CENTRO UNIVERSITÁRIO INTERNACIONAL UNINTER
MESTRADO E DOUTORADO PROFISSIONAL EM EDUCAÇÃO E NOVAS TECNOLOGIAS

MOACIR GOMES DA SILVA

PROPOSTA DE SIMULADOR COMPUTACIONAL COM RECURSOS DA MINERAÇÃO DE DADOS PARA PREDIÇÃO DE EVASÃO DISCENTE EM CURSOS DE ENGENHARIA – EAD: UM ESTUDO DE CASO

CURITIBA

2020

CENTRO UNIVERSITÁRIO INTERNACIONAL UNINTER
MESTRADO E DOUTORADO PROFISSIONAL EM EDUCAÇÃO E NOVAS TECNOLOGIAS

MOACIR GOMES DA SILVA

PROPOSTA DE SIMULADOR COMPUTACIONAL COM RECURSOS DA MINERAÇÃO DE DADOS PARA PREDIÇÃO DE EVASÃO DISCENTE EM CURSOS DE ENGENHARIA – EAD: UM ESTUDO DE CASO

CURITIBA

2020

MOACIR GOMES DA SILVA

PROPOSTA DE SIMULADOR COMPUTACIONAL COM RECURSOS DA MINERAÇÃO DE DADOS PARA PREDIÇÃO DE EVASÃO DISCENTE EM CURSOS DE ENGENHARIA – EAD: UM ESTUDO DE CASO

Dissertação apresentada ao Programa de Pós-Graduação – Mestrado e Doutorado Profissional em Educação e Novas Tecnologias, como parte dos requisitos necessários para obtenção do grau de Mestre em Educação e Novas Tecnologias.

Área de Concentração: Educação

Orientadores:

Profº.Dr. Mario Sergio Cunha Alencastro

Profª Drª. Siderly do C. Dahle de Almeida

CURITIBA

2020

S586p Silva, Moacir Gomes da
Proposta de simulador computacional com recursos da mineração de dados para predição de evasão discente em cursos de engenharia – EAD: um estudo de caso / Moacir Gomes da Silva. - Curitiba, 2020.
104 f. : il. (algumas color.)

Orientador: Prof. Dr. Mário Sérgio Cunha Alencastro
Orientadora: Profa. Dra. Siderly do Carmo Dahle de Almeida
Dissertação (Mestrado Profissional em Educação e Novas Tecnologias) – Centro Universitário Internacional UNINTER.

1. Evasão universitária. 2. Mineração de dados (Computação). 3. Simulador de predição de evasão. 4. Ensino à distância. 5. Tecnologia educacional. I. Título.

CDD 371.334

Catalogação na fonte: Vanda Fattori Dias - CRB-9/ 547

CENTRO UNIVERSITÁRIO INTERNACIONAL UNINTER
PRÓ-REITORIA DE PÓS-GRADUAÇÃO, PESQUISA E EXTENSÃO-PGPE
PROGRAMA DE MESTRADO E DOUTORADO PROFISSIONAL EM EDUCAÇÃO E NOVAS TECNOLOGIAS
Secretaria do Mestrado e Doutorado Profissional em Educação e Novas Tecnologias

Defesa Nº 039/2020

ATA DE DEFESA DE DISSERTAÇÃO PARA CONCESSÃO DO GRAU DE MESTRE EM EDUCAÇÃO E NOVAS TECNOLOGIAS

No dia 11 de dezembro de 2020, às 13h, reuniu-se via web conferência a Banca Examinadora designada pelo Colegiado do Programa de Mestrado e Doutorado Profissional em Educação e Novas Tecnologias, composta pelos professores doutores: Siderly do Carmo Dahle de Almeida (Presidente-Orientador-PPGENT/UNINTER), João Manuel Nunes Piedade (Integrante Externo/UNIVERSIDADE DE LISBOA), Nelson Pereira Castanheira (Integrante Interno Institucional/UNINTER), Ademir Aparecido Pinhelli Mendes (Integrante Interno Titular-PPGENT/UNINTER), Marcia Maria Fernandes de Oliveira (Integrante Interno Suplente-PPGENT/UNINTER), para julgamento da dissertação: "PROPOSTA DE SIMULADOR COMPUTACIONAL COM RECURSOS DA MINERAÇÃO DE DADOS PARA PREDIÇÃO DE EVASÃO DISCENTE EM CURSOS DE ENGENHARIA – EAD: UM ESTUDO DE CASO", do mestrando Moacir Gomes Da Silva. O presidente abriu a sessão apresentando os professores membros da banca, passando a palavra em seguida ao mestrando, lembrando-lhe de que teria até vinte minutos para expor oralmente o seu trabalho. Concluída a exposição, o candidato foi arguido oralmente pelos membros da banca.

Concluída a arguição, a Banca Examinadora reuniu-se e comunicou o Parecer Final de que o mestrando foi:

(x) APROVADO, devendo o candidato entregar a versão final no prazo máximo de 60 dias.

( ) APROVADO somente após satisfazer as exigências e, ou, recomendações propostas pela banca, no prazo fixado de 60 dias.

( ) REPROVADO.

O Presidente da Banca Examinadora declarou que o candidato foi aprovado e cumpriu todos os requisitos para obtenção do título de Mestre em Educação e Novas Tecnologias, devendo encaminhar à Coordenação, em até 60 dias, a contar desta data, a versão final da dissertação devidamente aprovada pelo professor orientador, no formato impresso e PDF, conforme procedimentos que serão encaminhados pela secretaria do Programa. Encerrada a sessão, lavrou-se a presente ata que vai assinada pela Banca Examinadora.

Recomendações: Dar continuidade a pesquisa em um doutorado e publicar.

Siderly do Carmo Dahle de Almeida
Presidente da Banca

Assinado por : **João Manuel Nunes Piedade**
Num. de Identificação: BI11601207
Data: 2020.12.17 20:22:18 +0000

João Manuel Nunes Piedade
Integrante Externo

Nelson Pereira Castanheira
Integrante Interno Institucional

Ademir Aparecido Pinhelli Mendes
Integrante Interno Titular

Marcia Maria Fernandes de Oliveira
Integrante Interno Suplente

Moacir Gomes Da Silva
Mestrando

Dedico este trabalho a duas pessoas especiais em minha vida, minha mãe e meu pai. Sempre estiveram comigo em todos os momentos da minha vida, apoiando e fazendo eu acreditar que eu posso.

AGRADECIMENTOS

Com certeza meus agradecimentos aos professores que tornaram minha vida especial, uma vida com muita alegria e sonhos realizados. Sei o quanto cada um deles se dedicou para a minha formação, desta forma, meus agradecimentos eternos a cada um.

Neste momento, o agradecimento especial, a dois professores que sempre estiveram ao meu lado nos momentos mais difíceis nesta caminhada: ao Professor Dr. Mario Sergio Cunha Alencastro que me ensinou o que é manter o foco nos estudos, e a Professora Drª Siderly do Carmo Dahle de Almeida como inspiração e admiração, pela sua dedicação e carinho com cada aluno, demonstrando o quanto podemos transformar a vida das pessoas pela educação.

O tempo é muito lento para os que esperam
Muito rápido para os que têm medo
Muito longo para os que lamentam
Muito curto para os que festejam
Mas, para os que amam, o tempo é eterno.

Henry Van Dyke

RESUMO

Esta pesquisa, desenvolvida em um programa que tem por área de concentração Educação e Novas Tecnologias, aborda o fenômeno da evasão na educação superior, especialmente na modalidade EAD, e tem por objetivo desenvolver um modelo computacional utilizando mineração de dados, para predizer os alunos em risco de evasão escolar dos Cursos de Engenharias e contribuir para o processo decisório das IES na mitigação deste fenômeno. Neste sentido, este estudo se justifica pela relevância do tema, o qual poderá propiciar benefícios tanto às instituições, que compreenderão por que seus alunos evadem, quanto aos próprios alunos, pois, em tempo hábil, a predição pode propiciar um alinhamento entre a IES e os alunos em busca de soluções para os problemas que levam à sua evasão. Metodologicamente, trata-se de uma pesquisa qualitativa, de natureza descritiva, exploratória e interpretativa, que fez uso de estudo bibliográfico e documental. Em um primeiro momento realizou-se um levantamento sobre o estado da arte da evasão escolar na educação superior no Brasil nos últimos dez anos e a aplicação da mineração de dados na área educacional. Na sequência, realizou-se estudos sobre os indicadores da evasão nos cursos de engenharia na modalidade à distância no Brasil. Por fim, desenvolveu-se um estudo de caso em cursos de engenharia de uma Instituição de Educação Superior com representação nacional. O produto obtido nos estudos foi o desenvolvimento de um simulador computacional preditivo de evasão discente, que permite as instituições de educação superior, de forma antecipada, no momento do ingresso do aluno na instituição, compreender o comportamento possível de sua evasão escolar e assim, definir suas estratégias de gestão. O simulador desenvolvido considerou a análise das taxas de evasão dos cursos de Engenharia Elétrica, Engenharia da Computação e Engenharia da Produção da Instituição de Educação Superior pesquisada, considerando nesta pesquisa o uso de oito variáveis que podem impactar nas taxas de evasão. O simulador obteve acurácia de 87,2%. O recurso utilizado no desenvolvimento do simulador preditivo de evasão discente foi o pacote de software do *WEKA.*

Palavras-chave: Evasão nas Engenharias, Simulador de predição de evasão. Mineração de dados.

ABSTRACT

This research, developed in a program that has as concentration Area Education and New Technologies, addresses the phenomenon of dropout in higher education, especially in the EAD modality, and aims to develop a computational model using data mining, to predict students at risk of school dropout of engineering courses and contribute to the decision-making process of HEI in mitigating this phenomenon. In this sense, this study is justified by the relevance of the theme, which can provide benefits both to institutions, which will understand why their students evade, and the students themselves, because, in a timely manner, prediction can provide an alignment between the HEI and students in search of solutions to the problems that lead to their evasion. Methodologically, this is a qualitative research, descriptive, exploratory and interpretative in nature, which made use of bibliographic and documentary study. At first, a survey was conducted on the state of the art of school dropout in higher education in Brazil in the last ten years and the application of data mining in the educational area. Next, studies were carried out on the indicators of evasion in the engineering courses in the distance modality in Brazil. Finally, a case study was developed in engineering courses of a Higher Education Institution with national representation. The product obtained in the studies was the development of a computational simulator predictive of student dropout, which allows higher education institutions, in advance, at the time of the student's entry into the institution, to understand the possible behavior of their school dropout and thus to define their management strategies. The developed simulator considered the analysis of the evasion rates of electrical engineering, computer engineering and production engineering courses of the Higher Education Institution researched, considering in this research the use of eight variables that can impact the evasion rates. The feature used in the development of the student dropout predictive simulator was the WEKA software package.

Keywords: Evasion in Engineering, Evasion Prediction Simulator. Data mining.

LISTA DE GRÁFICOS

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE TERMOS E ABREVIAÇÕES

ABED	Associação Brasileira de Educação a Distância
ARFF	*Atribute Relation File Format*
AVA	Ambiente Virtual de Aprendizagem
CSV	*Comma Separated Values*
EAD	Educação a Distância
EC	Engenharia da Computação
EDM	Mineração de Dados Educacionais (*Educational Data Mining - EDM*)
EL	Engenharia Elétrica
EP	Engenharia da Produção
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IES	Instituição de Ensino Superior
JAVA	*JavaScript*
KDD	Descoberta de Conhecimento em Dados *(Knowledge Discovery in Databases)*
MD	Mineração de Dados
MEC	Ministério da Educação
TIC	Tecnologias da Informação e Comunicação
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

A partir de 1990 o Brasil passou por uma importante expansão na educação superior, fruto de um período novo da democratização nacional, nova constituição (1988), novas políticas e leis na área da educação, desde a educação infantil até a superior, explicitas na LDB 9394/96. Este movimento promoveu a necessidade da especialização da área às novas normativas, assim como a capacitação de seus gestores frente aos novos desafios. As decisões estratégicas, que estão associadas ao futuro de uma organização, sua continuidade operacional, ganha relevância no processo decisório.

Com o crescimento das matrículas na educação superior nos últimos vinte anos, os desafios se multiplicaram, oportunidades surgiram, assim como os problemas. Um destes problemas é a evasão discente. Os gestores da área educacional, diante desta realidade, procuram tomar as melhores decisões para minimizar os impactos negativos para os alunos, sociedade e instituições que a evasão promove.

Associada a esta expansão, inúmeros estudos são realizados todos os anos e apresentados em congressos, fóruns, revistas acadêmicas, entre outros, para observar e discutir a permanência ou não dos estudantes nas Instituições de Educação Superior (IES). Destacam-se aqui que não somente estudos acadêmicos são delineados, mas estudos e soluções corporativos representados por consultorias, softwares, aplicativos para reduzir a evasão de alunos. Importante salientar que este problema impacta a instituição de ensino, que envida esforços para não perder alunos uma vez que isso se reflete como perda financeira, e impacta o próprio aluno, que investiu seu tempo, dinheiro e dedicou-se aos estudos e não conseguiu obter êxito em seus objetivos, necessitando, por algum motivo, abrir mão de seus sonhos.

Os processos decisórios, tanto das IES quanto de seus alunos, abalam, para o bem ou para o mal, o destino dos envolvidos. Estes processos decisórios têm níveis distintos dependendo das variáveis analisadas, variáveis estas associadas a dimensões culturais, econômicas, sociais, geográficas e pedagógicas.

A evasão discente é um dos fenômenos mais complexos na área da educação, envolvendo a interrupção de um ciclo de estudos, com impactos sociais e econômicos

negativos, não somente para o aluno, mas para toda uma sociedade. A complexidade surge da quantidade expressiva de variáveis determinantes deste fenômeno. Encontrar saídas para este problema traduz-se em contribuição para o desenvolvimento humano e uma sociedade mais inclusiva.

