> Uma rede neural artificial é formada pela combinação de diversos neurônios artificiais. As entradas podem ser conectadas em muitos neurônios, resultando, assim, em uma série de saídas, onde cada neurônio representa uma saída. Comparando com o sistema biológico, essas conexões representam o contato dos dendritos com outros neurônios, formando assim as sinapses (BRUMATTI, 2014, p. 3).

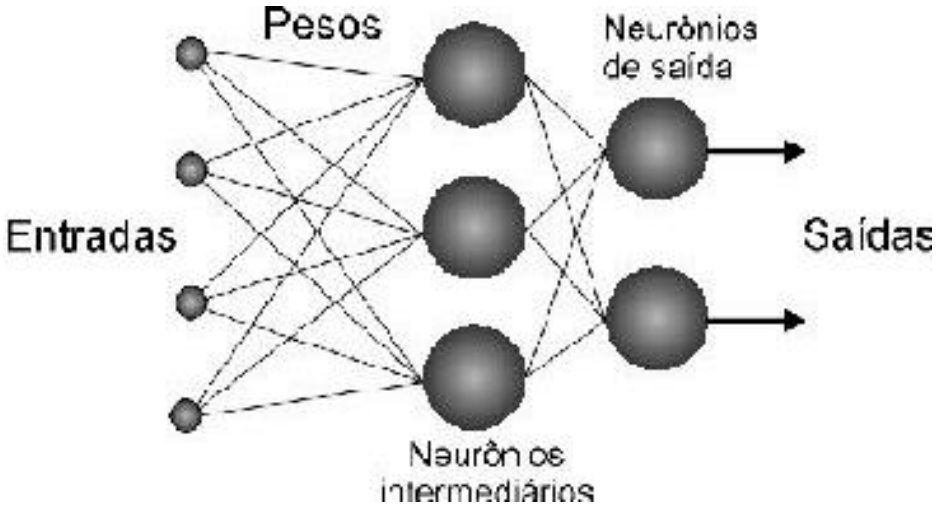

Figura 6: Estrutura de uma rede neural artificial que contém 4 entradas, duas camadas e duas saídas. Fonte: BRUMATTI, 2014.

Nas redes neurais artificiais, as sinapses (comunicação entre neurônios) de um neurônio são simuladas com a utilização dos pesos, que é o local onde a informação é armazenada. A seguir será dada uma explicação mais detalhadas a cerca deste assunto.

2.7.1 Pesos

Os neurônios biológicos utilizam processos químicos com a finalidade de transmitir sinais entre neurônios e outras células, processo denominado sinapse. No neurônio artificial, esse processo de sinapse é simulado através da utilização de pesos. O sinal que é utilizado para processamento no neurônio artificial é chamado de NET, cujo valor é calculado através de uma multiplicação entre o sinal de entrada do neurônio e o peso da sinapse (CARVALHO et al, 2015).

Os valores dos pesos podem ser ajustados durante uma etapa de treinamento da rede, esse treinamento consiste em fazer modificações sucessivas nos pesos das entradas, em busca dos valores que possam estimular neurônios artificiais e fazer com que respondam mais adequadamente a determinado peso, tornando a rede capaz de reconhecer padrões (BRUMATTI, 2014).

Brumatti diz que:

> Através de um processo de treinamento, as redes neurais passam a ser capazes de responder a estímulos de entrada. Conforme o aprendizado, a rede se torna capaz de reconhecer padrões classificando então as entradas, ou seja, quando uma entrada é aplicada à rede, esta fornece uma resposta de saída indicando a classe a qual a entrada pertence. Numa outra forma de aprendizado a rede aprende sobre o relacionamento que há entre pares entrada-saída, assim quando uma entrada for aplicada à rede, esta responde com a saída correspondente àquela entrada. A habilidade de manipular dados imprecisos faz com que as redes neurais sejam extremamente eficazes em tarefas onde especialistas não estão à disposição ou um conjunto de regras não pode ser facilmente formulado (BRUMATTI, 2014, p. 1).

As redes neurais contam com diversos algoritmos para treinamento da rede, sendo que um dos mais utilizados é o *Backprogation,* devido à metodologia utilizada. Este algoritmo utiliza multi camadas para fazer o treinamento da rede. De acordo com Affonso:

> Primeiramente apresenta-se um padrão à camada de entrada da rede. Esse padrão é processado camada por camada até que a camada de saída forneça a resposta processada. A resposta é comparada com a resposta desejada e se estiver errada, o erro é calculado. Os valores são retropropagados da camada de saída para a camada de entrada e conforme isso acontece, os pesos são ajustados e o processamento é feito novamente, até que se obtenha a resposta desejada (AFFONSO et al, 2010, p.5).

O algoritmo *Backprogation* é capaz de resolver diversos problemas devido a sua metodologia de aprendizado. Devido a isso, foi criado o algoritmo *Multilayer Perceptron* (MLP), que utiliza frequentemente o *Backprogation* para treinar a rede, minimizando os erros (MORAIS, 2009). Tal algoritmo será detalhado a seguir.

2.7.2 Multilayer Perceptron

O MLP é uma rede neural artificial que utiliza várias camadas de neurônios artificiais divididos em 3 categorias diferentes: camadas de entrada, camadas ocultas (ficam entre as camadas de entrada e saída) e camadas de saída. O principal papel das MLP é servir como uma espécie de “supervisor”, onde sua função é fornecer as entradas à rede e comparar as saídas geradas com um resultado esperado e ajustar os pesos a partir dos erros gerados (MORAIS, 2009).

O erro de uma camada de saída é utilizado para ajustar o peso da camada oculta anterior. O erro calculado dessa camada anterior é utilizado para calcular o erro da camada anterior a essa, e assim sucessivamente. Esse processo é repetido inúmeras vezes, até que o erro da primeira camada seja ajustado, tentando encontrar o menor erro possível (CORDON e MULLER).

Segundo Brumatti:

> Dizemos que a rede neural "aprendeu" quando ela passa a reconhecer todas as entradas apresentadas durante a fase de treinamento. Esta é a tradução do aprendizado da rede neural, pois, havendo pelo menos um neurônio que represente uma determinada informação (um estímulo apresentado na entrada), sempre que este estímulo for apresentado a esta rede neural, aquele neurônio que foi treinado para representá-lo, automaticamente irá ser disparado, informando assim, qual o estímulo que foi apresentado para a rede neural. (BRUMATTI, 2014).

Ou seja, a partir do momento que a rede termina o treinamento ajustando seus pesos, e consegue reconhecer e classificar corretamente os dados do treinamento, pode-se dizer que a rede está treinada para ser usada na classificação de novos dados.

As pesquisas na área de inteligência artificial tem ganhado força nos últimos anos. Com a popularização da mineração de dados, é comum encontrar empresas que as utilizam com foco na criação de metodologias e ferramentas que auxiliam na tomada de decisões. Há hoje em dia ferramentas que disponibilizam diversas metodologias e algoritmos prontas para uso, como a Weka, uma ferramenta grátis disponibilizada pela Pentaho Corporation. Os detalhes desta ferramenta serão dados a seguir.

2.7.3 Weka

A *Waikato Environment for Knowledge Analysis* (Weka) é uma ferramenta de mineração de dados, disponibilizada e criada pela Universidade Waikato, que começou a ser criada em 1993, e que foi integrada a *Pentaho Corporation* em 2006 (WEKA, 2015).

Ao iniciar a ferramenta é possível escolher entre 4 aplicações diferentes:

- Explorer: disponibiliza formas de visualização e exploração de um conjunto de dados, além de disponibilizar ao usuário uma grande quantidade de metodologias de mineração de dados, como classificação, agrupamento (clusterização), associação e seleção de atributos, com uma infinidade de algoritmos distintos para uso em cada uma das metodologias disponíveis;
- Experimenter: esta aplicação tem como foco a avaliação de diferentes tipos de algoritmos, sendo possível definir diversas tarefas com diferentes grupos de dados para a execução, viabilizando uma comparação posterior entre a performance de cada algoritmo;
- KnowledgeFlow: disponibiliza ao usuário formas de análise de dados utilizando fluxogramas, arrastando e configurando uma rotina de execução, com as atividades desejadas;
- Simple CLI: proporciona ao usuário utilizar o weka utilizando linhas de comando, caso o usuário deseje (WEKA, 2015).

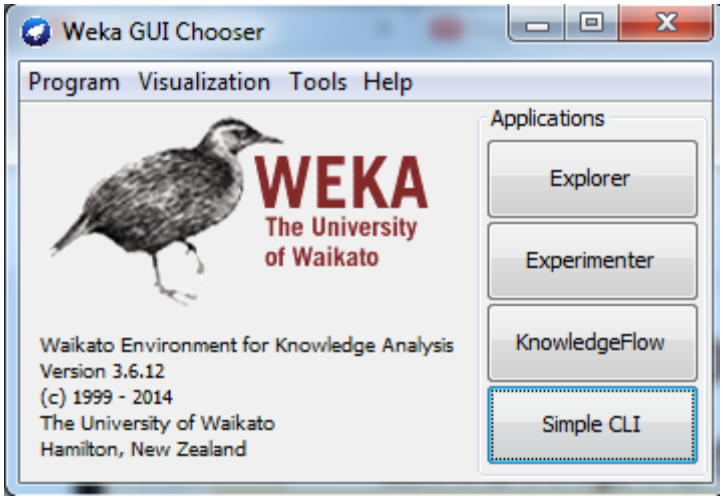

Figura 7: Aplicações disponibilizadas pelo Weka.
Fonte: Próprio autor

Ao selecionar a aplicação desejada, deve-se escolher o arquivo que contem os dados que serão utilizados, sendo possível escolher apenas uma fonte de dados por vez. Ao selecionar o arquivo, é possível visualizar diversas informações, como quantidade de linhas que possui o arquivo, quantidade de atributos e quantidade de elementos distribuídos em cada atributo. Além dessas informações, é possível utilizar outras opções, como edição do arquivo (WEKA, 2015).