Ao olhar de forma específica para cada área das ciências e do conhecimento, observa-se que cada uma apresenta um propósito distinto. A Engenharia, por exemplo, em seu âmbito geral, contribui para o desenvolvimento de novas tecnologias em suas mais diversas áreas para a transformação econômica e social das pessoas. Nesse sentido, a evasão discente nos cursos de graduação em engenharia em suas mais diversas ênfases, não somente interrompe sonhos, mas o desenvolvimento de uma nação.

Poder predizer, ou seja, possibilitar prever os acontecimentos futuros com certa acurácia no campo da evasão na educação, pode contribuir também na análise complexa do conjunto de variáveis sobre a evasão discente dos alunos nos cursos de graduação em engenharias, por exemplo, empoderando os gestores de conhecimento para a tomada da melhor decisão no combate a este problema.

O pesquisador, que atua em cargo de gestão em uma instituição de educação superior e que atuou em outros níveis educacionais, em instituições públicas e privadas de ensino, trouxe, ao longo da jornada do mestrado, algumas questões que o nortearão para o desenvolvimento desta pesquisa. Entre estas questões, como encontrar informações estruturadas e de qualidade para o processo decisório a fim de mitigar a evasão escolar? Esta pergunta inicial, de acordo com o olhar do pesquisador, estimula o estudo de um fenômeno de impactos perversos em uma sociedade: a evasão discente.

Diversas pesquisas são realizadas, publicadas e apresentadas sobre este fenômeno, desde a educação infantil a educação superior, público e privada, principalmente nos últimos anos com o advento das tecnologias que proporcionam acesso e análise a informações antes inacessíveis.

Em geral estes estudos buscam uma essência central: como contribuir para mitigar este fenômeno?

Interessante nestes estudos a busca pelos fatores que impactam na redução ou aumento da evasão, ou seja, causa e efeito. Cada pesquisa aborda um conjunto

de variáveis do processo de causa e efeito, e, assim, como encontrar os alunos que estão propensos a evasão? Como identificar este aluno com informações estruturadas e acuradas? É possível encontrar respostas destas questões usando as tecnologias?

Pesquisar faz-se necessário para identificar se os recursos tecnológicos atuais como métodos, aplicativos, ferramentas de gestão, sistemas computacionais, podem evidenciar os alunos propensos a evasão de forma preventiva.

Para esta pesquisa, em sua delimitação investigatória, mesmo compreendendo que o sistema de ensino é composto de vários níveis e modalidades e que todos sofrem os impactos da evasão discente, fez-se necessário focar a investigação em um recorte temporal. Assim, o olhar do pesquisador recaiu sobre a educação superior, na modalidade à distância, em cursos de graduação em engenharia, buscando identificar um conjunto de variáveis (atributos) que impactam na evasão e que tornassem possível responder a seguinte pergunta de pesquisa:

Como conceber um simulador computacional preditivo, produto desta dissertação, utilizando mineração de dados, para predizer os alunos em risco de evasão escolar dos cursos de engenharia na modalidade de ensino à distância de uma Instituição de Educação Superior (IES)?

De modo a tornar possível encontrar resposta a este problema de pesquisa, o objetivo geral do estudo é propor um modelo computacional utilizando mineração de dados, para predizer os alunos em risco de evasão escolar dos Cursos de Engenharias na modalidade de educação à distância e contribuir para o processo decisório das IES na mitigação deste fenômeno.

Para o alcance do objetivo geral faz-se necessário delinear outros objetivos específicos, que são:

a) Realizar um levantamento bibliográfico sobre a evasão discente e suas possíveis causas, assim como caracterizar o problema da evasão na educação superior na modalidade à distância;
b) Analisar as informações do banco de dados recebidas da instituição de ensino pesquisada dos alunos dos cursos de graduação em Engenharia Elétrica, da Computação, da Produção na modalidade à distância, identificando as variáveis fornecidas;

c) Projetar o simulador de predição em uma arquitetura computacional capaz de identificar os alunos em risco de evasão utilizando como base o pacote de software de apoio do *WEKA*[1] *(Waikato Environment for Knowledge Analysis)*.
d) Realizar simulações e experimentos para medir a acurácia dos resultados encontrados e identificar as variáveis no estudo de caso que impactam na evasão.

Um grande pesquisador da área do planejamento estratégico, Mintzberg (2003), em sua obra intitulada “Criando Organizações Eficazes”, descreve a dinâmica das organizações nas dimensões da estrutura (organograma), configurações organizacionais, divisão do poder e suas complexidades. As organizações da área da educação para Mintzberg são descritas com maior nível de complexidade em sua gestão comparadas com outras organizações, pois são organizações com certo grau de autonomia em seus processos operacionais, dificultando em determinados momentos o processo decisório ao encontro de sua maior eficiência e eficácia, tanto no segmento público ou privado, em seus diversos contextos.

A educação no Brasil, frente aos seus diversos desafios históricos e presentes, precisa encontrar soluções viáveis neste cenário de complexidade para seus problemas, sejam de ordem política, reguladora, participação da sociedade, econômica, pedagógica e gestão. São os alunos, em suas diversas instâncias, do ensino infantil ao superior, que irão colher os frutos do equilíbrio desta complexidade.

Interessante quando observamos os objetivos da educação, por exemplo na educação superior:

> O ensino universitário visa a autonomia intelectual, e principalmente a formação de profissionais com a construção e disseminação do conhecimento, contribuindo para a formação do cidadão crítico, reflexivo e que saiba promover o desenvolvimento humano sustentável, com responsabilidade ecológica. (SAVIANI, 2013, p.34).

Formar um cidadão crítico, reflexivo e que saiba promover o desenvolvimento humano sustentável, para si e para uma sociedade, torna-se um enorme desafio. Mas

[1] *O WEKA,* conforme *Waikato* (2017) é um sistema que possui um conjunto de algoritmos de aprendizado de máquina para a execução de tarefas de mineração de dados. Ao longo do trabalho será melhor abordado.

quando o contexto social e econômico não permite que isto ocorra em sua plenitude, surgem efeitos colaterais, problemas preponderantes, e, um deles é a evasão escolar.

Quando o aluno não pode dar continuidade aos seus estudos perde a oportunidade da sua formação e isto acarreta danos humanos e sociais significativos com impacto relevante nos recursos concentrados para este fim.

Quando dos projetos e estudos de viabilidade econômica e financeira, que buscam demonstrar o investimento na operação das organizações escolares, público ou privado, o aluno é o objeto do estudo, sua formação é objetivo principal da organização escolar. A previsão da quantidade de alunos, uma das variáveis mais importante nestes projetos de viabilidade econômica e financeira, determina os recursos necessários, sejam humanos, tecnológicos e financeiros.

Neste sentido, esta dissertação se justifica, pois, a evasão escolar é uma distorção neste processo, por mais que a mesma possa ser prevista, ou seja, predizer a quantidade de alunos que irão evadir no momento da elaboração do plano de viabilidade econômica e financeira, da oferta de um novo curso, novos esforços se fazem necessários para reduzir os impactos da evasão prevista. Mas qual valor utilizar para predizer a quantidade de alunos que irão evadir em todo o processo? Qual a fonte de informação a ser utilizada? Qual a qualidade desta informação? E se a evasão for maior que a prevista nos estudos?

Ao chegar neste ponto, constatamos a necessidade de encontrar maneiras, aprimorar modelos, estimular pesquisas, desenvolver instrumentos de avaliação e de predição, aperfeiçoar a gestão desde o planejamento até o acompanhamento e atingir se possível, sua máxima eficiência e eficácia para a solução deste problema, ou seja:

> Com o ambiente organizacional mudando constantemente, fazer escolhas certas com base em critérios adequados e alinhados com o direcionamento estratégico, torna-se um fator crítico de sucesso ou até sobrevivência das organizações. Um dos principais desafios das organizações é tomar decisões certas, consistentes e alinhadas com seus objetivos estratégicos em uma situação específica (TRIANTAPHYLLOU, 2002, p. 23).

Nesse contexto, conforme Martinho (2014, p. 67):

> Diante da complexidade do fenômeno e da necessidade de encontrar soluções, é imprescindível realizar estudos sistemáticos, observar os sinais de evasão iminente, desenvolver e implementar estratégias para favorecer a identificação precoce dos estudantes propensos à

> evasão. Isso, com o intuito de possibilitar a articulação de um conjunto de medidas e ações proativas no sentido de reverter as intenções de abandono, melhorar o sistema educacional e mitigar o fenômeno da evasão, possibilitando a manutenção do estudante na instituição.

O uso adequado das tecnologias neste momento pode ser fundamental para a permanência do aluno na instituição.

Nos últimos anos a transformação da educação em todos os níveis e modalidades, influenciados pelo uso das novas tecnologias da informação e comunicação (TIC) é notória. No campo da educação à distância o surgimento e avanço tecnológico das plataformas digitais de aprendizagem e diversos outros recursos permitem aos alunos acesso com maior flexibilidade temporal e espacial, diferentes tipos de interação e compartilhamento de informações ofertados por entidades públicas e privadas.

Acompanhando este movimento tecnológico no segmento da educação, a geração de conteúdo, de informações e dados computacionais são expressivos todos os dias.

O processo de descoberta do conhecimento em banco de dados computacionais (*Knowledge Discovery in Databases - KDD*) citados por Fayyad *et al.* (1996) contribui na compreensão de um potencial estratégico em informações que podem tornar-se útil ao processo decisório. Encontrar modelos que possam analisar e tornar as informações compreensíveis e úteis torna-se um grande desafio, porque estas novas modelagens informacionais precisam trazer algum benefício novo, agregar valor para uma rápida e assertiva tomada de decisão.

A área da educação não poderia ser diferente neste movimento para compreender a sua realidade diária de enfrentamentos operacionais e estratégicos para a sua continuidade e transformação dos seres humanos que passam por seus sistemas de ensino.

Observa-se grandes esforços das organizações do segmento educacional na captação de alunos para justificar seus fins, com ou sem fins lucrativos, público ou privado. Porém em muitas situações não temos a mesma preocupação em reter este aluno, ou seja, temos grandes investimentos sendo canalizados para a captação do aluno e parte deste investimento se perde no processo de evasão (FAZOLIN,2018).

Considerando-se todo o contexto ora apresentado, o objeto de pesquisa desta dissertação é o desenvolvimento de um simulador computacional preditivo de evasão com base na mineração de dados com auxílio do pacote de softwares do *WEKA* que possibilita a predição de evasão dos alunos das Engenharia Elétrica, Computação e Produção de uma IES objeto desta pesquisa.

Conceber um simulador computacional de predição da evasão discente, utilizando tecnologias computacionais e de mineração de dados, com acurácia significativa, torna-se um poderoso instrumento de gestão na compreensão da evasão discente e contribui no processo decisório para a mitigação deste fenômeno.

Esta pesquisa está delineada da seguinte forma:

No capítulo 1, destaca-se o delineamento desta pesquisa, iniciando pela contextualização do problema da evasão no Brasil, o problema e a proposta da solução, os objetivos, a justificativa da pesquisa e o objeto de estudo.

No capítulo 2, inicia-se pela fundamentação bibliográfica sobre o tema da evasão escolar no Brasil no ensino superior, assim como a mineração de dados, destacando os métodos propostos para predição da evasão e um resumo dos modelos encontrados em outras pesquisas. Os autores que embasam este capítulo são Tontini e Walter (2014); Sepúlvida (2018); Cunha e Morosini (2014); Silva Filho (2007); Lima e Zago (2016): Gottardo *et al*. (2012) e Romero e Ventura (2013), entre outros.

No capítulo 3, apresentam-se os aspectos metodológicos, fundamentados em autores como Yin (2015); Bogdam *et al*. (1994) e no modelo de evasão proposto por Tinto (1975).

No capítulo 4, ocorre a descrição da implementação do simulador computacional para predição da evasão com auxílio do pacote de software do *WEKA (Waikato Environment for Knowledge Analysis)*, considerando-se as etapas do desenvolvimento da base de dados, programação, treinamento e diagnóstico da evasão. Neste capítulo, foram utilizadas as contribuições de autores como Silva (2019); Waikato (2017); Hall *et al.* (2009) Bouckaert *et al.* (2010); Martinho (2014); Manhães (2015) e Sepúlvida (2016).

No quinto capítulo, considerando-se o percurso trilhado a análise dos dados, evidenciam-se os resultados obtidos com o simulador computacional de predição.

Finalizando, no capítulo sexto, apresentam-se as considerações finais acerca do estudo desenvolvido, descrevem-se as conclusões, tendo em vista o aporte teórico adotado, e as contribuições desta pesquisa e as possíveis sugestões e indicações para pesquisas futuras.

2 EVASÃO NA EDUCAÇÃO SUPERIOR

Este capítulo apresenta os resultados da pesquisa bibliográfica realizada que constitui a base deste trabalho de investigação. Aborda o tema da evasão, a evasão na educação superior e na educação à distância, os fatores que promovem e reduzem a evasão e as tendências. Sobre os aspectos da tecnologia aborda o tema da mineração de dados, algoritmos, sistema de predição e aplicações computacionais.

2.1 DEFINIÇÃO DE EVASÃO

Dentre diversos temas ao entorno da temática educação, um merece destaque pela complexidade de compreensão, devido a quantidade de variáveis que o definem, a evasão.

A evasão estudantil, desde a educação infantil à educação superior, seja presencial ou à distância, desafia seus gestores, tanto em organizações públicas como privadas e, reduzir, torna-se um grande desafio. A evasão consome esforços e recursos diversos que as vezes são escassos.

O Ministério da Educação, em estudos durante as décadas de 70 e 80 demostrou, através das políticas educacionais um aumento significativo das oportunidades de escolarização na educação básica. Todavia, os altos índices de repetência e evasão apontam problemas que evidenciam a grande insatisfação com o trabalho realizado pela escola (BRASIL, 1997).