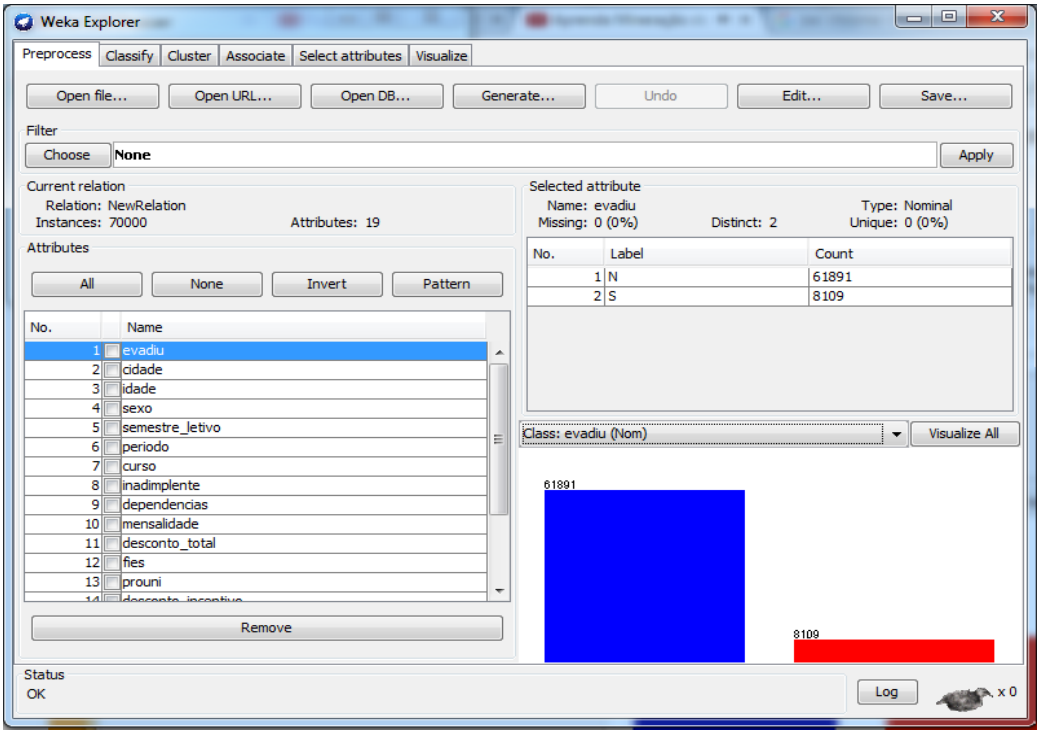

Figura 8: página principal da aplicação Explorer.
Fonte: Próprio autor

A ferramenta Weka utiliza um formato de arquivo próprio, como fonte de dados na extensão .arff. Esse arquivo é dividido em duas partes principais, uma delas contendo o cabeçalho, que possui as descrições dos atributos utilizados, e a outra contendo os dados, separados em colunas, sendo que cada coluna contém o dado referente a um dos atributos (WEKA, 2015).

Conhecendo os passos básicos para a mineração de dados tais como métodos e ferramentas, será possível definir se a utilização destas técnicas é viável para a criação de metodologias com a finalidade de minimizar da taxa de evasão de uma instituição de ensino.

A próxima seção descreverá a metodologia utilizada no estudo em questão em busca de resultados, com a finalidade de alcançar o objetivo citado acima.

3 METODOLOGIA

Quando um aluno efetua sua matrícula, não se sabe se ele virá a evadir. Devido a isso, atualmente diversos estudos são realizados na área, em busca de uma metodologia que possa ser capaz de prever a evasão de um aluno com o intuito de criar estratégias preventivas de evasão, com a finalidade de mantê-lo na instituição.

Considerando este fato, tomou-se como propósito deste trabalho a criação de uma metodologia/ferramenta capaz de prever a evasão de um aluno em uma instituição de ensino. Ou seja, uma ferramenta que seja capaz de prever se um aluno matriculado evadirá ou não, e quais as chances de um determinado aluno evadir, para que possam ser criadas metodologias preventivas de evasão, e assim possa reter o aluno na instituição de ensino.

Com a pesquisa bibliográfica feita, foi possível constatar que é viável a utilização de métodos computacionais para a mineração de dados acadêmicos, na busca de um padrão de evasão, e o quão importante é a criação de ferramentas que proporcionarão à instituição meios de identificar e minimizar a evasão do discente.

Com a finalidade de descobrir qual o padrão de evasão do aluno, elaborou-se uma pesquisa para definir quais métodos e ferramentas seriam utilizadas para a mineração dos dados acadêmicos, além de quais dados seriam utilizados para a análise em busca destes padrões. Durante a pesquisa foram encontrados alguns estudos já feitos na área, sendo que um destes estudos foi feito na própria instituição em 2009 pela aluna de ciência da computação Érika Raposa, intitulado “A utilização de técnicas de descobertas de conhecimento em ambiente acadêmico, aplicada ao problema de evasão escolar”.

Os trabalhos encontrados usavam diversas metodologias e algoritmos diferentes, como o feito por Raposa (2009), que utilizava o algoritmo de árvore J48. Santos (2014) utiliza o método de agrupamento e o algoritmo *Expectation Maximization* (EM). Outro trabalho encontrado, feito por Martinho (2014) também utilizava a classificação como método de mineração e em contrapartida ao trabalho executado por Raposa, Valquíria utilizou o algoritmo de redes neurais artificiais *ARTMAP-FUZZY.*

Para a construção deste trabalho, foi escolhido o algoritmo de redes neurais artificiais MLP. Escolheu-se este pelo fato de que algoritmos de redes neurais dão ao usuário a possibilidade de utilizar uma grande quantidade de entradas, pois o mesmo consegue tratar todas as entradas mais facilmente do que outros. Outro motivo foi o fato de que este algoritmo utiliza a retropropagação para treinar a rede, em busca de quais entradas são mais relevantes na decisão do aluno.

Para chegar ao resultado foram usadas ferramentas de mineração de dados, para analisar e encontrar padrões desconhecidos de evasão, que podem ser identificados apenas com a utilização de métodos computacionais, devido a grande massa de dados utilizada para o estudo de caso.

A seguir serão listadas as etapas que foram necessárias para o estudo e posteriormente serão explicadas mais detalhadamente.

As etapas utilizadas neste projeto foram:

a) Pré processamento: nesta etapa foram definidos quais dados seriam captados, tratados, organizados e armazenados no banco de dados criado para a mineração de dados;

b) Criação do *Data Warehouse*: foi criado um novo banco de dados para armazenar os dados pré processados na etapa anterior, facilitando assim a análise futura dos dados, este banco de dados continha dados de aproximadamente 90 mil alunos;

c) Mineração dos dados: nesta etapa foi utilizada a ferramenta *Open Source Weka*, que disponibiliza ao usuário diversos métodos de análise de dados, além de diversos algoritmos distintos para esta análise;

d) Entrada de dados: criação dos arquivos de entrada para análise dos dados. Foi criado um arquivo de entrada com aproximadamente 80% dos dados para treinamento do algoritmo, e após o treinamento o arquivo contendo o restante dos dados são utilizados para confirmação;

e) Análise dos dados utilizando redes neurais: foi utilizado um algoritmo de redes neurais para a mineração de dados. Para este trabalho foi utilizado o MLP, que utiliza *backpropagation* para treinamento da rede;

f) Resultado do processamento: são os resultados que foram obtidos ao final da análise dos dados.

3.1 PRÉ-PROCESSAMENTO DOS DADOS

A primeira etapa do trabalho é caracterizada pelo pré-processamento dos dados dos alunos, que estão armazenados nos bancos de dados disponibilizados pela instituição. O pré-processamento é uma das principais etapas, pois é nela que foi feita a captação, a organização, o tratamento e a preparação dos dados, para que posteriormente fossem migrados e armazenados em um novo banco de dados, que contém apenas os dados relevantes para o estudo, facilitando assim a análise e busca de informações relevantes que poderão ser usadas pela instituição.

Inicialmente pretendia-se utilizar os dados socioeconômicos dos alunos da instituição. Contudo, esses dados começaram a ser armazenados apenas a partir do ano de 2010. Outro fator que contribuiu para a não utilização de tais dados, é que apenas uma pequena parcela dos alunos matriculados a partir de 2010 e que fizeram o requerimento dos descontos cedidos pela instituição os possuem. Sendo que o percentual de alunos que se matricularam e encaixam-se nesta categoria, é de aproximadamente 30%. Dessa maneira, observou-se que sua utilização é inviável pelo fato que tão poucos alunos se encaixam em tais condições de uso dos dados socioeconômicos. Devido a isso um percentual muito grande da massa de dados armazenada deixaria de ser utilizada nos estudos, por isso optou-se pela não utilização dos mesmos.

O próximo passo foi definir qual o intervalo de tempo seria utilizado para captação e migração dos dados, sendo que no banco de dados disponibilizado é possível obter dados de 2004 em diante. O sistema ADX foi implantado nesta mesma época e devido a isso, foi feita a extração de 2006 em diante, pois já havia uma grande massa de dados armazenada e o período de teste de implantação já havia acabado.

Durante os estudos para a construção do referencial teórico foram encontrados alguns trabalhos já feitos na área, e analisando-os, notou-se que alguns dados eram usados com mais frequência para a análise, como nota (Raposa, 2009; Martins et al, 2012), sexo (Martinho at al, 2014; Santos at al, 2014; Junior, 2015), idade (Santos at al, 2014; Junior, 2015). Baseando-se nisto determinou-se que estes dados poderiam ser relevantes para os estudos, e então foram selecionados para a análise.

Logo após, foi feita uma análise detalhada dos bancos de dados que foram disponibilizados para estudo, para definir os dados que seriam utilizados na mineração e poderiam ser influentes na decisão de evasão do aluno. Com a nova análise, optou-se pela utilização de dados como: cidade, período, curso, inadimplências, dias inadimplentes, dependências, mensalidade, FIES (Fundo de Financiamento Estudantil), PROUNI (Programa Universidade para todos), desconto de incentivo, desconto total e um campo que informa se o aluno havia ou não evadido. Cada um destes dados será explicado de forma mais detalhada posteriormente.

Dados como CPF, nome, e-mail, entre outros, são atributos não valorados, devido a isso a rede neural não consegue utilizá-los para classificação. Estes atributos são utilizados apenas para ajudar na descrição do registro nos bancos de dados tradicionais, utilizados por setores operacionais das empresas. Em consequência disso, dados com essas características não foram utilizados para a mineração.