A complexidade que envolve este antigo problema atinge indistintamente todas as instituições de ensino. As pesquisas sobre o tema são constantes e o governo continua o acompanhamento acerca do assunto. A definição de evasão pode ser compreendida sobre três eixos (BRASIL, 1997):

1. Evasão de curso - quando o estudante se desliga do curso em situações diversas: abandono (deixa de se matricular), desistência (oficial), transferência (mudança de curso) ou exclusão por norma institucional;
2. Evasão da instituição - quando o estudante se desliga da instituição na qual está matriculado;

3. Evasão do sistema - quanto o estudante abandona de forma definitiva ou temporária o ensino superior.

Conforme o dicionário DICIO (dicionário online de português), evasão significa:

> Ação de abandonar algo; desistência, abandono, evasão escolar. Ação de escapar da prisão ou do local em que se estava preso, fuga. Ação de argumentar de modo vago, de utilizar pretextos para evitar uma resposta objetiva; evasiva. Deslocação de um lugar para outra; saída: evasão de dólares. Ação ou efeito de evadir.

Os termos na língua inglesa se referem a *scholl dropout* e *abandonment*, que traduzido significa: abandono, desistência, renúncia, evasão escolar.

Este termo causa preocupação para professores, gestores, organizações e toda comunidade.

De acordo com Tontini e Walter (2014), sob a ótica da educação, evasão é a ação de abandonar ou desistir de alguma atividade-fim da universidade, podendo existir evasão no ensino, na pesquisa e na extensão.

Outras definições de evasão são contempladas no trabalho de pesquisa de Sepúlvida (2018), que elaborou um quadro com a síntese de vários pesquisadores, no Quadro 1 é possível contemplar está síntese:

QUADRO 1 - DEFINIÇÃO DE EVASÃO

Autor	**Definição**
Gaioso (2005)	Interrupção no ciclo de estudos, em qualquer nível de ensino.
Brasil/MEC (2010)	Saída definitiva do estudante de seu curso de origem, sem concluí-lo
Kira (1998)	Perda ou fuga de estudantes da universidade.
Baggi (2011)	Saída do estudante da instituição antes da conclusão de seu curso.
Tinto (1975 apud SANTOS 2014)	Abandono voluntário (não motivado pelo fracasso escolar) e permanente.
Martins (2007)	Saída do estudante de uma IES ou de um de seus cursos de forma temporária ou definitiva por qualquer motivo, exceto a diplomação.
Laguardia e Portela (2009)	Saída do estudante de um curso ou programa educacional sem tê-lo completado com sucesso.
Favero (2006)	A evasão se caracteriza pela desistência do curso pelos estudantes.
Santos et al (2008)	**A evasão refere-se à desistência definitiva do estudante em qualquer etapa do curso.**

Fonte: Sepúlvida (2018)

Com relação a evasão no ensino superior Cunha e Morosini (2014, p.83) destaca:

> O estudo da evasão/abandono escolar tem se constituído, no âmbito da educação superior, numa temática “nova” e instigante, que tem conduzido diferentes estudiosos a enveredarem na busca de maiores informações e de dados consistentes que possam subsidiar de forma particular ou coletiva (nesse caso, as instituições de ensino) a adotar estratégias que busquem minimizar os efeitos danosos que o fenômeno causa tanto para os estudantes como para as instituições.

Na pesquisa de Silva Filho et al (2007, p.642), o autor destaca que:

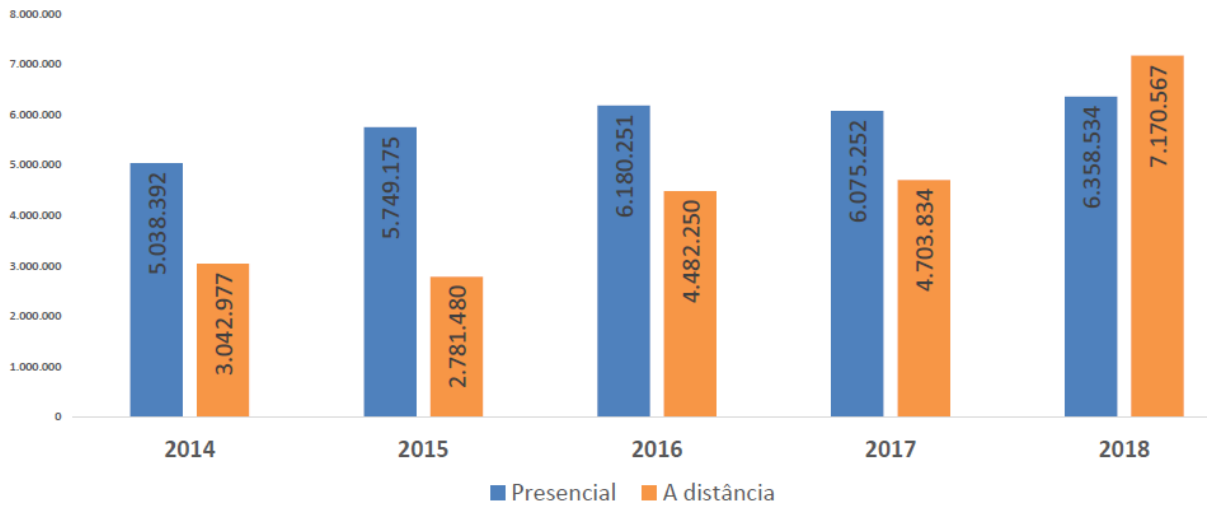
> A evasão estudantil no ensino superior é um problema internacional que afeta o resultado dos sistemas educacionais. As perdas de estudantes que iniciam, mas não terminam seus cursos são desperdícios sociais, acadêmicos e econômicos. No setor público, são recursos públicos investidos sem o devido retorno. No setor privado, é uma importante perda de receitas. Em ambos os casos, a evasão é uma fonte de ociosidade de professores, funcionários, equipamentos e espaço físico.

Neste sentido, observa-se que muitas são as razões pelas quais a evasão merece destaque em pesquisas, na busca por apontar caminhos que favoreçam a sua redução.

2.2 EVASÃO NO BRASIL

Conforme as bases do INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) em 2018 no Brasil foram ofertadas 13.529.101 vagas no ensino superior no Brasil, no Gráfico 1 a representação das vagas ofertadas de 2014 a 2018 por modalidade de ensino:

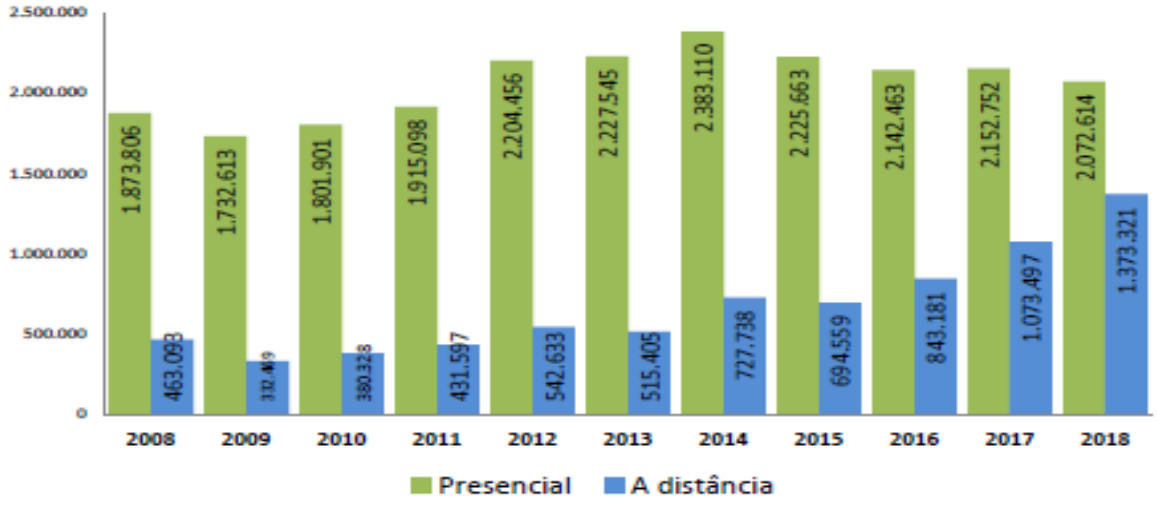
GRÁFICO 1 – NÚMERO DE VAGAS OFERTADAS EM CURSOS DE GRADUAÇÃO POR MODALIDADE DE ENSINO 2014 A 2018



Fonte: Censo da Educação Superior 2018

Entretanto observa-se que as vagas ocupadas ano a ano são bem menores que as vagas ofertadas, no Gráfico 2, a seguir, observa-se os números absolutos:

GRÁFICO 2 - NÚMERO DE INGRESSOS EM CURSOS DE GRADUAÇÃO – 2008 A 2018

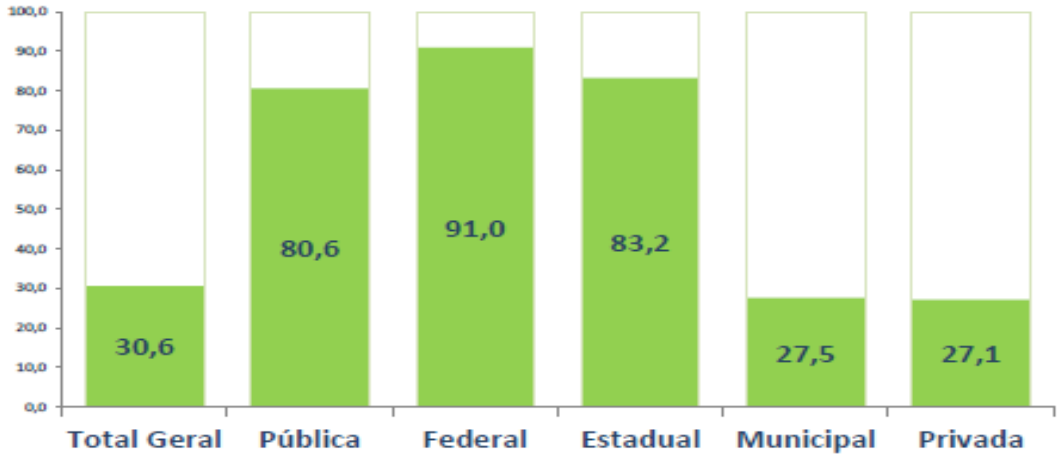


Fonte: Censo da Educação Superior 2018

O crescimento dos alunos ingressantes em cursos de graduação ano a ano na modalidade à distância começa a ter representatividade significativa e assim como as pesquisas de evasão na modalidade presencial, novas pesquisas precisam ser realizadas para compreender este fenômeno na modalidade à distância.

O INEP informa também que a taxa média de ocupação das vagas ofertadas no ano de 2018 foi de aproximadamente trinta por cento (30,6%), conforme Gráfico 3 uma taxa preocupante, demonstrando que após as matrículas muitos destes alunos irão evadir.

GRÁFICO 3 - % DE VAGAS OCUPADAS POR ESFERA DE ENSINO – ANO DE 2018

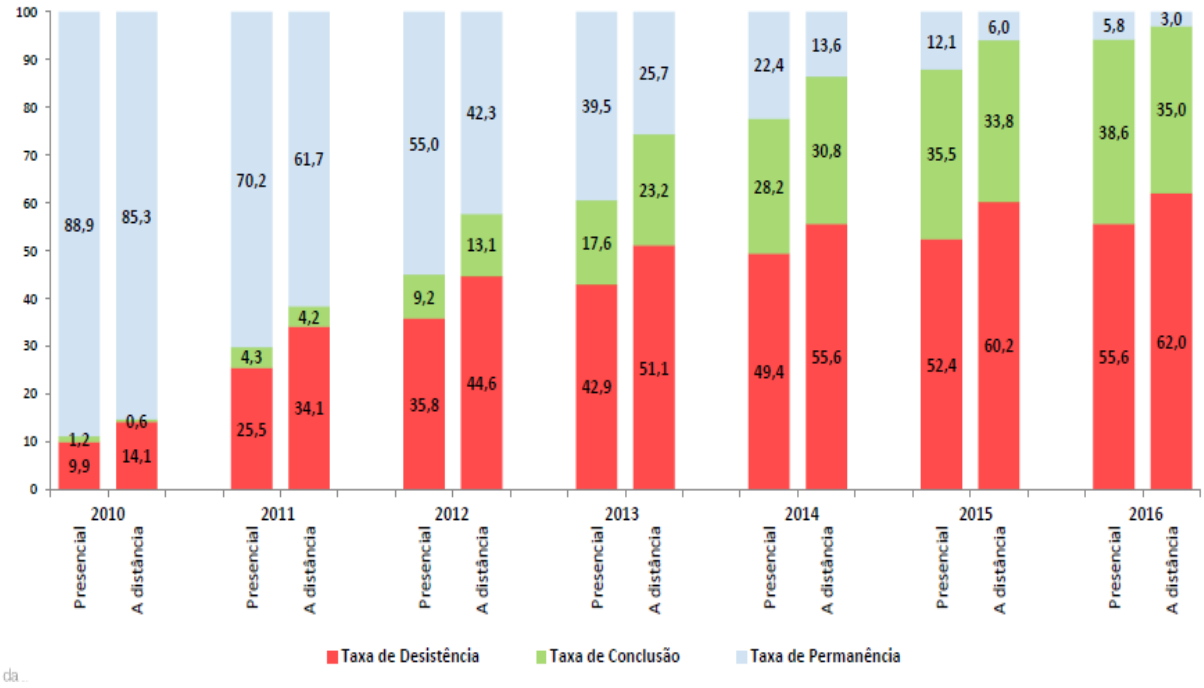


Fonte: Censo da Educação Superior 2018

Importante destacar que pelo Plano Nacional de Educação do Brasil (2014-2024), existem algumas metas definidas para acelerar o ritmo e a direção da educação superior, em destaque a meta número 12 (doze) que deseja elevar a taxa de matrículas na educação superior para 50% em 2024 (em 2018 em 30,6%) (INEP,2014).

Outro indicador do INEP é a taxa de desistência (evasão) no ensino superior, ou seja, o fluxo de entrada do aluno em determinado momento até o momento da sua conclusão ou evasão. Para alunos que ingressaram (matricularam-se) em 2010 no ensino superior, nas modalidades presencial e à distância, quando observado o ano de 2016 a taxa estava em 56,8%, ou seja, para cada 100 alunos matriculados em 2010 no ensino superior, em média 57 alunos desistiram do curso entre 2010 e 2016. No Gráfico 4 é possível acompanhar a evolução desta taxa no decorrer dos anos.