Após alguns testes iniciais os resultados obtidos não foram satisfatórios, tendo um percentual de acerto abaixo de 80%, constatou-se então que somente estes dados não seriam suficientes, tornando necessário refazer a análise em busca de novos dados, obtendo os seguintes dados: total de inadimplências, total de dias inadimplentes, inadimplência máxima e mínima, total de dependências e percentual de nota alcançada.

Após a inclusão destes novos dados, o percentual de acerto foi melhorado, chegando a ser superior a 90%. Esse percentual de acerto foi considerado aceitável, pois trabalhos encontrados na área possuíam um percentual de acerto que variam entre 87% e 96%, devido a isso os testes posteriores foram focados em tentar encontrar quais seriam os dados mais relevantes e quais poderiam ser descartados, por serem considerados pouco relevantes para o trabalho.

Durante estes testes notou-se que alguns dados possuíam pesos relativamente baixos, sendo inferiores a 0,1. Inadimplência mínima foi o dado que possui menor peso, sendo que o mesmo não chegou a 0,001. Devido a isso, optou-se por retirar os dados que tivessem os valores mais baixos.

A seguir cada um dos dados que foram selecionados, serão explicados de forma detalhada.

3.1.1 Evasão

Este dado é determinante para o estudo, pois é o critério utilizado para a classificação do aluno. É atribuído ao aluno um “sim” caso ele tenha evadido e “não” caso ele tenha concluído os estudos ou ainda esteja matriculado.

Para determinar se o aluno veio a evadir, foi analisado se ele efetuou a rematrícula no próximo semestre letivo. Se o aluno não rematriculou-se, foi considerado abandono e classificado como evadido, se o aluno trancou, transferiu ou cancelou a matrícula, é atribuído a ele o dado que o classifica como evadido. Caso ele não se encaixe nestas condições, ele é classificado como não evadido.

3.1.2 Cidade

Este dado informa se o aluno mora na cidade onde se situa o campus da instituição, e classifica o aluno em sim, se ele for da cidade, e não se for de outra cidade.

Este dado foi utilizado procurando saber se a distância que o aluno percorre para frequentar as aulas pode ser um dos fatores determinantes na decisão de evasão do aluno.

3.1.3 Idade

Foi utilizada a idade que o aluno no momento da matrícula. Este dado foi selecionado, pois espera-se que a idade possa ser um fator de grande relevância na decisão de evasão do aluno.

Para os estudos foram excluídos os alunos que possuem idade abaixo de 13 anos e acima de 100 anos, pois há grandes chances dos dados terem sido lançados de forma errada na matrícula do aluno.

3.1.4 Sexo

Dado que indica o sexo do aluno, caso seja feminino é classificado com ”F”, caso seja masculino, é classificado com “M”. Este dado é utilizado com a finalidade de identificar se o sexo pode influenciar na evasão do aluno.

3.1.5 Período

O período do curso na qual o aluno veio a se matricular para cursar o semestre letivo. Considerando que ao rematricular-se para um novo semestre, o aluno adquire dados diferentes, e portanto a sua probabilidade de evasão se altera. Espera-se definir que este dado possa identificar qual período o aluno é mais propenso à evasão.

3.1.6 Curso

Este dado informa em qual curso o aluno está matriculado, auxiliando na determinação de qual curso tem maior taxa de evasão.

3.1.7 Inadimplências

Este dado busca informar quantas mensalidades o aluno atrasou ou deixou de pagar durante o semestre em que se matriculou. Utilizou-se este dado com a finalidade de se definir o quanto o fator inadimplência foi determinante na evasão do aluno.

3.1.8 Total de inadimplências

Este campo informa o total de inadimplência do discente. É um somatório da quantidade de vezes que o aluno deixou de pagar ou atrasou uma de suas mensalidades, levando em consideração todos os boletos inadimplentes dos semestres anteriores em que o mesmo esteve matriculado.

Por exemplo, um aluno se matriculou no primeiro período, e veio a ter 2 mensalidades inadimplentes, posteriormente, ao se rematricular no semestre letivo seguinte, veio a atrasar outras duas mensalidades. O Campo Inadimplências indicaria o valor 2 nas duas ocasiões, mas o Total de inadimplências indicaria o valor 2 na primeira ocasião e 4 na segunda, totalizando todos os boletos inadimplentes do discente.

Espera-se que ao final, este campo possa ser um fator diferencial indicando que o discente que tenha um alto valor de inadimplências, esteja mais propenso à evasão.

3.1.9 Dias inadimplentes

Somatório dos dias inadimplentes de cada mensalidade que teve seu pagamento atrasado ou não efetuado pelo discente, durante o semestre letivo. Espera-se que o discente que tenha dificuldades de pagar suas mensalidades, chegando ao ponto de atrasá-las ou até mesmo deixando de pagá-las venha a ter eventuais dificuldades de terminar seus estudos.

3.1.10 Total de dias inadimplentes

É o Somatório dos dias inadimplentes de cada mensalidade que teve seu pagamento atrasado ou não efetuado, levando em consideração os semestres anteriores em que o discente esteve matriculado. Este dado utiliza a mesma

metodologia do total de inadimplências, mas somando o total de dias inadimplentes do aluno dos semestres anteriores.

3.1.11 Inadimplência máxima

Este campo considera apenas o maior tempo inadimplente de uma das mensalidades que teve seu pagamento atrasado, ou não efetuado no semestre em que o discente esteve ou está matriculado. Este tempo é referente aos dias inadimplentes.

3.1.12 Inadimplência mínima

Este campo considera apenas o menor tempo inadimplente de uma das mensalidades que seja diferente de 0 e que teve seu pagamento atrasado, ou não efetuado no semestre em que o discente esteve ou está matriculado. Este tempo é referente aos dias inadimplentes.

3.1.13 Dependências

Este campo se refere a quantidade de disciplinas em que o discente não obteve nota suficiente para ser aprovado no semestre em que esteve matriculado, sendo retido e obrigado a ter que refazer a disciplina em questão. Espera-se que o discente que tenha dificuldades de concluir as disciplinas do semestre letivo, venha a ter dificuldades para concluir o curso em que esteja matriculado, devido a isso acredita-se que este dado seja de relevância para a decisão de evasão.

3.1.14 Total de dependências

Este dado informa a soma de todas as inadimplências do discente, considerando os semestres anteriores em que ele esteve matriculado. A metodologia é semelhante a utilizada em Total de Inadimplências.

3.1.15 Valor da mensalidade

Este dado busca informar qual o valor da mensalidade do aluno, na tentativa de identificar se tal informação pode ser um fator relevante para a evasão do aluno. Mas devido ao fato de que anualmente todos os valores de mensalidade são reajustados, foi utilizada uma metodologia para generalizar os valores da mensalidade. Mensalidades abaixo do valor de 350 recebem o valor de -1, mensalidade que estão entre 350 e 700 são classificados como 0 e acima de 700 é atribuído o valor 1.

3.1.16 FIES

Este campo não informa o valor de desconto que o discente possui, ele apenas indica se o discente é beneficiado pelo desconto. Caso seja, é atribuído o valor 1 a este campo, caso contrário, é atribuído o valor 0.

O FIES é um desconto cedido ao aluno, mas que é repassado pelo governo federal mensalmente para a instituição. Este desconto é semelhante a um empréstimo, onde o aluno deve pagá-lo posteriormente. Diferente dos demais descontos, este ainda implica em uma divida ao discente. Ao selecionar este dado, espera-se identificar quanto o FIES pode influenciar na decisão do aluno de evasão.

3.1.17 PROUNI

Utiliza a mesma metodologia do FIES, apenas indicando se o aluno possui o desconto. O PROUNI é um desconto cedido ao aluno em troca de isenção de impostos a instituição, onde o discente não terá que arcar com custos posteriormente, mas possui vagas limitadas, sendo que apenas os alunos que possuíam as maiores notas no Enem e utilizaram essa nota concorrendo a vaga tem chances de se beneficiar com o desconto em questão.

3.1.18 Desconto de incentivo

Da mesma forma que os demais descontos apresentados anteriormente, este item recebe o valor 1, caso seja beneficiado pelo desconto e 0 caso não o possua. Este desconto é cedido pela própria instituição, como uma forma de incentivar o discente a pagar suas mensalidades antes do vencimento, dando um desconto extra em caso o pagamento seja efetuado.

3.1.19 Desconto total

É a soma dos percentuais de descontos que beneficiam o discente no semestre em que esteja matriculado, variando de 0 a um limite de 100%. O discente que possuir 0% de desconto estará sujeito a pagamento dos valores integrais de suas mensalidades, já o que tiver 100% de desconto estará isento do pagamento de suas mensalidades.

3.1.20 Inadimplência da primeira etapa

A Rede Doctum, antes do ano de 2009, utilizava uma metodologia de distribuição de notas que dividia o semestre letivo em duas etapas, com duração de aproximadamente 3 meses cada. Levando em consideração este fato, o campo em questão procurou informar quantas mensalidades o aluno deixou de pagar ou atrasou seu pagamento nos 3 primeiros meses de cada semestre letivo, sendo que estes três primeiros meses seriam equivalentes a primeira etapa de notas. Com isso, este dado busca informar se o aluno possuía dificuldades de pagamento das mensalidades nos primeiros meses de cada semestre letivo, caso sim, pode ser um indicativo de que os meses subsequentes poderiam ter o mesmo fim.

3.1.21 Dias inadimplentes na primeira etapa

Utiliza uma metodologia baseada na mesma lógica utilizada para a captação dos dados do campo anterior, mas informando a quantidade de dias inadimplentes do discente no semestre letivo em questão.

3.1.22 Nota semestral

Este campo informa qual a média da nota que o aluno obteve no semestre. Devido ao fato da Rede Doctum ter mudado a metodologia de notas a partir do ano de 2009, foi necessário criar uma estrutura capaz de generalizar as duas metodologias de notas utilizadas na instituição, tornando-as equivalentes.

A primeira metodologia utilizada era composta por duas etapas de distribuição de notas, contendo 60 e 40 pontos, onde o discente necessitava obter 60 pontos para estar apto à aprovação na disciplina. A segunda metodologia que é utilizada nos dias atuais, é composta por 3 etapas que distribuem 30, 40, 30 pontos

respectivamente, sendo necessário a obtenção de 70 pontos para a aprovação na disciplina.