GRÁFICO 4 - EVOLUÇÃO DOS INDICADORES DA TRAJETÓRIA DOS ESTUDANTES COM INGRESSO EM 2010



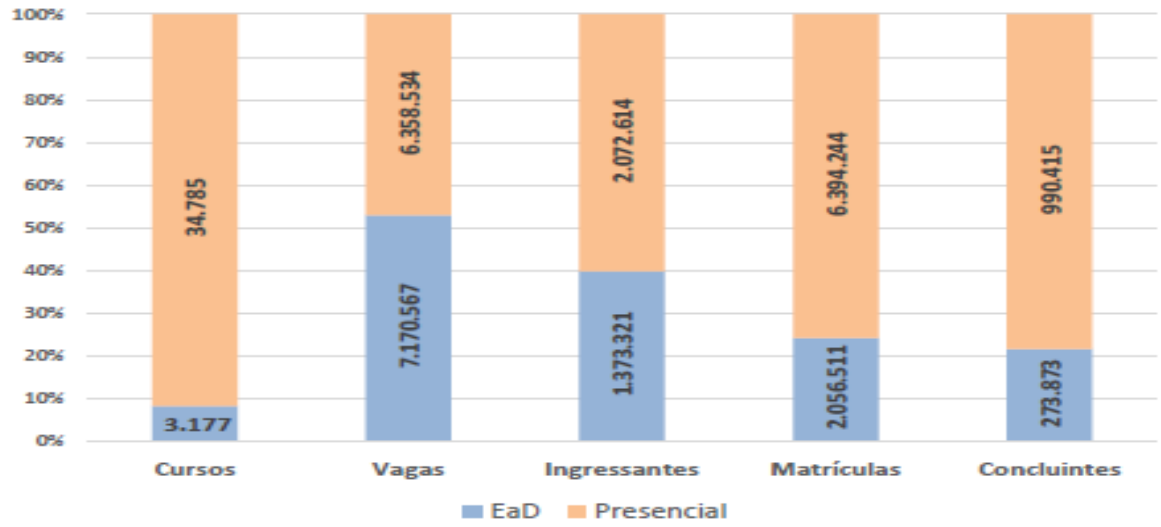
Fonte: Censo da Educação Superior 2018

Ao analisar o gráfico observa-se que no ano de 2010, de 100% dos alunos que se matricularam em 2010 na modalidade presencial já existiam 9,9% dos alunos que desistiram no próprio ano (2010). Observa-se em 2011, de 100% dos alunos que se matricularam em 2010, em 2011 25,5% dos alunos da modalidade presencial desistiram entre os anos de 2010 e 2011, ou seja, o fluxo de acompanhamento dos alunos que se matricularam em 2010.

Ao observar as taxas de desistência (evasão) ano a ano de alunos ingressos em 2010 no decorrer do tempo, o ano que possui a maior evolução da taxa é no segundo ano, um salto da taxa de 9,9% na modalidade presencial em 2010 para 25,5% em 2011 na mesma modalidade, entre os outros anos não observamos um crescimento maior que os 15,6% (25,5% - 9,9%) entre um ano e outro.

Por fim, neste cenário de indicadores do ensino superior a síntese dos dados absolutos em números de cursos, vagas, ingressantes, matrículas e concluintes no Brasil no Gráfico 5 a seguir:

GRÁFICO 5 - ESTATÍSTICA DA EDUCAÇÃO SUPERIOR POR MODALIDADE DE ENSINO – BRASIL 2018



Fonte: Censo da Educação Superior 2018

O INEP possui uma grande quantidade de indicadores para acompanhar a evolução da educação no país, mas os dados demonstrados nesta pesquisa subsidiam sobre a baixa eficiência do sistema de ensino superior no Brasil em relação a evasão, ou seja, a evasão atual compromete a possibilidade de cumprir as metas previstas no Plano Nacional de Educação do Brasil (2014-2024).

2.3 FATORES QUE LEVAM À EVASÃO NO ENSINO SUPERIOR

Compreender o tema dos fatores que levam a evasão é pesquisado a muito tempo, Bueno (1993) em seu trabalho de título, “A Evasão de Alunos”, realizou uma pesquisa levando-se em conta questões ligadas a escolha profissional, as expectativas de realização pessoal e o sucesso profissional do aluno. Em um trecho de seus estudos relatou:

> A falta de prestígio social de certas profissões reduz os incentivos para que estas sejam buscadas com persistência; o aviltamento salarial e as dificuldades de obter condições adequadas de trabalho levam os cursos de licenciatura e de bacharelado a serem considerados uma atividade secundária na ordem do reconhecimento social. As possibilidades limitadas de sucesso financeiro como empregados ou no magistério se mostram palpáveis já no início da vida universitária. Com chances limitadas de emprego, com falta de prestígio, de condições de trabalho, de sucesso financeiro, a realização profissional passa a ser apenas uma fantasia na cabeça dos estudantes de cursos que levam a profissões com estas características (magistério secundário, empregados em áreas técnicas e de pesquisa, etc); à

> primeira dificuldade, a evasão do candidato a estas profissões é a consequência natural. (BUENO, 1993, p. 11)

Neste sentido, Barroso e Falcão (2004), em pesquisa realizada com alunos de graduação do curso de licenciatura em Física da Universidade Federal do Rio de Janeiro, enfatizam que são três os fatores que levam a evasão: o fator econômico, o vocacional e o institucional.

Outra abordagem é em relação a distância geográfica entre a moradia do estudante e a instituição de ensino. Esta questão, foi apontado por Parente (2014) e Lima e Zago (2016):

> A evasão no ensino superior ocorre de diferentes formas, mas uma proporção importante (38,2%) das interrupções de curso se configura por mobilidade (tanto do curso quanto da instituição) e não de desistência do sistema de ensino. (LIMA; ZAGO 2016, p.1)

Também Lima e Zago (2016, p.5), em suas pesquisas identificaram tendências da evasão no ensino superior

> Portanto, observamos que as desigualdades sociais e culturais marcam o perfil do estudante evadido, sobretudo pelo baixo capital econômico e cultural familiar, e que o ingresso no ensino superior não é suficiente para equilibrar estas desigualdades, o que nos leva a questionar se o simples fato de dar oportunidades iguais de acesso aos desiguais é suficiente para garantir a concretização dos estudos de ambos. (LIMA; ZAGO, 2016, p.8)

Os autores destacam que a análise individual dos fatores não é possível de ser realizada:

> Verificando as principais causas citadas em relação à evasão, observamos que não é possível analisá-las isoladamente, pois entendemos que cada fator está ligado a um contexto institucional e pessoal e que reduzir as explicações acerca das causas da evasão associando-as unicamente a falta de condições financeiras do estudante em pagar a mensalidade nas IES cujo ensino não gratuito, ou ainda que o universitário não conseguiu acompanhar o currículo escolar mais exigente nas IES públicas, seria negar a presença de desigualdades sociais e tensões entre elas e, por várias razões, o impasse vivenciado pelo estudante entre prosseguir ou não seus estudos. (LIMA, ZAGO, 2016, p.8)

Para Santos (2014, p. 253) o fator do comprometimento do estudante é superior em relação a outros fatores na educação à distância:

> Embora apareçam aspectos em relação à gestão acadêmica, estes surgem em proporção significativamente menor do que os olhares em relação ao comprometimento do estudante. Acredita-se ser necessário realizar estudos que triangulem a qualidade do ensino, o comprometimento institucional e o comprometimento do estudante em relação à Educação Superior.

No Censo realizado pela ABED (Associação Brasileira de Educação à Distância) em 2013, a evasão é colocada como uma das maiores preocupações dos gestores e que diversas pesquisas são realizadas para compreender os fatores que levam a evasão e mecanismos de predição, prevenção e correção.

A dispersão da atenção da vida urbana dificulta a autonomia e organização pessoal, indispensável para o processo de aprendizagem a distância, segundo Moran (2003). O autor relata que estudantes com dificuldade de organizar sua agenda terá dificuldade de acompanhar os estudos e evade. Informações como idade, sexo, polo que estuda, curso escolhido podem ser fatores que podem determinar também a evasão.

O pesquisador Kember et al. (1992) propôs um modelo de evasão que no contexto da educação à distância pressupõe um processo linear do progresso do aluno adulto conforme suas características de entrada no curso como nível educacional, estado civil, idade, sexo, emprego e status familiar. Estas características promovem dois cenários, o primeiro da atribuição externa e incompatibilidade acadêmica e a segunda da integração social e acadêmica.

Laguardia e Portela (2009) destacam a expansão da educação à distância e a relação com a evasão:

> O destaque dado à evasão e as razões para a sua ocorrência decorrem, em grande medida, da expansão da educação online e a sua vinculação à qualidade dos processos de aprendizagem em ambientes virtuais. Entretanto, as diversas perspectivas teóricas utilizadas nas pesquisas para descrever ou predizer a evasão, a extensa lista de variáveis apontadas como preditoras e o número expressivo de estudos que apresentam resultados divergentes sinalizam para o caráter complexo e multidimensional desse fenômeno.

Segundo os autores algumas estratégias são necessárias para o combate da evasão, sendo:

a) A primeira etapa corresponde ao momento em que o aluno potencial expressa seu interesse pelo curso e requer informações mais detalhadas para tomada de decisão. A escolha de se matricular e, após algum tempo, descobrir que escolheu o curso errado é mais comum nos cursos que não exigem requisitos formais para a matrícula;
b) A segunda etapa é definida como o período entre a matrícula e o início do curso, em que as estratégias de registro temporário, o direito ao reembolso por desistência dentro de um prazo estipulado, o envio de material de apoio e cartas de boas-vindas, assim como o contato individual do tutor com cada aluno ajudam a estabelecer a percepção de pertencimento à instituição.
c) A terceira etapa se estende do envio da primeira atividade à avaliação final, na qual a evasão do aluno depende, em certa medida, da garantia de processos de aprendizagem de qualidade e uma equipe amigável, entusiasmada e profissional
d) A quarta e última etapa está baseada no retorno do aluno após a conclusão do curso, quando ele escolhe a mesma instituição para um próximo curso e essa escolha depende da percepção da experiência prévia de estudo a distância e se essa experiência valeu o dinheiro e o tempo investidos.

Segundo Gottardo et al. (2012 p. 4):

> Em resumo, diversos trabalhos investigaram a aplicação de técnicas de Mineração de Dados Educacionais com o objetivo de realizar inferências ou previsão a respeito do desempenho de estudantes. Entretanto, de forma geral, são analisados os resultados considerando dados envolvendo todo o período de realização do curso, que só podem ser obtidos ao final do mesmo. Desta forma, as informações obtidas, embora consideradas importantes ao contexto educacional, somente poderão ser utilizadas como apoio para ações envolvendo estudantes de cursos futuros.

Nesse sentido, Gottardo 2012) analisa que com o uso da mineração de dados, enfatiza-se a necessidade de que modelos de predição precisam abordar informações iniciais que permitam prever a evasão ao início do processo e não ao final ou para turmas futuras.

2.4 MINERAÇÃO DE DADOS NA EDUCAÇÃO

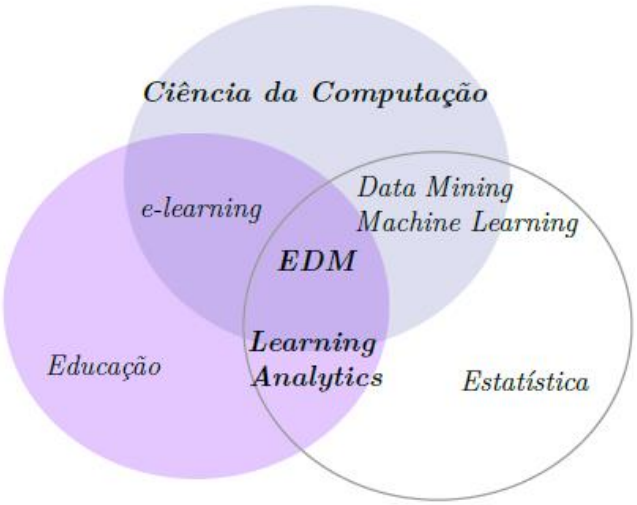
Com a mineração de dados é possível a descoberta de conhecimento de dados, conforme Han (2006) é a descoberta de informações úteis em um grande conjunto de informações, que podem contribuir ao processo decisório. Com as novas tecnologias nos últimos 20 anos presentes em todas as áreas do conhecimento a mineração de dados ganha destaque.

Na área da educação a importância da mineração de dados não poderia ser diferente de outras áreas, e a aplicabilidade encontrou um potencial de geração de informações úteis para o desenvolvimento de soluções.

Alguns autores conceituam EDM *(Education Data Mining*), como mineração de dados para a educação (ROMERO; VENTURA 2013, PAIVA et al., 2014), uma área com o objetivo de investigar o desenvolvimento de métodos para fazer descobertas de conhecimento com base em estudos e pesquisas na área de educação. Através do acesso ao portal da *Education Data Mining*[2] é possível acompanhar as pesquisas, fóruns, congressos, eventos etc sobre o tema EDM.

Na Figura 1 observa-se a interdisciplinaridade das áreas do conhecimento da EDM:

Figura 1 - Áreas que envolvem o KDD e EDM



Fonte - Romero e Ventura (2013)

[2] *Education Data Mining*. Disponível em https://educationaldatamining.org/

Nas pesquisas de Romero e Ventura (2013) uma das abordagens da aplicação do EDM é a predição, ou seja, uma forma apropriada de estimar o valor desconhecido de uma variável que descreve o estudante.