Devido a isso, foi necessário a criação de um método que tornava as notas equivalentes, chegando ao seguinte resultado, caso a média das notas fosse suficiente para a aprovação do discente no semestre, é atribuído a este campo o valor 1, caso a nota seja suficiente apenas para deixa-lo apto a fazer a prova final, é atribuído a ele o valor 0, caso contrário, a nota seja insuficiente para os dois casos anteriores, é atribuído a ele o valor -1.

Com este campo espera-se obter informações sobre o quanto as notas podem ser decisivas na decisão de evasão do aluno, pois observa-se que discentes que possuem notas baixas estejam propensos a sair.

3.1.23 Percentual de nota alcançada

Da mesma forma que o campo anterior, para a obtenção deste dado foi necessário criar uma estrutura que fosse capaz de equivaler as notas das duas metodologias utilizadas pela instituição.

Este campo busca informar qual o percentual da nota lançada na primeira etapa, o aluno foi capaz de obter. A primeira etapa de notas da antiga metodologia distribuía 60 pontos, e a segunda 40 pontos. Um aluno que obtivesse 30 pontos dos 60 já lançados da primeira etapa na antiga metodologia, obteria 50% da nota lançada. Já na segunda, o aluno que obtivesse 24 pontos por exemplo, teria alcançado 80% da nota lançada na primeira etapa.

Com a média da nota alcançada na primeira etapa, é possível visualizar quais os alunos obtiveram melhor desempenho desde o início do semestre letivo.

Após o término da seleção de dados, que é a etapa de fundamental importância para o estudo em questão, foi dado início a etapa de pré processamento dos dados, que é caracterizada pela extração, tratamento e transformação dos mesmos, para a criação dos Data Warehouse específico para a aplicação. Esta etapa será explicada na sessão a seguir.

3.2 DATA WAREHOUSE

Nesta etapa se inicia o pré processamento dos dados, onde os dados são captados, tratados, transformados e armazenados em um novo banco, dando origem ao Data Warehouse usado para a mineração de dados.

Inicialmente foram criados apenas 4 data marts, contendo a maior a maior parte dos campos citados na sessão anterior e a partir destes foi possível a criação de um novo banco de dados, com dados mais complexos, como o somatório das inadimplências, dos dias inadimplentes e dependências totais do discente.

O custo computacional para obtenção destes dados em uma base dados tradicional seria muito alto, tornando necessária a criação de um novo banco que facilitasse a obtenção dos mesmos.

Ao final, obteve-se 5 bancos de dados, cada um contendo informações específicas, simplificando a obtenção dos dados que serão minerados nas próximas etapas. Os bancos criados foram os seguintes:

- dados_dos_alunos: este banco armazena dados pessoais do aluno, como código, idade, sexo, período e se o aluno evadiu.
- dados_economicos_dos_alunos: este armazena informações financeiras dos alunos, como descontos que o mesmo possui, inadimplências, e valor do boleto e tempo inadimplente.
- dados_economicos_dos_alunos_3_primeiros_boletos: este armazena os dados dos três primeiros boletos de cada semestre do aluno, como inadimplências e tempo inadimplente.
- dados_economicos_somatorio_semestral: este armazena a soma dos dados financeiros dos semestres anteriores de cada aluno, como total de inadimplências e de dias inadimplentes.
- dados_notas_dos_alunos: este banco armazena as notas, médias semestrais, média alcançada na primeira etapa de cada semestre e dependências do aluno.

Para o estudo optou-se pela criação de 5 bancos com tipos específicos de dados, como notas ou dados financeiros, pois seria mais fácil a análise em busca de possíveis erros para correção, tornando-os mais confiáveis.

Após a criação destes bancos de dados, iniciou-se a fase de captação extração e transformação dos dados utilizando a ferramenta disponibilizada pela *Pentaho Corporation*, denominada *Pentaho Data Integration* também conhecida como Kettle.

3.2.1 Pentaho Data Integration

Esta ferramenta tem se tornado cada vez mais comum no mercado, devido aos diversos recursos disponibilizados para captação, migração e transformação de dados, que podem ser feitas utilizando diversas fontes de dados, como arquivos de texto, Excel ou a captação direta em um banco de dados.

Com os componentes cedidos pela ferramenta, criou-se rotinas de execução, que tornaram possíveis a obtenção dos dados desejados. Os componentes utilizados nesta etapa foram:

- Conexões
- Input Tables ou Tabela de Entrada
- Output Tables ou Tabela de Saída
- Java Script
- Gerador de arquivo no formato .ARFF

Para a etapa de pré-processamento, foi necessário a utilização de apenas 3 componentes, sendo eles as conexões, Input Tables e Output Tables. Os demais foram utilizados nas fases finais do estudo. A utilização de cada um dos componentes será explicado a seguir.

3.2.1.1 Conexões com os bancos de dados

O primeiro passo a ser desta etapa, é a criação de conexões com os bancos de dados de onde será feita a extração. Para este trabalho, foram utilizados bancos de dados de sete cidades onde se situam sedes da Rede Doctum, sendo elas Cataguases, Caratinga, Juiz e Fora, Iuna, Leopoldina, Manhuaçu e Teófilo Otoni.

O critério utilizado para a escolha destas sete filiais foi o fato de quea partir de 2006 todas estas filiais possuíam dados que poderiam ser utilizados para a mineração e os dados são atualizados periodicamente até os dias atuais, além disso, algumas filiais só foram integradas à rede a partir do ano de 2010 e o processo de implantação do sistema demandou algum tempo para se estabilizar. Outro fato é que as filiais mais antigas possuíam mais dados armazenados para estudo.

Após a criação das conexões entre os bancos de dados que tiveram os dados extraídos e a ferramenta, iniciou-se a captação dos dados.

3.2.1.2 Transformações

Na mineração de dados, a transformação é a etapa onde é feita a captação, organização e preparação dos dados que são utilizados nas próximas etapas. Utilizando o Kettle, é gerado um arquivo .ktr. e esse arquivo é chamado de Transformação e contém uma rotina de execução feita pelo usuário. Nesse arquivo o usuário escolhe quais serão os recursos utilizados, seleciona, arrasta para a tela e liga os passos, sendo que a execução de um passo só acontece ao terminar o anterior.

Após criar as conexões, a ferramenta começou a ser utilizada de forma efetiva na extração dos dados citados anteriormente dos bancos disponibilizados. Nessa etapa foram criadas 5 transformações diferentes, e cada uma delas era responsável por executar um conjunto de passos responsáveis pela busca de um determinado grupo de dados, que compõe o *Data Warehouse* criado para o estudo, como dados do aluno, notas e financeiro.

Cada rotina de uma transformação iniciava-se definindo qual seria a fonte que forneceria os dados para extração. O recurso utilizado para essa etapa foi o *Table Input* - que será especificado mais detalhadamente à seguir - que se conecta a um banco de dados e utiliza consultas para retornar os dados desejados. Logo após, esses dados captados são repassados para a próxima etapa, que é a etapa de saída: armazenamento de dados, que capta os dados extraídos e os armazena no Data warehouse.

3.2.1.3 Table Input

A ferramenta *Pentaho Data Integration* disponibiliza para o usuário diversos recursos para captação de dados, mas para este trabalho foi necessário a utilização de apenas um, denominado *Table Input*. Este recurso utiliza as conexões feitas anteriormente buscando nos bancos de dados apenas os dados que o usuário deseja. Para cada uma das 7 sedes, utilizou-se um *Table Input*, responsável por captar os dados necessários de cada banco de dados e repassa-los para o *Table output*. O recurso que dá ao usuário a possibilidade de armazenar os dados obtidos é denominado *Table output*, e será explicado a seguir.

3.2.1.4 Table Output

Diferente do *Table Input*, que necessita de uma conexão com o banco de dados de onde os dados foram extraídos, para o uso do *Table output* é necessário estabelecer uma conexão com o banco de dados onde os dados serão armazenados. Após estabelecer a conexão com o banco de dados desejado, foi feita uma conexão entre todos os *Table Inputs* e o *Table output* para que os dados sejam migrados.

Como já foi dito, inicialmente foram obtidos os dados de 4 bancos criados para a mineração de dados e o quinto foi criado apenas após o termino de extração destes. Essa etapa será explicada a seguir.

3.2.2 Passos finais de criação do data warehouse

Apesar de terem sido usados 5 bancos de dados diferentes, inicialmente foram obtidos dados de apenas 4 bancos, sendo eles:

- dados_dos_alunos;
- dados_economicos_dos_alunos;

- dados_economicos_dos_alunos_3_primeiros_boletos;
- dados_notas_dos_alunos.

Após a criação dos 4 bancos iniciais, foi criada uma nova transformação que utilizava os demais bancos criados como fonte de dados, onde foram extraídos dados como somatório de inadimplências e dependências. O resultado gerado desta nova transformação foi então armazenado no quinto banco de dados. Após o término desta extração, o Data Warehouse foi concluído, possuindo todos os dados que foram levantados no primeiro ciclo e que foram assinalados como importantes para o estudo da mineração de dados institucionais.

3.3 MINERAÇÃO DE DADOS

A mineração de dados é uma metodologia que tem sido cada vez mais utilizada atualmente. Ela consiste em uma análise detalhada de uma grande massa de dados com o auxílio de ferramentas computacionais em busca de padrões ou informações que não poderiam ser encontradas através de métodos comuns.

Apesar de existirem diversas ferramentas para mineração de dados, como a Oracle Data Mining, IBM Intelligent Miner, entre outras, optou-se pela utilização da ferramenta Weka para a construção deste trabalho, pois é uma ferramenta grátis que disponibiliza todos os recursos necessários para o estudo, como metodologias de classificação, algoritmos de inteligência artificial e apresentação de resultados clara para o usuário.

Para este trabalho foi necessário a utilização da metodologia de classificação, que utiliza classes já definidas pelo usuário para classificar os dados de entrada. No caso, neste trabalho as classes utilizadas foram “Evadidos” e “Não evadidos”.

Para a análise dos dados foi utilizado o algoritmo de inteligência artificial Multilayer Perceptron, um algoritmo que utiliza a retropropagação para treinar a base de dados, atribuindo pesos a cada entrada ao final da análise de cada aluno. Esses pesos foram reajustados a cada dado analisado, tentando definir quais entradas teriam uma maior influência na decisão de evasão do aluno, e classificando-os

posteriormente. Cada etapa relacionada ao uso da ferramenta será explicada nas próximas seções.