Em relação as principais categorias do EDM, Baker (2011) elaborou um quadro demonstrando os métodos, o objetivo de cada método e a principal aplicação, abaixo o Quadro 2 adaptado por Sepúlvida (2016):

QUADRO 2 - PRINCIPAIS CATEGORIAS DO EDM

Atividade/Método	Objetivo do Método	Principal Aplicação
Predição – Classificação – Regressão – Estimação	O objetivo é desenvolver modelos que deduzam aspectos específicos dos dados, conhecidos como variáveis preditivas (*predicted variables*), através da análise e fusão dos diversos aspectos encontrados nos dados, chamados de variáveis preditoras (*predictor variables*).	Detectar comportamentos dos estudantes, como por exemplo, comportamentos fora do padrão; possíveis resultados educacionais.
Agrupamento	Encontrar dados que naturalmente se agrupam, dividindo dados completos em conjuntos de categorias.	Descobrir novos padrões de comportamentos e investigar diferenças e semelhanças entre os grupos.
Associação – Mineração de Regras de associação – Mineração de Correlações – Mineração de Padrões Sequenciais – Mineração de Causas	Descobrir relações entre as variáveis.	Descobrir associações curriculares na sequência do curso; descobrir quais estratégias pedagógicas levam a uma aprendizagem mais robusta e eficaz.
Descoberta de Modelos	Um modelo de um fenômeno desenvolvido com predição, aglomeração, ou conhecimento de engenharia, é utilizado como um componente adicional em relação a predição.	Descoberta de relações entre os comportamentos dos estudantes e características dos estudantes ou variáveis contextuais; análise da questão de pesquisa em toda a grande variedade de contextos.
Destilação de dados para facilitar as decisões humanas	Os dados são refinação para permitir a um ser humano identificar ou classificar rapidamente características dos dados.	Identificação humana de padrões na aprendizagem dos estudantes, comportamento, ou de colaboração; dados de rotulagem para uso no desenvolvimento posterior do modelo de previsão.

Fonte: Sepúlvida (2016)

O método por predição, através de classificação é a abordagem desta pesquisa que está detalhado mais adiante.

Para Han e Kamber (2006): mineração de dados é o processo de descoberta do conhecimento útil escondido em alguma base de dados ou repositório de informação, assim como encontrar um determinado padrão nas informações.

Para Baker e Isotani (2011):

> A mineração de dados educacionais (EDM)[...] tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Atualmente ela vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino. (BAKER e ISOTANI, 2011, p 1).

Conforme Carvalho (2005, p. 45) mineração de dados relaciona-se ao “Uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano. ”

Para Manhães et al. (2015), as técnicas de mineração de dados podem ser classificadas em duas categorias, a descritiva e preditiva. A categoria descritiva tem por objetivo analisar os dados, descrever suas características e apresentar propriedades interessantes gerais dos dados. A categoria preditiva tem por objetivo analisar os dados, a fim de construir um ou um conjunto de modelos, e tentar fazer inferências sobre os mesmos de modo que o sistema possa fazer predições ou prever o comportamento de novos conjuntos de dados.

Os modelos descrevem aspectos específicos dos dados, desta forma necessita-se de uma certa quantidade de dados que devem possuir um conjunto de características (atributos), que descrevem corretamente grupos ou classes distintas.

Assim Manhães et al. (2015) relaciona as pesquisas em EDM em duas grandes linhas de pesquisa:

a) Identificar atributos relevantes para caracterizar estudantes ou;

b) Identificar e comparar o desempenho de algoritmos classificadores.

O autor reforça que a primeira linha está relacionada em utilizar a EDM para identificar, em bases de dados educacionais, os atributos mais relevantes para caracterizar os grupos de estudantes. A segunda está relacionada ao estudo e

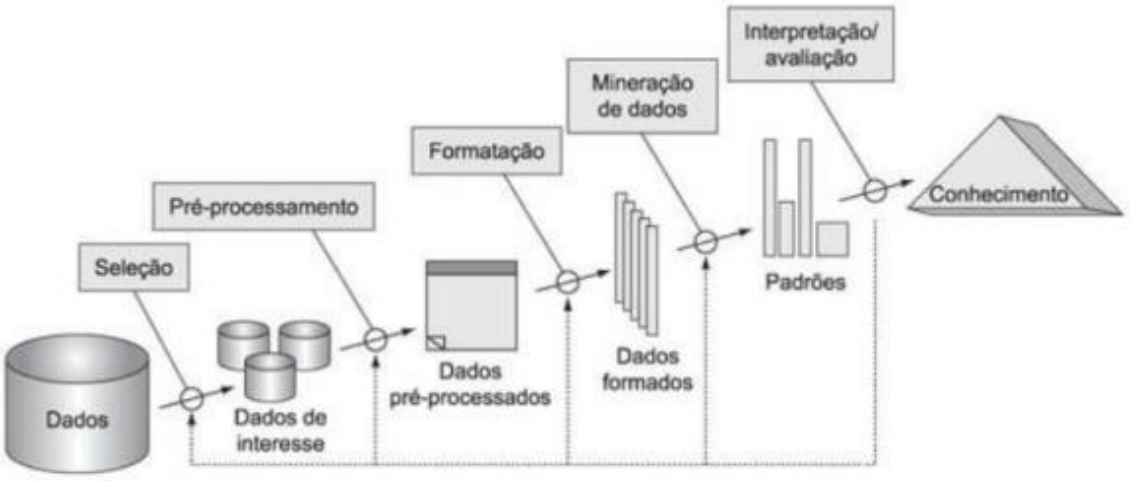
seleção dos algoritmos mais apropriados para a solução de problema, estimando o desempenho quando aplicados a dados educacionais.

Importante ressaltar que quanto maior a quantidade e a qualidade das variáveis de entrada (preditoras), melhor a qualidade das predições do algoritmo. A fonte de dados precisa ser segura e confiável para investigar o problema.

Fayyad (1996, p. 43), define KDD como um “processo não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados”. O autor descreve que estas informações implícitas são de difícil detecção utilizando métodos tradicionais que tratam de informações, ou seja, com métodos da extração do conhecimento é possível identificar informações implícitas e contribuir ao processo decisório.

Para a geração da descoberta do conhecimento, conforme Fayyad (1996) algumas etapas são necessárias conforme a Figura 2, desde a seleção até a interpretação e avaliação dos dados.

FIGURA 2 - ETAPAS DO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS



Fonte: Fayyad (1996)

Silva (2019) descreve este processo para a geração do conhecimento em duas grandes etapas, pré-processamento e pós-processamento. A etapa do pré-processamento é mais trabalhosa, com as seguintes ações:

1. Limpeza dos dados, que consiste em remover ruídos e dados inconsistentes.

2. Integração dos dados, quando é necessário fazer fusão de dados de mais de uma fonte de dados.

3. Seleção dos dados, na qual os dados relevantes para a tarefa de análise são selecionados a partir da base de dados, sendo ela integrada ou não.

4. Transformação dos dados, os dados são transformados e consolidados em formato adequado para a mineração.

5. Na fase de mineração de dados são aplicadas as técnicas ou métodos para extrair padrões dos dados.

A fase do pós-processamento que deve apresentar resultados válidos e úteis e inclui as ações de:

1. Avaliação: que consiste em identificar os padrões que representam conhecimento útil, avaliados a partir de medidas de interesse.

2. Apresentação do conhecimento extraído.

Silva (2004) indica a mineração de dados como uma das etapas do KDD:

> As ferramentas e técnicas empregadas para análise automática e inteligente dos imensos repositórios de dados de indústrias, governos, corporações e institutos científicos são os objetos tratados pelo campo emergente da Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases*-KDD). Mineração de dados é a etapa em KDD responsável pela seleção dos métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão. (SILVA, 2004, p. 1)

Por fim, Romero e Ventura (2010) identificaram as tendências nas pesquisas e desenvolvimentos em EDM:

a) O desenvolvimento de ferramentas de EDM para educadores e gestores acadêmicos que não são peritos em mineração de dados;
b) As operações de pré-processamento das informações, facilidades de configurações dos algoritmos e interpretação dos resultados dos algoritmos estão a parte do interesse dos educadores, por isso a necessidade de criação de ferramentas mais genéricas, configuráveis e de simples manipulação;

c) Não há ferramentas de EDM que possam ser reutilizadas em qualquer sistema educacional, em especial no contexto das IES brasileiras; e
d) Não há uma padronização para entrada de dados e resultado dos modelos obtidos, após as fases de pré-processamento, mineração de dados e pós-processamento dos dados educacionais.

Deste modo é possível dizer que a mineração de dados contribui ao processo de predição é torna-se um instrumento importante de gestão no processo decisório.

2.5 MODELOS DE PREDIÇÃO

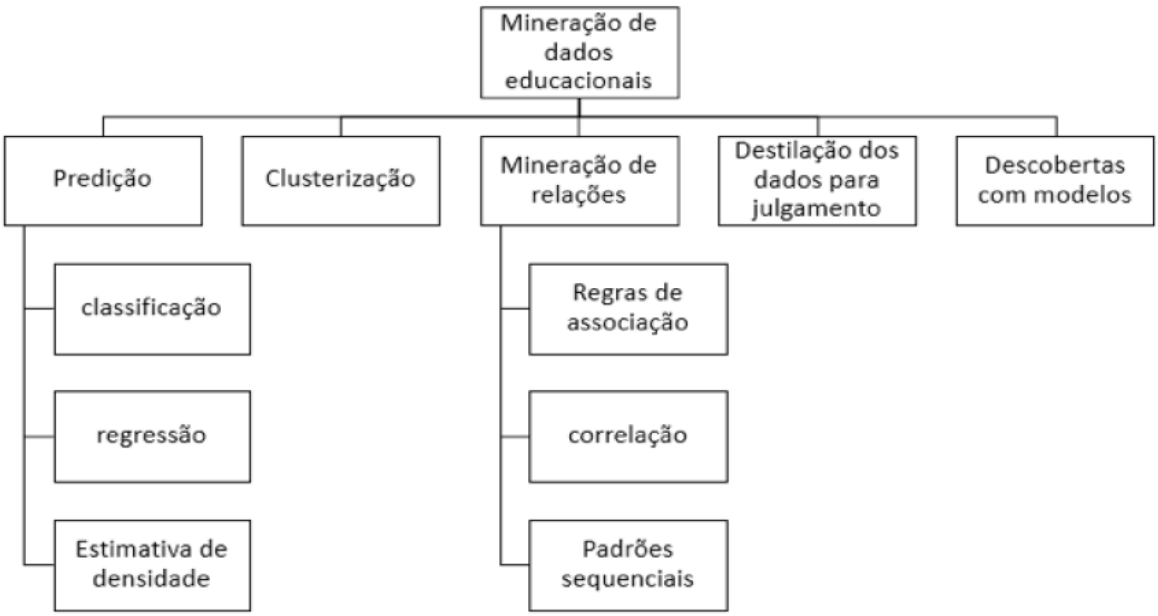
Silva (2019) define que as pesquisas em EDM concentram-se em Mineração de Dados, Aprendizado de Máquina, Psicometria, Estatística, Visualização de Informações e Modelagem Computacional.

As técnicas que se aplicam na EDM são extraídas de diversos conhecimentos que Baker (2009, 2010, 2011 e 2014) no desenvolvimento de sua taxonomia dividiu em 5 categorias, sendo:

a) Predição;
b) Clusterização;
c) Mineração das relações;
d) Destilação de dados para julgamento;
e) Descoberta de modelos.

Na Figura 3, observa-se também a subdivisão quando aplicável destas técnicas:

FIGURA 3 – TAXONOMIA DAS PRINCIPAIS SUBÁREAS DE PESQUISA EM EDM



Fonte: Silva (2019)

Conforme a Figura 3, a Predição subdivide-se em Classificação, Regressão e Estimativa de Densidade.

A tarefa de Regressão é parecida à tarefa de Classificação definida mais abaixo, onde são buscadas funções que relacionem os registros de uma base de dados em um intervalo de valores reais (FURLAN,2018). A principal diferença é que o atributo-alvo adota valores numéricos. A tarefa de Regressão utiliza-se da Estatística, Rede Neurais entre outras técnicas que oferecem os recursos para sua implementação (MICHIE et al., 1994).

A classificação, como um dos itens da predição, pode ser conceituada conforme Manhães et al. (2015, p.65):

> Como um processo de encontrar um modelo ou função que descreve e distingue classes de dados com o propósito de utilizar o modelo encontrado para predizer a classe de um novo elemento cuja identificação da classe é desconhecida. O modelo gerado baseia-se na análise de um conjunto de treinamento com os rótulos das classes bem definidos e conhecidos. A classificação utiliza algoritmos supervisionados para inferir (predizer) o grupo ou classe dos novos exemplos (registros). O algoritmo precisa de um conjunto de dados, na qual os exemplos (registros) possuem classes conhecidas, para aprender a identificar quais valores de atributos são importantes para definir ou caracterizar exemplos de cada classe.

Na conceituação de algoritmo Farrer (1989, p. 10) descreve como um “conjunto de comandos que, obedecidos, resultam numa sucessão finita de ações”. Podemos exemplificar que um computador executa comandos a fim de solucionar problemas.

Segundo Saliba (1992), dá-se o nome de algoritmo à especificação da sequência ordenada de passos que deve ser seguida para a realização de uma tarefa, garantindo a sua repetibilidade; ou seja, um algoritmo é uma lista de instruções para a execução, passo a passo, de algum processo. Há inúmeros casos que podem exemplificar o uso (involuntário ou não) de algoritmos para a padronização do exercício de tarefas rotineiras, como por exemplo; uma receita de bolo, trocar um pneu furado ou uma lâmpada queimada.