3.3.1 Entrada de dados na ferramenta

A ferramenta *Weka* utiliza como padrão arquivos de texto na extensão .arff como fonte dos dados para mineração. Devido a isso, foi necessário armazenar os dados do *Data Warehouse* em um arquivo .arff. Para esta etapa tornou-se necessária a utilização da ferramenta *Kettle* novamente, pois a mesma disponibiliza um recurso de criação de arquivos nesta extensão como saída de dados.

Tornou-se necessário a criação de dois arquivos, um deles possuindo 80% dos dados armazenados no banco de dados que foram usados para treinamento e outro contendo os 20% dos dados restantes que foram utilizados para teste e confirmação, sendo que, os dados foram distribuídos de forma aleatória entre os dois arquivos.

O primeiro arquivo, que contem a maioria dos dados, foi utilizado para treinamento da base de dados em busca de quais dados possivelmente foram de maior influência na decisão de evasão. Ao terminar o treinamento obteve-se quais entradas tem maior peso, ou seja, mais influenciaram na evasão do aluno. Logo após o término da etapa de treinamento, os resultados obtidos foram aplicados no restante dos dados, sendo que a própria ferramenta se torna responsável pela classificação do discente, em evadido ou não evadido, baseando-se nas entradas do aluno e nos pesos obtidos de cada informação.

Na etapa de criação dos arquivos, foi necessária a utilização de dois *Table Input*, onde cada um captaria os dados que compõem os arquivos. O primeiro *Table Input* captou aproximadamente 80% dos dados e estes dados foram armazenados no arquivo .arff que foi utilizado para treinamento, totalizando dados de aproximadamente 70.000 alunos. O segundo *Table Input* captou dados de aproximadamente 19.000 alunos para armazena-los no arquivo para teste e confirmação dos resultados, mas antes de armazená-los no arquivo .arff. Foi necessário utilizar outro recurso da ferramenta *Kettle*, denominado *Modified Java Script Value*, que utiliza Java script para manipular os dados necessários para a

aplicação. O processo de criação dos arquivos que possuem dos dados utilizados na mineração de dados é evidenciado pela figura a seguir

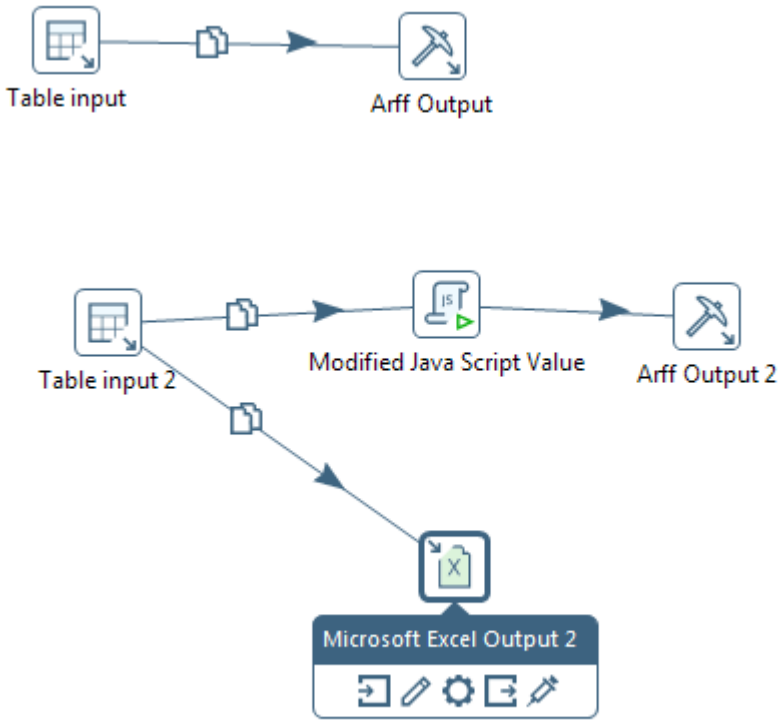

Figura 9: criação dos arquivos que contém os dados utilizados na mineração.
Fonte: Próprio autor

Para esse trabalho utilizou-se este recurso com o intuito de trocar o valor que define se o aluno havia ou não evadido por “?”, com o objetivo de esconder da ferramenta este dado, para que a mesma utilizasse os padrões encontrados durante o teste para classificar o aluno. Estes mesmos dados foram armazenados em uma planilha, mas contendo a informação de evasão do aluno, para que após o término da análise fosse feita uma comparação, identificando qual o percentual de acerto da informação.

3.4 RESULTADOS PRÉVIOS DA ANÁLISE

Ao terminar a análise dos dados, a ferramenta disponibiliza para o usuário algumas informações, como o percentual de chance de evasão, a qual classe o algoritmo o classificou, qual o percentual de erro de classificação dos dados utilizados para o treinamento, e quais entradas possuem o maior peso, ou seja, possivelmente mais influenciaram na decisão de evasão do aluno.

Ao final da análise, utilizando os dados disponibilizados pela ferramenta é possível fazer uma comparação sobre a taxa de acerto da mesma, analisando os dados de saída com os dados reais, que estão armazenados na planilha que contém os dados do discente, e informam se ele realmente evadiu ou não. A seguir será feita uma análise detalhada dos resultados.

4 RESULTADOS

Os resultados obtidos foram divididos em duas partes, sendo um deles referente ao treinamento da rede, com todos os dados disponíveis, e o outro a classificação do restante que tiveram seus dados de evasão omitidos.

Os primeiros resultados obtidos são os pesos de cada dado de entrada, informando através da análise quais foram provavelmente os dados mais influentes na decisão de evasão. Durante os primeiros testes, alguns dados apresentaram-se com um peso relativamente baixo, devido a isso optou-se por sua retirada. Estes dados foram:

- Dias inadimplentes
- Total de dias inadimplentes
- Inadimplência máxima e mínima
- Dias inadimplentes na primeira etapa

Logo após a retirada destes dados, iniciaram-se algumas análises mais detalhadas com os dados para treinamento da rede, os pesos obtidos durante a etapa de treinamento da rede são evidenciados na figura a seguir:

Classifier output

```
Sigmoid Node 301
    Inputs    Weights
    Threshold    -16.28382025966776
    Attrib cidade    7.9563740611757385
    Attrib idade    0.7329881964592608
    Attrib sexo    -8.817999457587069
    Attrib semestre_letivo    9.527032023039741
    Attrib periodo    13.322690509262214
    Attrib curso    22.18831374746157
    Attrib inadimplente    -10.668704285903972
    Attrib dependencias    -5.007990982744983
    Attrib mensalidade    18.024183164877503
    Attrib desconto_total    -1.6821180363689434
    Attrib fies    8.886355263982916
    Attrib prouni    9.126840887941913
    Attrib desconto_incentivo    -7.647390920591633
    Attrib inadim_pri_etapa    -5.514071371944424
    Attrib soma_inadimplente    8.355922599428334
    Attrib soma_dep    -4.261807824996183
    Attrib nota_pri_etapa    17.125067378964367
    Attrib media_semestral    13.904191434197681
```

Figura 10: Peso de cada entrada ao final da análise.
Fonte: Próprio autor.

Ao analisar as saídas, é possível observar que os dados que obtiveram maior peso foram:

- Curso que o discente estuda: 22
- O valor da mensalidade: 18
- Nota alcançada na primeira etapa: 17
- Média semestral: 13
- Período: 13

Apesar de ser esperado que o curso e o valor da mensalidade fossem fatores de grande influência, a nota da primeira etapa se mostrou tão relevante quanto, indicando que alunos que tem um bom desempenho no início de cada semestre tendem a uma menor evasão.

Outra informação importante que se pode extrair desta análise, é que a idade e o total de desconto do aluno tiveram um peso relativamente baixo, indicando que a idade e a quantidade de desconto do aluno podem ser pouco relevantes para a decisão de evasão do mesmo.

O segundo resultado que se pode extrair da ferramenta é o percentual de chance de evasão e a qual classe o algoritmo o classificou. Esta informação é altamente relevante, pois possibilita a criação de grupos de evasão, baseando-se na probabilidade de evasão dos discentes. Ao criar estes grupos, a instituição tem a possibilidade de realocação de recursos e criação de métodos de retenção, com foco na retenção dos discentes que estiverem nos grupos que apresentam os maiores riscos. Esta probabilidade pode ser consultada na coluna “probability distribution” da FIGURA 11:

Numeração dos registros	Valor Atual	Valor da previsão	Erro de predição	Probabilidade	
				Não evadir	Evadir
inst#	actual	predicted	error	probability distribution	
1	1:N	1:N		*0.58	0.42
2	2:S	1:N	+	*0.735	0.265
3	1:N	1:N		*0.945	0.055
4	1:N	1:N		*0.991	0.009
5	1:N	1:N		*0.963	0.037
6	1:N	1:N		*0.936	0.064
7	1:N	1:N		*0.92	0.08
8	1:N	1:N		*0.967	0.033
9	1:N	2:S	+	0.228	*0.772
10	1:N	1:N		*0.99	0.01
11	1:N	1:N		*0.934	0.066
12	1:N	1:N		*0.994	0.006
13	1:N	1:N		*0.982	0.018
14	1:N	1:N		*0.942	0.058
15	1:N	1:N		*0.993	0.007
16	1:N	1:N		*0.888	0.112
17	1:N	1:N		*0.991	0.009
18	1:N	1:N		*0.982	0.018
19	1:N	1:N		*0.992	0.008
20	1:N	1:N		*0.587	0.413
21	1:N	1:N		*0.58	0.42

Figura 11: Classificação e percentual de evasão.
Fonte: Próprio autor

Como é possível observar na figura acima, a coluna “actual” indica o valor de entrada do dado, ou seja, se o aluno evadiu o não. A terceira coluna, “predicted”, indica a qual classe o algoritmo classificou o discente; a coluna “error” indica em quais linhas houve erro de classificação, ou seja, quais alunos foram classificados de forma errada.

A coluna de “*probability distribution*” é subdividida em duas colunas e é a mais importante. A primeira coluna deste campo indica o percentual de chance de o aluno ser classificado como não evadido, e a outra o percentual de chance de classificação como evadido, sendo que esta probabilidade varia de 0 a 1. Como foi dito, esta informação possibilita a criação de grupos de probabilidade de evasão, sendo que os alunos que tiverem os maiores valores no primeiro campo serão os que têm a menor probabilidade de evadir, fato esse que pode ser comprovado analisando a segunda coluna, pois estes mesmos discentes terão o menor valor na coluna que indica a probabilidade de evasão. Em contrapartida, o discente que tiver

o maior valor na segunda coluna deste campo terá a maior probabilidade de evasão, possibilitando que seja feito um acompanhamento destes discentes.

O próximo campo que se tem acesso a uma informação relevante é o percentual de acerto na classificação do discente pela ferramenta.

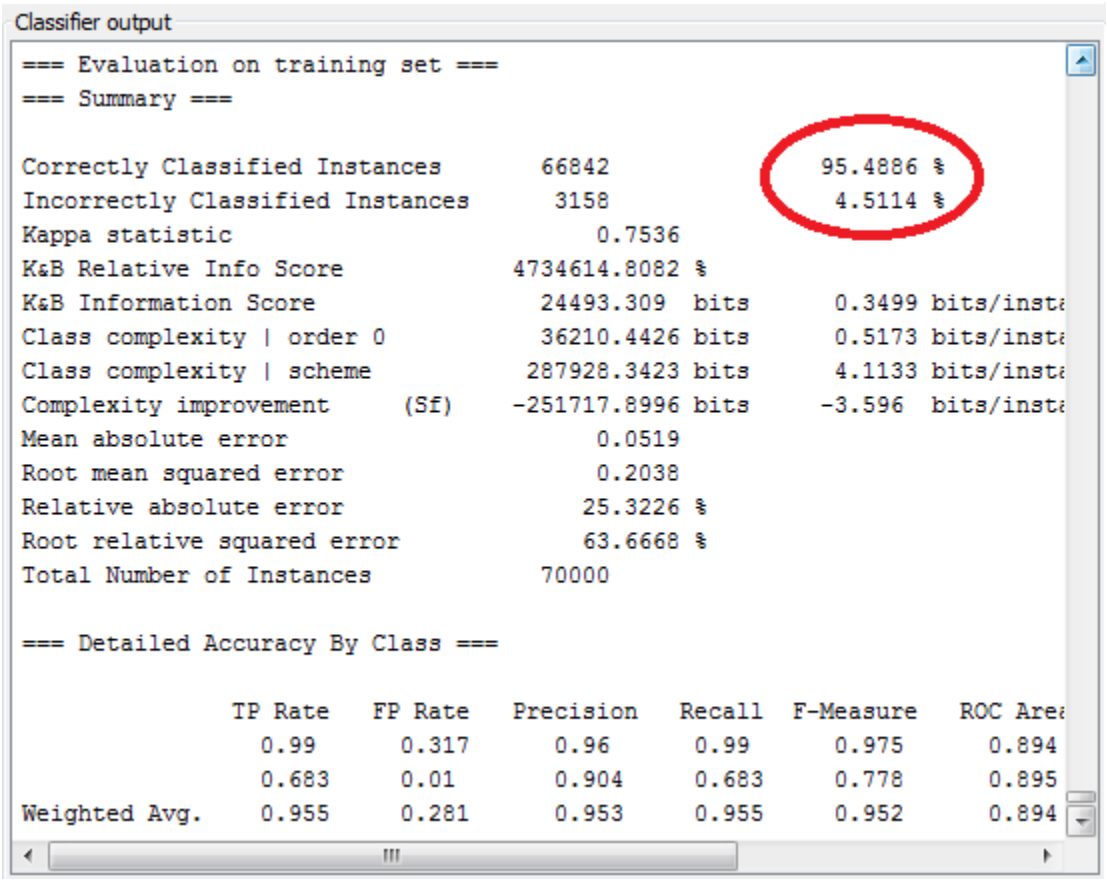

Figura 12: Percentual de acerto e erro de classificação.
Fonte: Próprio autor

A informação apresentada na imagem acima é de suma importância, pois é através dela que se define se os resultados do treinamento estão aptos para utilização. Caso seja considerado bom, a próxima etapa é a validação dos resultados através dos dados onde é omitido o dado de evasão do aluno.

A análise dos resultados deve ser cautelosa, pois apesar do alto percentual de acerto, outras informações devem ser levadas em consideração, sendo elas o falso positivo, onde aluno que não evadiu foi classificado como evadido, e o falso negativo, onde aluno evadido foi classificado como não evadido, que são os erros de classificação do algoritmo.

Dos 70000 alunos analisados, 8109 são evadidos e 61891 não evadidos. Ao analisar mais detalhadamente os resultados de classificação, observou-se que, apesar de um percentual de erro de apenas 5%, somente este resultado não é suficiente para afirmar que a metodologia seria eficiente para o uso prático.

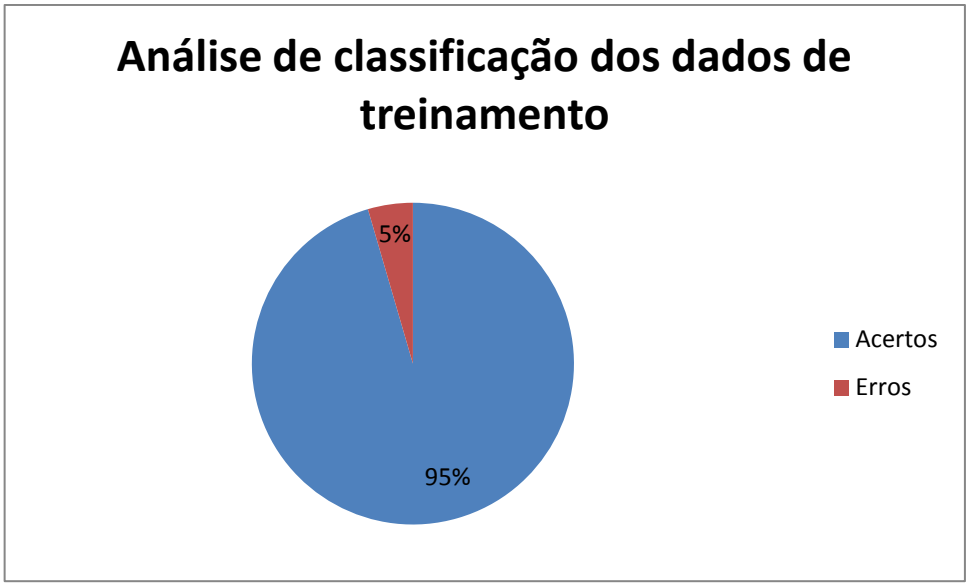

Gráfico 5: Total de acertos e erros de classificação.
Fonte: Próprio autor

Devido a isso, foi feita uma análise ainda mais detalhada sobre os erros e acertos de classificação da ferramenta, obtendo-se a seguinte tabela:

Total de discentes analisados	70000 alunos
Total de acertos	66842 alunos
Total de erros	3158 alunos

Quadro 1: Total de acertos e erros de classificação
Fonte: Próprio autor

Os erros e acertos podem ser distribuídos em 4 classes distintas, sendo elas:

- Discentes que não evadiram e foram classificados como não evadidos (verdadeiro negativo);
- Não evadidos classificados como evadidos (falso positivo);
- Evadidos classificados como não evadidos (falso negativo);
- Evadidos classificados como evadidos (verdadeiro positivo).

Ao detalhar os resultados utilizando estas classes, foi possível construir o gráfico abaixo:

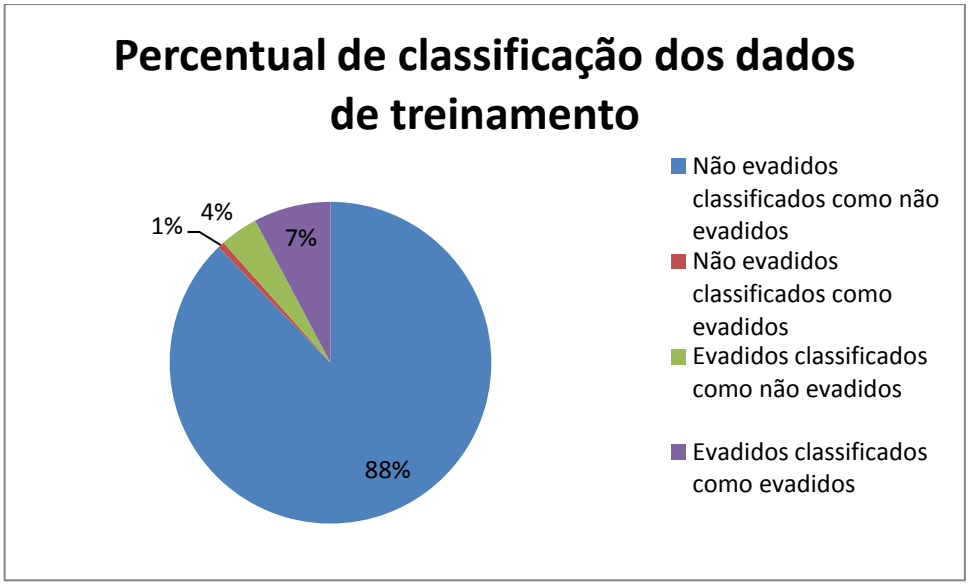

Gráfico 6: Percentual detalhado da classificação da ferramenta.
Fonte: Próprio autor

Analisando o gráfico, é possível notar que o percentual de erro no item de alunos não evadidos é mínimo, mas o percentual de erro no item de alunos evadidos classificados de forma errada é relativamente alto, sendo que aproximadamente um terço destes alunos foi classificado como não evadido.

Não evadidos classificados como não evadidos	61409 alunos
Não evadidos classificados como evadidos	482 alunos
Evadidos classificados como não evadidos	2676 alunos
Evadidos classificados como evadidos	5433 alunos

Quadro 2: Resultado da classificação de treinamento dos dados.
Fonte: Próprio autor

Apesar de um erro de 33% na classificação dos discentes evadidos, os resultados encontrados durante a etapa de treinamento foram considerados bons para dar continuidade a pesquisa, utilizando tais resultados para a classificação dos discentes que possuem o dado de evasão omitidos.