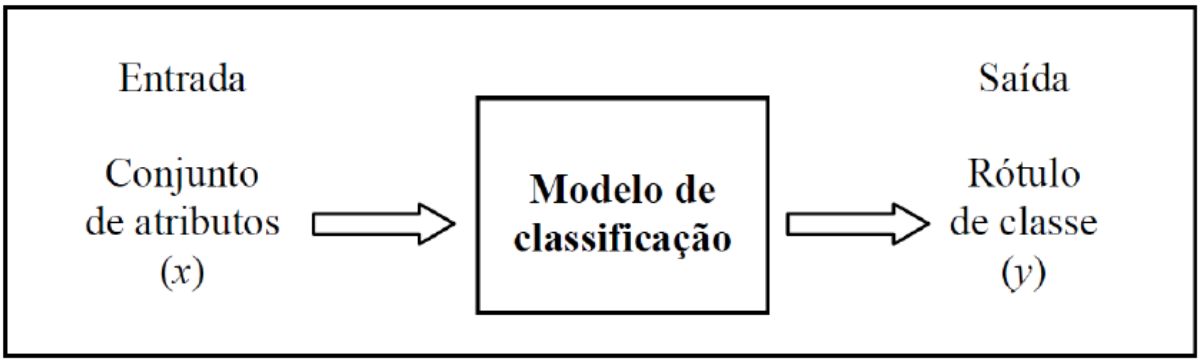
Neste processo Manhães et al. (2015, p.61) descreve os algoritmos classificadores:

> Há diversos algoritmos classificadores e diversas formas de representar o conhecimento. A escolha dos algoritmos para aprendizado do modelo depende das diversas características encontradas nos dados de entrada. A qualidade e a quantidade de dados influenciam diretamente no aprendizado do modelo. Neste caso, a qualidade representa o quanto o conjunto de entrada é significativo para descrever a classe a ser aprendida e a quantidade representa um número adequado de exemplos na base para treinamento e teste do modelo aprendido.

A tarefa de classificação faz uso do aprendizado supervisionado que é um tipo de aprendizado indutivo. Este tipo de aprendizado utiliza os dados mostrados na forma de pares ordenados (entrada, saída desejada), e necessita que os registros do conjunto de dados tenham suas classes pré-definidas (GOLDSCHMIDT, 2015).

A Figura 4 ilustra o processo de classificação:

FIGURA 4 - TAREFA DE CLASSIFICAÇÃO



Fonte: Tan (2009)

A tarefa de classificação consiste em organizar objetos em uma categoria pré-definida, sendo a aprendizagem de uma função, denominada modelo de classificação, onde mapeia-se cada conjunto de atributo x a um único rótulo y (classe) ou objeto de saída (SILVA,2019).

Este processo ocorre em duas etapas (HAN, 2006), sendo:

a) na primeira etapa constrói-se um modelo de classificação com base nos dados históricos. Nesta etapa ocorre o processo de aprendizagem ou de treinamento onde, a partir do algoritmo de classificação constrói-se o modelo, analisando ou "aprendendo com" um conjunto de treinamento.
b) na segunda etapa, denominada etapa de teste, utiliza-se o modelo gerado para classificação, cujo objetivo é avaliar a precisão do modelo para classificar novos dados . Para a etapa de teste usa um conjunto de atributos, cujos dados não foram utilizados na etapa de treinamento para evitar a sobrecarga dos dados.

Ao final deste processo, avalia-se se os registros de saída foram classificados de forma correta ou não, esta avaliação se denomina de desempenho do modelo de classificação e assim avalia-se a qualidade do classificador escolhido (acurácia).

A qualidade do classificador direciona a qualidade das decisões que serão tomadas pelos gestores e o algoritmo precisa corresponder às expectativas. Mas na prática podemos ter algoritmos respondendo diferente das expectativas e testes precisam ser realizados para garantir a acurácia ou não. Han e Kember (2006) apresentam métricas para medir a qualidade da classificação do algoritmo:

a) *Acurácia* – é a precisão de um classificador, dado um determinado conjunto de teste, obtém-se a porcentagem dos que estão corretamente classificadas no conjunto todo de dados pelo classificador;
b) *Taxa de erro ou acerto* – significa o quanto o modelo acertou ou errou na predição dos exemplos de cada classe analisada;
c) *Matriz de confusão* – é um recurso muito útil para análise do resultado do classificador, pois mostra o quantitativo para as diferentes classes investigadas;
d) *Kappa* – utiliza-se a medida estatística *Kappa* para medir o número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre o previsto e o observado em um conjunto de dados. O coeficiente *Kappa* é calculado levando-se em consideração todas as classes e é útil para mensurar o grau de concordância ou qualidade do classificador.

Neste processo a escolha adequada do modelo de classificação torna-se importante para garantir taxas significativas de desempenho do algoritmo classificador escolhido.

2.6 ALGORITMOS DE CLASSIFICAÇÃO

Os principais algoritmos de classificação aderentes a esta pesquisa para um modelo de predição para encontrar grupos de risco, ou seja, alunos propensos a evadir são:

a) *Naive Bayes*: classificador apoiado na teoria das probabilidades, calcula a probabilidade de uma variável pertencer a uma determinada classe pré-definida sobre um evento já ocorrido. Parte do princípio de que as variáveis preditoras são independentes, desta característica surge o uso do termo *Naive* (ingênuo) (ZHANG, 2004), uma abordagem muito utilizada nas pesquisas investigadas;
b) K Vizinhos Mais Próximos *(Nearest-Neighbor – KNN*): é um classificador denominado de preguiçoso, pois a princípio não se cria um classificador, mas sim um modelo baseado em instâncias, o que significa que para cada dado

novo inserido o mesmo é comparado com uma base de dados com classe atribuída do conjunto de dados do treinamento. A nova instância será atribuída a classe mais próxima (vizinha) (YANG,1994).

c) *Support Vector Machine* (SVM): técnica utilizada quando é possível a separação linear das classes, no momento do treinamento define-se a maior margem de separação entre as classes (TAN, 2009).

2.7 WEKA

O WEKA, conforme Waikato (2017) é um sistema que possui um conjunto de algoritmos de aprendizado de máquina para a execução de tarefas de mineração de dados. Com estas ferramentas internas no sistema é possível realizar pré-processamento de dados para classificação, clusterização, regras de associação, regressão, correlação entre outras funcionalidades que permitem também o desenvolvimento e novos sistemas e modelagem de aprendizagem de máquina. Abaixo a Figura 5 representa a tela inicial do sistema *WEKA*.

FIGURA 5 - TELA INICIAL DO SISTEMA *WEKA*

Fonte - Autor

Observa-se na Figura 5 que na tela inicial demonstra a versão do software WEKA, neste caso a versão 3.8.4. O sistema não possui versão em português.

As aplicações possíveis do *WEKA* são:

a) *Explorer:* neste módulo é possível operacionalizar um arquivo de dados em um dos algoritmos escolhido disponíveis;
b) *Experimenter*: Possibilita aplicar vários métodos de classificação sobre um conjunto de dados e comparar o desempenho de cada método;
c) *Knowledge Flow*: Interface que demonstra a operação interna do programa *WEKA,* assim como algumas parametrizações possíveis;
d) *Workbench*: Uma interface de bancada de trabalho que permite de forma mais amigável observar em uma única tela as funcionalidades do sistema, desta versão e versões anteriores;
e) *Simple CLI:* abreviatura de *Simple Client*, interface na forma de console que possibilita para os conhecedores da programação do *WEKA* uma forma direta de realizar a programação.

Com a utilização do pacote de software do *WEKA,* como sistema base para a programação de simuladores computacionais de predição, utiliza-se o JAVA como linguagem de programação, linguagem raiz do *WEKA* (código-fonte).

2.8 COMPARATIVO OUTRAS PESQUISAS

Na realização do levantamento bibliográfico, identificou-se estudos que apresentaram suas metodologias e resultados sobre a predição da evasão com uso de técnicas de mineração de dados e modelos de predição, a seguir o descritivo destes estudos:

Silva (2019) em sua pesquisa com alunos do ensino superior na modalidade à distância, com o uso de alguns indicadores de mineração de dados, em um modelo indutivo, encontrou relevante grau de confiança em seus estudos com a média de 84,03% de acurácia em suas predições.

Martinho (2014) propôs um sistema inteligente de predizer alunos do ensino superior presencial no grupo de risco de evasão, utilizando para o desenvolvimento de uma rede neural inteligente, o *ARTMAP-Fuzzy*. A pesquisa fundamentou-se em dados socioeconômicos e acadêmicos que atingiu grau de eficiência entre 76% e 95%.

Sepúlvida (2016), estudou o ensino superior na modalidade à distância, e a acurácia dos seus estudos demonstraram valor de 80,95% em seu modelo de predição, aplicando a mineração de dados principalmente nas informações iniciais geradas pelo sistema AVA da instituição pesquisada nos primeiros 30 dias do aluno no ensino superior.

Manhães et al. (2015), em uma abordagem para fornecer conhecimento aos gestores acadêmicos sobre os alunos em grupo de risco de evasão, estudou na Universidade Federal do Rio de Janeiro, graduandos pelo período de 16 anos. Desenvolveu uma arquitetura *EDM Wave* com o uso do pacote de software do *WEKA* e chegou a uma acurácia próxima de 80% em suas predições.

Araújo (2019) em sua tese na área de inteligência artificial em educação, fez diversos experimentos no ensino superior, nas modalidades presencial e à distância, no curso de Engenharia da Computação, utilizando o pacote de software do *WEKA* para modelos de predição e chegou em 92% de acurácia no ensino à distância.

Fazolin (2017) com o uso do algoritmo classificador *Naive Bayes* obteve uma acurácia entre 81% a 83% em seus experimentos no ensino fundamental.

Santos (2015), com a utilização de técnicas de mineração de dados, estudou a predição do curso de Letras no ensino à distância com dados a partir do AVA (dados acadêmicos) e chegou em acurácia próxima de 86% com a utilização do pacote de software do *WEKA.*

Fernandes (2019), utilizando a mineração de dados na educação (EDM) estudou a evasão no curso de computação da Universidade Federal do Pampa, com atributos socioeconômicos dos alunos. No algoritmo estudado *KNN* o índice de acurácia chegou-se em 98%.

Silva (2019) em estudo no Instituto Federal do Rio Grande do Norte, em cursos técnicos utilizou-se de técnicas de mineração de dados seguindo a metodologia *CRISP-DM* para descobrir padrões implícitos e possíveis correlações entre eles, com o auxílio do software Orange. As taxas de acurácia encontradas são próximas de 85%.

Schimitt (2018) desenvolveu dois experimentos abrangendo três cursos de graduação em que foram empregados dados de interações dos alunos no AVEA e dados do sistema de gestão acadêmicos da instituição pesquisada. Como resultados foram obtidos bons índices de acertos, permitindo assim, a conclusão de que a

abordagem proposta é factível para detecção de alunos com tendência a evasão em cursos de graduação EAD. Também foi desenvolvido um aplicativo para dispositivos móveis que permitem os gestores acompanharem os indicadores de evasão e assim contribuir no processo decisório a fim de mitigar a evasão.

Kraemer (2018), com o intuito de auxiliar a inferir os alunos com maior tendência a evadirem dos cursos, desenvolveu um sistema de estimativa da probabilidade de evasão, utilizando-se de algoritmos de classificação já existentes e um sistema web de *Business Intelligence*, tornando assim possível a tomada de decisão a partir de informações quantitativas mais precisas. Este sistema foi desenvolvido e testado sobre uma base de 10.371 registros de alunos de dez cursos de graduação e obteve uma margem de sucesso na previsão da tendência de evasão do aluno variando de 92,34% no mínimo e 99,32% no máximo, com média de 95,81% entre todos os cursos analisados sobre a base de testes.

As pesquisas citadas acima, que abordam este tema da evasão com as técnicas de mineração de dados, ou seja, as investigações sobre os alunos propensos a evadir com o uso de métodos de predição, demonstram um índice de acerto entre 76% a 97%.

3 METODOLOGIA

Nesta fase os fundamentos dos métodos escolhidos no desenvolvimento da pesquisa são descritos, assim como a sustentação teórica dos critérios utilizados a fim de demonstrar todos os caminhos percorridos.

3.1 CARACTERIZAÇÃO E PROCESSO DE DESENVOLVIMENTO

O processo de desenvolvimento desta dissertação iniciou-se com um levantamento das publicações sobre o tema escolhido: simulador computacional preditivo de evasão no ensino superior à distância com uso dos recursos da mineração de dados suportado pelo pacote de software do *WEKA* em cursos de engenharia. O mecanismo de busca utilizou-se de alguns dos principais repositórios de pesquisas e estudos acadêmicos e científicos na internet, escolhendo-se: *Scielo*, CAPES e *Google Scholar*.

O espaço temporal da pesquisa referencia-se principalmente em estudos realizados nos últimos 10 anos, não deixando de utilizar-se de trabalhos e autores clássicos referenciados em muito dos materiais lidos.

No Quadro 3, observam-se as 5 (cinco) sentenças de busca utilizadas, com as palavras de busca (evasão, ensino superior, mineração de dados, *WEKA*, engenharia, data mining, ensino à distância e predição) assim como a quantidade de materiais resultado da busca de cada sentença.

QUADRO 3 - SENTENÇAS UTILIZADAS DE BUSCA

Sentenças	Definição da Sentença	Resultados da Busca (Quant.)		
		Scielo	CAPES	Google Scholar
Sentença 1	Evasão	239	1.781	22.100
Sentença 2	Evasão AND ("Ensino Superior" OR "Educação Superior")	62	468	14.800
Sentença 3	Evasão AND ("Ensino Superior" OR "Educação Superior") AND ("Mineração de Dados" OR "Data Mining")	-	6	790
Sentença 4	Evasão AND ("Ensino Superior" OR "Educação Superior") AND ("Mineração de Dados" OR "Data Mining") AND ("Weka")	-	1	191
Sentença 5	Evasão AND ("Ensino Superior" OR "Educação Superior") AND ("Mineração de Dados" OR "Data Mining") AND ("Weka") AND ("Ensino a distância" OR "educação a distância") AND ("Predição") AND ("Engenharia")	-	-	47

Fonte: do Autor

Scielo: *http://www.scielo.org*

CAPES: http://www.periodicos.capes.gov.br

Google Scholar: http://scholar.google.com

No quadro observa-se a diversidade da quantidade de trabalhos resultantes da busca para cada sentença, observando que no *Google Scholar* encontra-se trabalhos referente a busca mais criteriosa (sentença 5) de todos os termos principais abordados nesta pesquisa, sendo: evasão, educação superior, ensino superior, mineração de dados, *data mining*, *WEKA*, ensino á distância, educação à distância, predição e engenharia.