Ao término da classificação de aproximadamente 19000 discentes que tiveram o dado de evasão omitido, foram obtidos novos resultados. Nesta segunda etapa, o algoritmo não disponibiliza uma forma de consulta de percentual de acerto de forma automática, pois os dados de evasão não foram informados.

inst#,	actual,	predicted,	error,	probability distribution
1	?	1:N	+	*0.887 0.113
2	?	1:N	+	*0.986 0.014
3	?	1:N	+	*0.944 0.056
4	?	1:N	+	*0.956 0.044
5	?	1:N	+	*0.975 0.025
6	?	1:N	+	*0.988 0.012
7	?	1:N	+	*0.84 0.16
8	?	1:N	+	*0.926 0.074
9	?	1:N	+	*0.957 0.043
10	?	1:N	+	*0.967 0.033
11	?	1:N	+	*0.967 0.033
12	?	1:N	+	*0.993 0.007
13	?	1:N	+	*0.967 0.033
14	?	1:N	+	*0.8 0.2
15	?	1:N	+	*0.967 0.033
16	?	2:S	+	0.364 *0.636
17	?	1:N	+	*0.983 0.017
18	?	1:N	+	*0.99 0.01
19	?	1:N	+	*0.967 0.033
20	?	1:N	+	*0.983 0.017
21	?	1:N	+	*0.972 0.028

Figura 13: Conjunto de dados que omitem a evasão.
Fonte: Próprio autor

A figura acima demonstra que a análise deve ser feita separadamente, consultando a classificação feita na coluna “predicted” e os dados que foram salvos na planilha que foi criada em conjunto com o arquivo que se omite os dados.

Após o término da análise obteve-se os resultados evidenciados no gráfico 07:

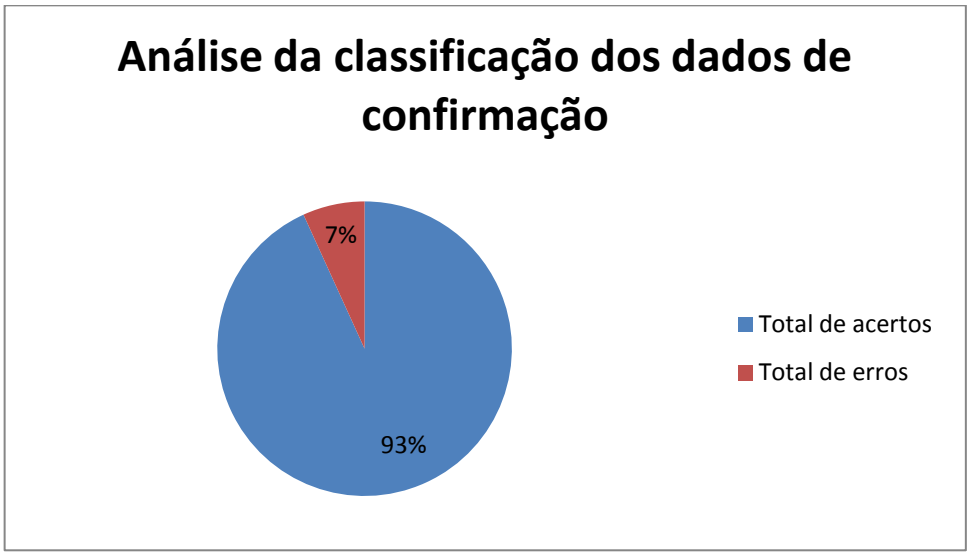

Gráfico 7: Percentual de classificação dos dados de confirmação.
Fonte: Próprio autor

Novamente o percentual de acerto foi relativamente alto sendo que ultrapassou a taxa de 90%. Os valores reais podem ser consultados na tabela 2:

Total de discentes analisados	19889 alunos
Total de acertos	18535 alunos
Total de erros	1354 alunos

Quadro 3: Resultado da classificação de confirmação dos dados.
Fonte: Próprio autor

Da mesma forma, não se deve considerar o resultado como bom apenas analisando o percentual de acerto, pois como foi apresentado, há outras variáveis que devem ser consideradas ao se detalhar os resultados. O percentual de classificação obtido nesta etapa pode se verificado no gráfico 8 e seus respectivos valores na tabela 4.

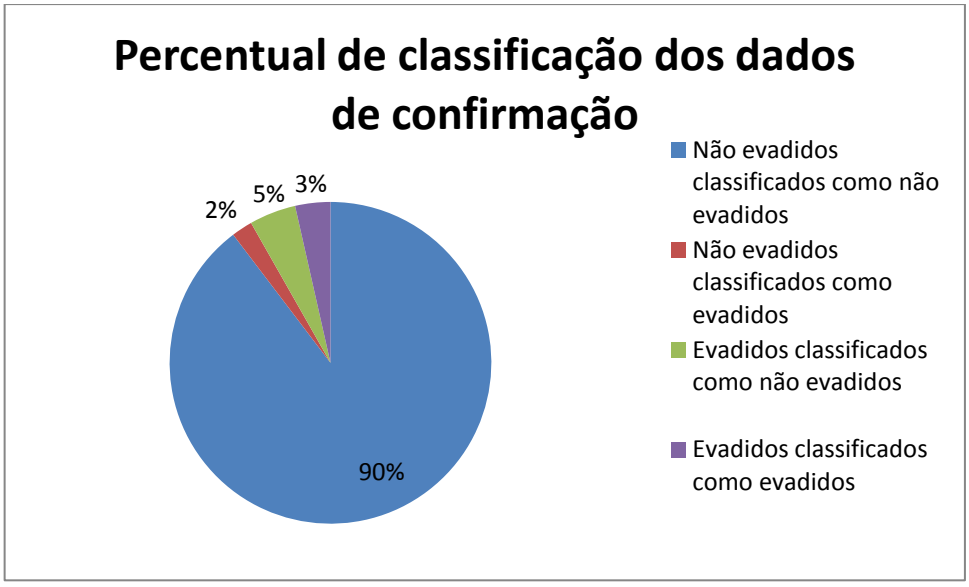

Gráfico 8: Percentual detalhado da classificação da ferramenta utilizando os dados de evasão omitidos.
Fonte: Próprio autor

Não evadidos classificados como não evadidos	17832 alunos
Não evadidos classificados como evadidos	422 alunos
Evadidos classificados como não evadidos	932 alunos
Evadidos classificados como evadidos	703 alunos

Quadro 4: Resultado da classificação que possui os dados de evasão omitidos.
Fonte: Próprio autor

O percentual de acerto considerando apenas os discentes que não vieram a evadir foi considerado satisfatório, pois a taxa de erro foi próxima de 2,3%. Mas a

taxa de erro de classificação de alunos que evadiram se mostrou um problema no uso desta metodologia, pois o percentual de erro está acima de 50%, mostrando que apesar do percentual total estar próxima de 93,3%, metade dos discentes que vieram a evadir não foram identificados de forma correta pela ferramenta.

CONCLUSÃO

Este trabalho aborda um problema recorrente em qualquer instituição de ensino: a evasão. Este fator está associado a diversas situações diferentes, tornando complexa a criação de medidas preventivas com o intuito de retenção do discente, proporcionando uma diminuição deste índice.

Tendo como base esta situação, o foco deste trabalho acadêmico foi a criação de um modelo baseado em redes neurais artificiais para mineração de dados, com a finalidade de usar os dados dos discentes que estão armazenados no banco de dados da instituição na tentativa de encontrar padrões de evasão discente, criando uma forma de minimizar este índice através de uma predição.

Segundo Rigo, at al (2012):

> Acredita-se que resultados nesta área de Mineração de Dados Educacionais possam apoiar efetivamente processos de detecção de comportamentos ligados à evasão escolar. Entretanto, destaca-se a importância de um amplo mapeamento de fatores associados, envolvendo os diversos setores das instituições, dado que modelos teóricos acerca da evasão escolar apontam para múltiplas causas, em diversas medidas interrelacionadas entre si (RIGO, et al, 2012, p.8).

Era esperado, portanto, que não fosse alcançado o percentual de 100% de acerto, pois as condições de evasão são várias, impossibilitando a predição ideal. Apesar disso, o índice de acerto geral da etapa de treinamento chegou a 95%, sendo que, ao analisar o percentual de acerto considerando apenas os alunos evadidos obteve-se uma taxa de acerto de 67%, indicando que esta metodologia poderia sim, ser válida para uso, mas ao se omitir os valores de evasão os resultados obtidos não mostraram-se tão promissores, indicando que seu uso poderia ser inviável para a instituição ao obter apenas 43% de acerto dos alunos que realmente evadiram.

Como o foco deste trabalho é identificar os alunos com alta probabilidade de evasão, para que seja possível monitorá-los e criar metodologias que os façam permanecer na instituição, torna-se inviável a sua utilização, pois menos da metade dos alunos que se encontram no grupo de risco são identificados de forma correta. Desta forma, os recursos que seriam alocados pela instituição com o intuito de minimizar a taxa de evasão, seriam aplicados à apenas uma parcela da população

de alunos a qual as metodologias se destinam. Outro fator agravante é o erro de classificação dos alunos que não vieram a evadir. Este erro implica na utilização de tais recursos à alunos que não evadiriam.

Durante a realização das pesquisas, foram raros os trabalhos encontrados nessa área que abordavam esse assunto de forma mais detalhada, dificultando a obtenção de parâmetros para comparação, com o objetivo de identificar qual metodologia seria a mais indicada para a mineração de dados educacionais.

Apesar dos resultados não alcançarem os valores ideais para uso, a ferramenta possibilitou a observação de alguns fatores que podem ser utilizados para uma análise prévia, como o peso de determinados dados. Os pesos da etapa de treinamento indicaram que a mensalidade, o curso e a nota da primeira etapa dos discentes podem ser os fatores mais relevantes para a decisão de evasão. Embora não seja este o resultado ideal, ele pode ser o primeiro passo para a criação de métodos de retenção discente.

Com a finalidade de se obter um resultado ainda melhor, é possível que sejam utilizados outros dados, como os dados socioeconômicos do discente, aumentando ainda mais o nível de detalhamento, mas no momento, devido ao pequeno percentual de alunos que possuem tais dados, este estudo torna-se pouco eficaz, podendo ser feito apenas em um futuro próximo.

Portanto, diante dos resultados alcançados, é possível observar que a mineração de dados institucionais pode ser uma poderosa ferramenta na busca por informações discentes, com foco na redução do índice de evasão das instituições de ensino, mas que deve ser feito um trabalho em conjunto com os diversos setores da mesma em busca de um melhor índice de acerto, tornando o resultado esperado ainda mais confiável.

TRABALHOS FUTUROS

Apesar do percentual de acerto relativamente alto, há algumas possibilidades de uma melhoria nos resultados, devido a isso uma das sugestões é a utilização de novos dados buscando um percentual de acerto maior, como os dados socioeconômicos do discente, uma análise detalhada por curso e uma análise da vida acadêmica do aluno do decorrer do curso.

Outra melhoria considerável é a possível implementação de uma ferramenta que viabilize o uso da ferramenta por usuários leigos no uso desta ferramenta, tornando-a mais acessível e facilitando seu uso.

Há também a possibilidade de utilização de outras metodologias de mineração de dados, e também uma possível comparação entre métodos, mostrando qual teria uma melhor eficiência na mineração dos dados da instituição.

REFERÊNCIAS

AFFONSO, E. T. F.; et al. Uso Redes Neurais Multilayer Perceptron (MLP) em Sistema de Bloqueio de Websites Baseado em Conteúdo. **Mecánica Computacional – Associação Argentina de Mecânica Computacional,** Buenos Aires, Argentina. Vol XXIX, págs. 9075-9090. 15-18 nov. 2010.

AGÊNCIA SENADO. Despreparo de alunos leva a evasão nos cursos superiores, alerta Cristovam. **Notícias Senado**. Brasília, DF; 9 set. 2015.

AMO, S. de. Técnicas de mineração de dados. In **Jornada de Atualização em Informática**, Uberlâncida. Faculdade de computação, Universidade Federal de Uberlândia. 2004.

AMORIM, T. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados**. Monografia (Bacharelado) Graduação em Ciência da Computação Centro de Informática. Universidade Federal de Pernambuco, 2006.

BARDAGI, M. P.; **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.

BARRA, F. **Redes neurais artificiais.** Programa de Educação Tutorial da Engenharia Civil (PET). Universidade de Juiz de Fora (UFJF). 05 jul. 2013.

BRASIL. DECRETO Nº 6.096, DE 24 DE ABRIL DE 2007.

BRUMATTI, M. **Redes Neurais Artificiais**. Departamento de Engenharia Elétrica - Universidade Federal do Espírito Santo (UFES). Vitória – ES, 2014.

CAMILO, C. O.; SILVA, J. C. da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Relatório Técnico (Graduação). Instituto de Indormática,universidaed Federal de Goiás. Goiânia, Ago. 2009.

CARDON, A.; MÜLLER, D N. **Introdução Às Redes Neurais Artificiais**. 1994. 31 f. Trabalho de Pós Graduação (Pós-Graduação em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul. Porto Alegre, 1994.

CARDON, A.; MÜLLER, D. N.; **Introdução Às Redes Neurais Artificiais.** Trabalho de conclusão de curso (Pós-Graduação em Ciência da Computação). Universidade Federal do Rio Grande do Sul. Instituto de Informática. Curso de Pós-Graduação em Ciência da Computação. Porto Alegre, 1994.

CARVALHO, E. C. C.; et al. **Redes neurais artificiais: essência e aplicações**. [S.I.]. 29 jul. 2015. Disponível em: <http://www.dropbug.com.br/redes-neurais-artificiais/>.

COMISSÃO ESPECIAL DE ESTUDOS SOBRE A EVASÃO NAS UNIVERSIDADES PÚBLICAS BRASILEIRAS**. Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas**. [S.l.] Out. 1996.

DOCTUM, REDE DE ENSINO. **História**. Potal Doctum. 2015. Disponível em: <http://www.doctum.edu.br:8080/portal/institucional/historia>. Acesso em: 08 out. 2015.

ELMASRI, R.; NAVATHE, S. R. **Sistemas de Banco de Dados**, 4ª ed., Pearson Addison Wesley, São Paulo-SP, 2005.

FÁTIMA, M.; FARIA, B.; FRANCO, A. L; Causas da evasão em curso de graduação a distância em administração em uma Universidade Pública Federal. **Rev. Teoria e Prática da Educação**, v. 14, n. 3, p. 43-56, set./dez. 2011.

FIALHO, M. G. D. , PRESTES, E. M. T. Evasão escolar no curso de pedagogia da UFPB: na compreensão dos gestores educacionais. **Rev do Mestrado Profissional Gestão em Organizações Aprendentes,** v.3, n.1, p. 42-63, João Pessoa, 2014.

GAIOSO, N. P. L.; **O fenômeno da evasão escolar na educação superior no Brasil.** Brasília, 2005. Dissertação de Mestrado – Universidade Católica de Brasília.
GOIS, A.; Metade dos universitários não se forma. **Folha de São Paulo**, São Paulo, 31 dez. 2006.

GOLDSCHMIDT, R.; PASSOS, E.; **Data Minig: conceitos, técnicas, ferramentas, orientações e aplicações**. Rio de Janeiro: Elsevier Editora Ltda, 2005. 261 p. INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), ASSESSORIA DE COMUNICAÇÃO SOCIAL DO. **Brasil teve mais de 7 milhões de matrículas no ano passado**. Portal Inep. Brasília, DF; 17 set. 2013.

INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), ASSESSORIA DE COMUNICAÇÃO SOCIAL DO. **Matrículas no ensino superior crescem 3,8%**. Portal Inep. Brasília, DF; 09 set. 2014.

JUNIOR, R. C., **Uso da mineração de dados na identificação de alunos com perfil de evasão do ensino superior**. Trabalho de Conclusão de Curso. 205. 136 f. (Bacharelado em Ciência da Computação). Universidade de Santa Cruz do Sul, Santa Cruz do Sul – RS. 2015.

MARTINHO, V. R. de C., **Sistema inteligente para a predição de grupo de risco de evasão discente**. 2014. 145 f. Tese (Doutourado em Engenharia Elétrica). Faculdade de Engenharia de Ilha Solteira – UNESP. Ilha Solteira - SP, 2014.

MARTINS, L. C., et al. **Um Assistente de Predição de Evasão aplicado a uma disciplina Introdutória do curso de Ciência da Computação.** In: ANAIS DO SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, (SBIE) 2012. Universidade do Vale do Itajaí (UNIVALI). Rio de Janeiro, 26-30 nov. 2012.

MARTINS, L. C.; LOPES, D. A.; RAABE, A.; **Um Assistente de Predição de Evasão aplicado a uma disciplina Introdutória do curso de Ciência da Computação**. 23º Simpósio Brasileiro de Informática na Educação. ISSN 2316-6533. 2012. Rio de Janeiro. Anais do SBIE, 2012, 26-30.

MORAES, S. A. S. de; SOUSA, P. de T. C.; **Sistemas de Bancos de Dados.** (Trabalho desenvolvido no Mestrado em Informática). Universidade Católica de Brasília UCB. Brasília, 2000.

MORAIS, P. F. T. B. de - 2009. **Estudo do uso da Informação Mútua na Seleção de Atributos para o Treinamento De Redes Neurais**. 2009. 51 f. Trabalho de Graduação. (Graduação em Ciência da Computação). Centro de Informática. Universidade Federal De Pernambuco (UFPE). Recife - PE, jun. de 2009.

NAPOLEÃO, J. F.; **Causas para a evasão dos alunos do curso de graduação a distância em ciências econômicas da Universidade Federal de Santa Catarina**. 2013. 215 f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Sócio Econômico. Programa de Pós-Graduação em Administração. Florianópolis, 09 dez. 2013.

OLIVEIRA, M.; **Data Warehouse.** [S.I.] 2009. Disponível em: <http://www.datawarehouse.inf.br/academicos/a%20publicar_data_warehouse_marcell_oliveira.pdf>. Acesso em: 15 out. 2015.

PRACIANO, E. **O que é um banco de dados relacional?**. [S.I.] 30 set. 2013 Disponível em: <http://elias.praciano.com/2013/09/o-que-e-um-banco-de-dados - relacional/>.

RAPOSA, É. O. B.; **A utilização de técnicas de descoberta de conhecimento em ambiente acadêmico, aplicada ao problema de evasão escolar.** 2009. 66 f. Dissertação – Faculdades Integradas de Caratinga, FIC/MG. Caratinga, MG – Brasil, 09 dez 2009.

REDE EDUCACIONAL DO GRUPO VIRTUOUS. **Células nervosas**. [S.I.]. 2008-2015 Disponível em: <http://www.sobiologia.com.br/conteudos/FisiologiaAnimal/nervoso2.php>.

RIGO, S. J.;et al. **Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades.** In: WORKSHOP DE DESAFIOS DA COMPUTAÇÃO APLICADA À EDUCAÇÃO. São Leopoldo – RS, 2012. Programa de Pós-Graduação em Computação Aplicada - Universidade do Vale do Rio do Sinos (UNISINOS). **Anais do...** p. 168-177. São Leopoldo, 2012.

RODRIGUES, M.; MORENO A. C.; **Matrículas no ensino superior sobem 3,8% e atingem 7,3 milhões de alunos.** Globo Comunicação e Participações, [S.I.] 09 set. 2014.

SALIBA, Nemre Adas et al. Organização curricular, evasão e repetência no curso de odontologia: um estudo longitudinal. **Revista de Odontologia da UNESP**, v. 35, n. 3, p. 209-214, 2006.

SANTOS, Á. P.; **A predição da evasão de estudantes de graduação como recurso de apoio fornecido por um assistente inteligente**. 2014. 69 f. (Dissertação de Mestrado). Universidade Católica de Brasília, Gestão do Conhecimento da tecnologia da Informação. Brasília, 2014.

SANTOS, V. V. dos; **Data warehouse: análise da performance de ferramentas de etl**. Universidade do Sul de Santa Catarina – Unisul. Florianópolis, 2013.

SCHMITT, J. et al. **Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo.** Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Ciência da Computação. Florianópolis – SC, 2005.

SOUZA, S. L**.; Evasão no ensino superior: um estudo utilizando a mineração de dados como ferramenta de gestão do conhecimento em um banco de dados referente à graduação de engenharia.** 2008. 114 f. (Tese de Doutourado). Universidade Federal do Rio de Janeiro, COPPE. Rio de Janeiro, 2008.

TAKAI, O. K.; et al.; **Introdução a banco de dados**. (Apostila educacional) Departamento de Ciência da Computação do *IME (Instituto de Matemática e estatística/USP (Universidade de São Paulo).* São Paulo, fev. 2005.

VARELLA, D.; **Corpo Humano: Neurônios**. 06 mai. 2014. Disponível em: <http://drauziovarella.com.br/corpo-humano/neuronios>.