Após a leitura do resumo dos trabalhos apresentados nas sentenças 4 e 5 do Quadro 3 dos portais de busca correspondentes, selecionou-se as seguintes pesquisas no Quadro 4 como referência aos estudos, comparações de resultados e bibliografias:

QUADRO 4 - PESQUISAS SELECIONADAS

Ano	Tipo	Título	Autor
2019	D	Um modelo descritivo para auxiliar o acompanhamento da evasão escolar nos cursos técnicos e superiores no instituto federal do rio grande do norte - campus são gonçalo do amarante	Silva, E.M.C.
2019	M	Modelo para predição de risco de evasão na educação a distância utilizando técnicas de mineração de dados	Silva, D.R.
2019	G	Estudo da evasão de alunos de graduação utilizando Educational Data Mining	Fernades, K.C.
2019	T	Estratégia de ensino e aprendizagem ativa aplicada ao aprendizado de algoritmos e programação : identificação e análise da motivação dos estudantes	Franzen, E.
2019	T	Uma Abordagem Computacional de Predição de Desempenho Acadêmico de Estudantes em Cursos Online de Programação	Araúlo, F.F.
2018	D	Identificação de alunos com tendência à evasão nos cursos de graduação a distância por meio de mineração de dados educacionais	Schimitt, J.A.
2018	G	Sistema de apoio a decisão para prevenção da evasão nas instituições de ensino superior	Kraemer, J.R.
2018	D	Sistema de Mineração de Dados para Apoiar a Tomada de Decisão em uma Instituição de Ensino Superior: o problema da evasão escolar no IFTM	Araújo, E.O.
2017	D	Tratamento temporal em mineração de dados educacionais para fidelização de estudantes	Fazolin, K.
2016	M	Predição de evasão na educação a distâncis como subsídio a tomada de decisão	Sepúlvida, W.R.
2015	T	Predição do desempenho acadêmico de graduandos utilizando mineração de dados educacionais	Manhães, L.M.B.
2015	T	Sobre a evasão estudantil na esola politécnica da universidade de são paulo: identificação e possíveis causas	Fiorani, L. A.
2015	M	Uma abordagem temporal para identificação precoce de estudantes de graduação a distância com risco de evasão utilizando técnicas de mineração de dados	Santos, R.N.
2014	T	Sistema inteligente para a predição de grupo de risco de evasão discente	Martinho,V. R. C.

Fonte: do Autor

Tipo; T - Tese; D – Dissertação; G – Monografia Graduação.

A pesquisa descrita neste trabalho se constituiu no desenvolvimento de um simulador computacional para predição de evasão discente considerando um conjunto de atributos de entrada que permitem, através de metodologias de mineração de dados, predizer a evasão com certa acurácia e, assim, conseguir identificar a influência das variáveis na evasão e contribuir ao processo decisório.

A motivação desta pesquisa iniciou-se com uma pesquisa exploratória, na busca de referenciais teóricos sobre os temas principais abordados e registrados no Quadro 3. Na sequência, quanto a natureza da pesquisa trata-se de uma pesquisa descritiva, que após estabelecer o aporte teórico necessário para o desenvolvimento desta pesquisa, a estratégia eleita para trabalhar com as evidências levantadas, foi o

estudo de caso (EDC). De acordo com Yin (2005), a estratégia de estudo de caso se relaciona ao entendimento de fenômenos sociais complexos, possibilitando averiguar processos organizacionais e administrativos.

Um outro relevante aspecto que qualifica o estudo de caso “reside em sua capacidade de lidar com uma ampla variedade de evidências – documentos, artefatos, entrevistas e observações” (YIN, 2001, p.27).

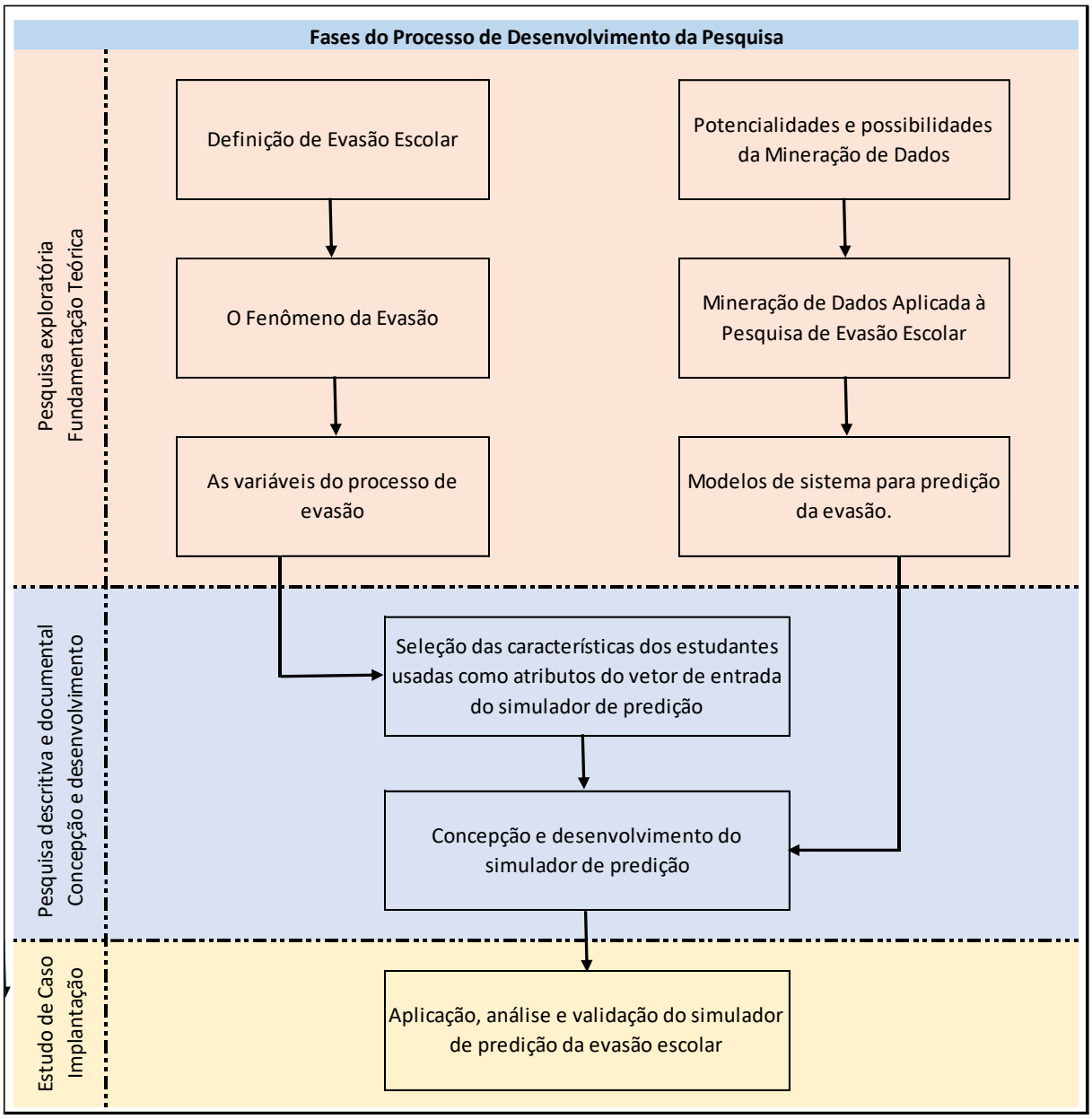
Nesse ponto, o estudo de caso

> Visa conhecer em profundidade o como e o porquê de uma determinada situação que se supõe ser única em muitos aspectos, procurando descobrir o que há nela de mais essencial e característico. O pesquisador não pretende intervir sobre o objeto a ser estudado, mas revelá-lo tal como ele o percebe. O estudo de caso pode decorrer de acordo com uma perspectiva interpretativa, que procura compreender como é o mundo do ponto de vista dos participantes, ou uma perspectiva pragmática, que visa simplesmente apresentar uma perspectiva global, tanto quanto possível completa e coerente, do objeto de estudo do ponto de vista do investigador (FONSECA, 2002, p. 33).

Para esta pesquisa, o estudo de caso pautou-se no levantamento sistemático das ferramentas sobre a evasão dos alunos, posteriormente foram estudados os documentos e dados que fundamentam os cursos de engenharia da instituição e concomitantemente foram realizadas as observações sob uma perspectiva interpretativa no próprio ambiente com conclusões aqui explicitadas apresentando o ponto de vista do pesquisador.

Quanto a coleta dos dados obtidos e suas bases documentais que fundamentam a concepção e desenvolvimento da solução proposta neste estudo, utilizou-se de dados disponíveis no sistema acadêmico da instituição pesquisada. Quanto a abordagem trata-se de uma pesquisa qualitativa. Para a análise dos dados obtidos utilizou-se de um simulador de predição de evasão desenvolvido nesta pesquisa. Na Figura 6 as fases do processo de desenvolvimento desta pesquisa:

FIGURA 6 - FASES DO PROCESSO DE DESENVOLVIMENTO DA PESQUISA



Fonte: Adaptado Martinho (2014)

O estudo de caso consiste na observação detalhada de um contexto ou indivíduo e suas relações e implicações, de uma única fonte de documentos ou de um acontecimento específico (BOGDAN et al.,1994).

3.2 ÂMBITO E UNIVERSO DA PESQUISA

Esta pesquisa foi desenvolvida no âmbito de uma Instituição de Educação Superior (IES), organização privada, com abrangência nacional, que oferece cursos de graduação na modalidade presencial, semipresencial e à distância, assim como cursos Lato Sensu e Stricto Sensu.

O Universo de interesse desta pesquisa são os alunos dos cursos de graduação na modalidade à distância de Engenharia Elétrica, Engenharia da Computação e Engenharia da Produção.

A escolha destes cursos e seus respectivos alunos se justifica pelos altos índices de evasão indicados pela Instituição comparados com outros cursos da própria instituição em seus controles internos.

O estudo da predição desenvolvido por esta pesquisa utilizou-se dos dados dos alunos dos cursos citados acima no período de 2015 a 2020.

Todos os alunos ingressam em um dos 7 (sete) processos seletivos realizados durante o ano pela instituição. Neste estudo de caso os dados dos alunos (variáveis/atributos) que ingressaram nos processos seletivos de 2015 e 2020 foram divididos em 3 bases de dados para o desenvolvimento do simulador, sendo:

1. Base de dados de treinamento;
2. Base de dados de testes, simulação;
3. Base de dados de validação da base.

Com o término das pesquisas deste trabalho, assim como com o término do desenvolvimento do simulador computacional para prever a evasão dos cursos de engenharia na modalidade de ensino à distância, este simulador ao alcançar a acurácia desejada, poderá ser aplicado em todos os cursos da instituição, assim como em outras organizações públicas ou privadas.

3.3 A COLETA DE DADOS

A instituição pesquisada autorizou o acesso as informações em março de 2020 após *email* encaminhado com a proposta dos estudos para elaboração desta dissertação.

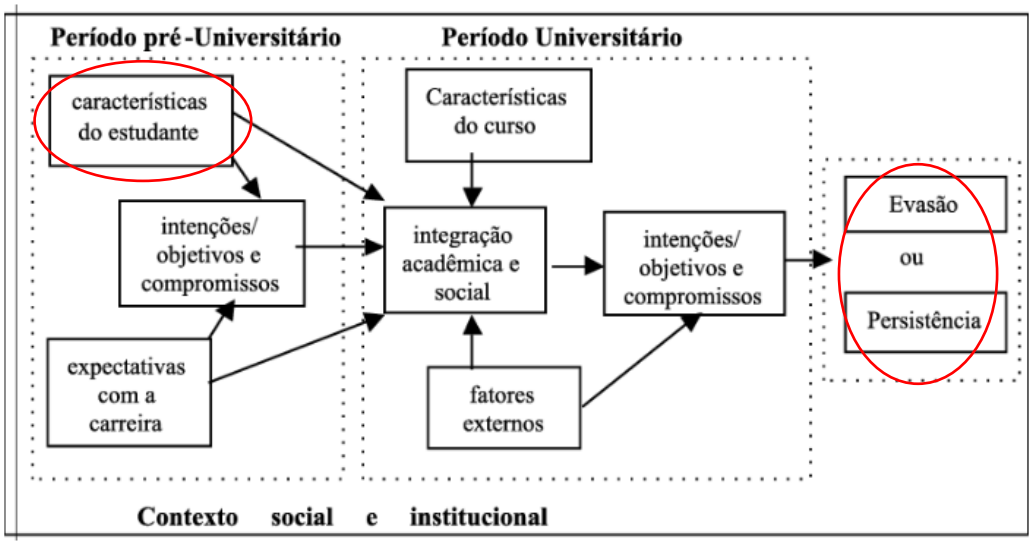
A solicitação contemplou acesso as bases de dados dos alunos dos cursos de alto índice de evasão, ao qual foram designados os cursos de graduação em Engenharia Elétrica, Engenharia da Computação e Engenharia da Produção no período de 2015 a 2020 na modalidade à distância.

A coleta dos dados, após análise documental (pesquisa documental) das informações constantes nas bases de dados permitidas dos alunos dos cursos referenciado acima, permitiu identificar atributos que pudessem ser utilizados em um modelo de predição de evasão, simulando no momento do ingresso do aluno na instituição se o mesmo possui ou não risco de evadir.

A autorização permitiu acesso a algumas informações (limitadas) constantes em um dos sistemas computacionais (sistema acadêmico “A”) de controle da instituição, ao qual a pesquisa teve acesso aos dados do aluno do seu processo seletivo, dados de desempenho acadêmico não foram fornecidos.

Com as informações de características do estudante fornecidas pela instituição é possível uma abordagem nesta pesquisa utilizando parte do modelo de evasão proposto por Tinto (1975), conforme Figura 7. Tinto propõe uma análise dividida em dois períodos, período pré-universitário e universitário em um contexto social e institucional. Esta pesquisa abordou, os dados sócios demográficos de entrada, do estudante, e, como saída, a evasão ou permanência do estudante (ATIVO OU EVADIDO).

FIGURA 7 - PROPOSTA DE MODELO EXPLICATIVO DA EVASÃO



Fonte: Tinto (1975)

Os dados sócios demográficos e de inscrição do ingresso de cada aluno nos processos seletivos promovidos pela instituição de ensino e fornecidos são os constantes no Quadro 5 a seguir:

QUADRO 5 – INFORMAÇÕES COLETADAS

Nº	Atributos/Características	Descrição
1	Data de Matrícula	Data da matrícula do aluno
2	Situação do Aluno	Situação Acadêmica do Aluno no momento do envio do arquivo para pesquisa
3	Código da Turma	Código da turma ao qual o aluno foi matriculado
4	Ano e Mês de Entrada	Ano e mês de entrada do aluno na respectiva turma
5	Nome da Turma	Indica o curso escolhido com o ano e mês de entrada
6	Situação da Turma	Situação da turma se ativa ou não no momento do envio do arquivo para pesquisa
7	Data Iníco da Turma	Data de início da turma para fins acadêmicos
8	Código do Polo	Código do Polo ao qual o aluno se matriculou
9	Estado Polo	Estado do Brasil onde o Polo localiza-se
10	Código Curso	Código do Curso escolhido pelo aluno
11	Nome do Curso	Nome do Curso escolhido pelo aluno
12	Escola	àrea do conhecimento da instituição ao qual o curso escolhido pelo aluno pertence
13	Nível do Curso	Grau acadêmico do curso
14	Modalidade do Curso	Modalidade do curso escolhido pelo aluno
15	Data da Última Atualização Ativa do Aluno	Última data de algum registro do aluno nos sistemas
16	Estado	Estado de moradia do aluno
17	Data Nascimento	Data de Nascimento do Aluno

Fonte - Autor

Após a análise da base de dados foi disponibilizado pela instituição uma planilha no software Excel com os dados acima de 25.354 alunos entre o período de 2015 a 2020.

Das informações recebidas, após análise, para o processo de predição, algumas informações foram consideradas para serem as variáveis de entrada do sistema de predição, ou seja, variáveis preditoras e sócio demográficas. As variáveis selecionadas foram: Situação do Aluno, Ano e Mês de Entrada, Código do Polo, Nome do Curso, Data da Última Atualização Ativa do Aluno, Estado e Data de Nascimento.

Esta pesquisa em suas análises de predição considerou as variáveis do processo seletivo do aluno e se o mesmo está ativo academicamente ou não com a instituição, ou seja, se evadiu e quando. Não são considerados dados acadêmicos do aluno nesta pesquisa para análises de predição, ou seja, o desempenho acadêmico nas disciplinas cursadas e se isto afetou a decisão de continuidade ou evasão, pois foram variáveis não fornecidas pela Instituição.

Na maioria dos trabalhos analisados que embasaram esta pesquisa as variáveis relevantes de análise de evasão eram acadêmicas em relação ao desempenho dos alunos em determinadas disciplinas e períodos do curso, conforme autores como Sepúlvida (2018); Silva (2019); Fernandes (2019); Araújo (2019); Fazolin (2017); Manhães et al. (2016); Fiorani (2015), entre outros.

Os dados recebidos na tabela de Excel, após a separação das variáveis escolhidas como preditoras, precisou passar por um processo de análise e validação das informações, excluindo-se aquelas que possuíam alguma inconsistência. Do total de informações recebidas de 25.354 alunos, 22.312 mantiveram-se para este estudo.

Para o uso destas informações de forma a utilizá-las como preditoras em um simulador de predição de evasão, com os recursos do pacote de software do *WEKA*, passaram por um processo de codificação para serem analisadas e processadas pelo modelo de algoritmo analisado pelo mesmo sistema.

Esse processo categoriza as variáveis de entrada e saída, para que possam ser inseridas nas linhas de comando de programação, ou seja, linguagem de programação. Esta categorização foi necessária para os anos que foram utilizados como base de treinamento do simulador de predição, os anos de 2015 e 2020, assim como para a base de diagnóstico (predição).

Neste trabalho, com a técnica de mineração de dados, com o uso de técnicas de algoritmos de classificação, busca-se obter em um primeiro momento um modelo particular para inferir a partir de um conjunto de variáveis preditoras (atributos de entrada) uma variável de saída (classe) que irá informar se o estudante evadiu ou não (fase do aprendizado do sistema). No segundo momento com outro conjunto de dados esta classificação é realizada pelo simulador que diagnostica se o aluno irá ou não evadir (fase do teste e validação).

O modelo de algoritmo parametrizado e analisados pelo simulador desenvolvido com os recursos do pacote de software do *WEKA* foi o *Naïve Bayes.* O *Naive Bayes* é uma técnica de classificação com suposição de independência entre os atributos preditores.

Do total de variáveis selecionadas para o modelo de predição, 9 foram selecionadas. Destas, 8 utilizadas como variáveis de entrada, sendo: Data de Nascimento, Nome do Curso, Mês e Ano de Entrada, Código do Polo, Estado e Região. Uma foi utilizada como variável de saída, a Situação do Aluno, ATIVO ou EVADIDO.

3.4 ANÁLISE DOS RESULTADOS

Alguns passos foram realizados nestes estudos e estão descritos em seus capítulos demonstrando o desenvolvimento e a execução do simulador de predição e a obtenção dos dados de saída (resultados), identificando os alunos com potencial de evasão ou não, simulações foram realizadas a fim de garantir níveis de acuracidade aceitáveis para um processo decisório.

Na etapa seguinte, com os resultados de saída do simulador de predição, realizou-se uma comparação entre a informação do simulador e os dados reais da planilha de Excel fornecida pela instituição, para identificar os acertos e erros de todos os alunos classificados como evadidos ou não.

Nas pesquisas utilizadas para embasar estes estudos, demonstradas no capítulo 2, foi realizada uma comparação também dos dados de acerto (acurácia) desta pesquisa com outras pesquisas. Importante considerar que não foi identificada nenhuma pesquisa similar a esta em seus dados de entrada (somente dados do processo seletivo) para o processo de predição na área de educação superior na modalidade à distância. A maioria das pesquisas comparadas levam em consideração dados acadêmicos (desempenho acadêmico), ou dados sócios econômicos, que as vezes também são completadas com parte dos dados do processo seletivo do aluno.

Por fim, com as comparações realizadas foi possível inferir em que medida o simulador desenvolvido pode contribuir na predição da evasão discente e, assim, ajudar os gestores em seus processos decisórios a fim de encontrar a máxima

eficiência e eficácia da organização com a identificação das variáveis com mais impacto na evasão.

4 CARACTERIZAÇÃO DA PESQUISA

Neste capítulo aborda-se a concepção e desenvolvimento do simulador computacional para predizer a evasão, tendo como base uma ferramenta de mineração de dados (*WEKA).* Também descrevendo as potencialidades do simulador, desde o desenvolvimento das bases de dados ao diagnóstico da evasão.

4.1 CONCEPÇÃO DO SIMULADOR

Silva (2019) em sua pesquisa elencou as principais ferramentas de mercado para uso da mineração de dados e suas características potenciais, aqui elaborou-se o quadro 6 representando as ferramentas elencadas

QUADRO 6 - FERRAMENTAS DE MINERAÇÃO DE DADOS

Nome da Ferramenta	Descrição
Clementine 7	É uma ferramenta proprietária, lider de mercado, desenvolvida pela SPSS e além de outras facilidades suporta o processo CRISP-DM
SAS Enterprise Miner Suite	É uma ferramenta proprietária bem conhecida desenvolvida pela empresa SAS 8. Possui diversos recursos, entre eles integração de código aberto com o R e módulos para trabalhar em todas as etapas do processo de DM
Pimiento	Ferramenta open source para mineração de textos
SAS Text Miner	Ferramenta para mineração de texto, desenvolvida pela empresa SAAS
Oracle Data Mining (ODM) e IBM Intelligent Miner	desenvolvidas pelas respectivas empresas (Oracle e IBM) para serem usadas em seus banco de dados: Oracle e DB2.
KNIME	é uma ferramenta de código open source, licença GPL, que implementa o paradigma de pipelining de dados. Possui um ambiente intuitivo e fácil de usar e realiza mineração de grandes volumes de dados.
Statistical Package for the Social Sciences - SPSS	Software usado para análises estatísticas que pode ser usado para mineração de dados e de textos. Os dados são definidos em forma de planilha Excel (xls), Access e outros. Necessita-se declarar cada variável usada, tipo de dados (número, strings), tamanho da variável, entre outras informações. O SPSS possibilita a realização de análise estatística, geração de gráficos, regressão linear, transformação de variáveis, geração de quartis, entre outras atividades
RapidMiner	Ferramenta desenvolvida pela empresa que leva o mesmo nome. Permite a aplicação de diversas técnicas para DM e mineração de textos. Fornece um ambiente integrado para preparação dos dados, possui interface gráfica que trabalha com diversos operadores para realizar a descoberta do conhecimento.
Waikato Environment for Knowledge Analysis - WEKA	Esta ferramenta foi desenvolvida pela Universidade de Waikato, na Nova Zelândia, na linguagem Java, multiplataforma e é distribuída através de licença GNU - GPL, e pode ser obtida por download direto do site da universidade . A ferramenta possui técnicas para pré-processamento de dados, classificação, regressão, regras de associação e visualização

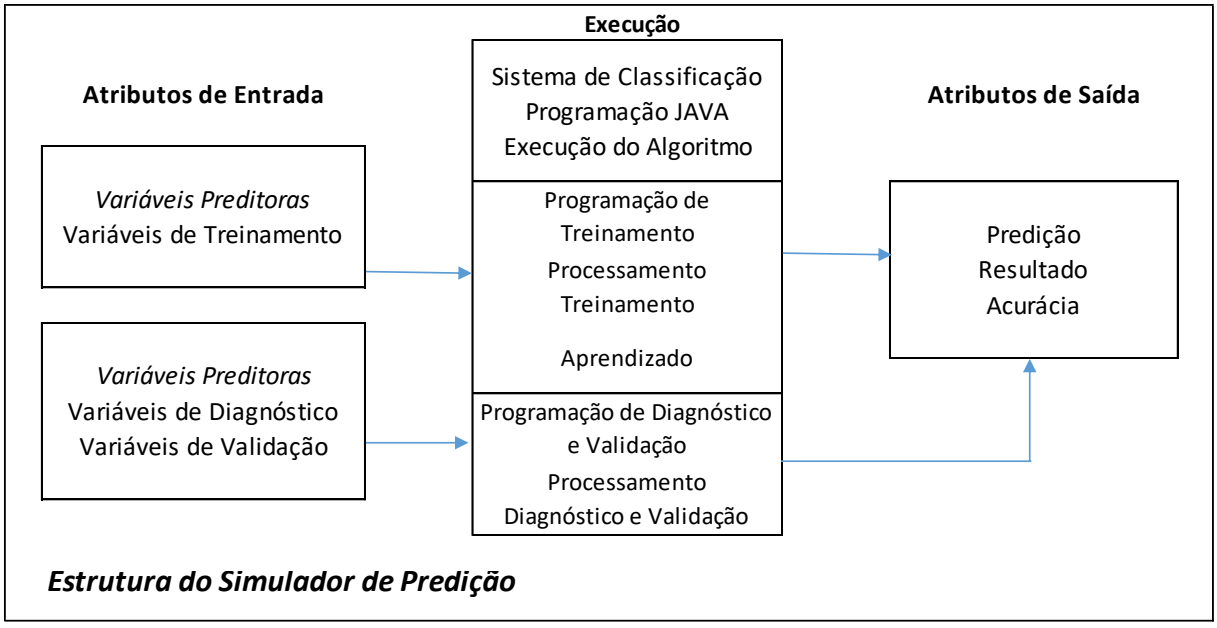
Fonte – Adaptado Silva(2019)

O mercado disponibiliza diversas ferramentas de mineração de dados, mas cada uma possui suas características e aplicabilidades com recursos específicos de classificação, algumas para sua aquisição com custo financeiro e outras de acesso livre.

Conforme referenciado no capítulo 1 desta pesquisa a ferramenta de base escolhida para o desenvolvimento do simulador computacional de predição é o *WEKA,* a escolha justifica-se por ser uma ferramenta de acesso financeiro gratuito, enquanto as outras com necessidade de pagamento financeiro. Outra escolha justifica-se pelo código aberto, permitindo como base para o desenvolvimento de simuladores desenvolvido pelo próprio usuário. Outro ponto é uma ferramenta de base que possui em seus pacotes de software os algoritmos necessários para o desenvolvimento de simuladores.

A Figura 8 representa abaixo a concepção macro da estrutura de funcionamento do simulador computacional desenvolvido para a predição da evasão com o uso do pacote de software do *WEKA* que este estudo se propôs.

FIGURA 8 - VISÃO MACRO FUNCIONALIDADES SIMULADOR DE PREDIÇÃO NO *WEKA*



Fonte – do Autor

O simulador desenvolvido, em sua programação, trabalha com variáveis de entrada (preditoras) de treinamento, diagnóstico e validação para a predição da evasão. Neste modelo desenvolvido as variáveis de entrada, suas descrições e valores possíveis são observados no Quadro 7:

QUADRO 7 - VARIÁVEIS DE ENTRADA

Nº	Atributos/Características	Descrição	Valores Possíveis
1	Mês	Mês de início do curso	01 a 12
2	Ano	Ano de início do curso	2015 a 2020
3	Polo	Código do Polo ao qual o aluno se matriculou	00001 a 99999
4	Estado	Estado do Brasil de origem do Aluno	Sigla do Estado
5	Curso	Nome do Curso escolhido pelo aluno	EC, EL e EP
6	Evasão	Mês de evasão do aluno	01 a 99
7	Idade	Idade do Aluno no momento da matrícula	01 a 99
8	Região	Região do Estado do Aluno	1 a 5
9	Situação	Situação Acadêmica do Aluno	Ativo e Evadido

Fonte: Autor

Na saída o sistema indica se o aluno está no grupo de risco e prediz sua evasão ou não (ativo ou evadido).

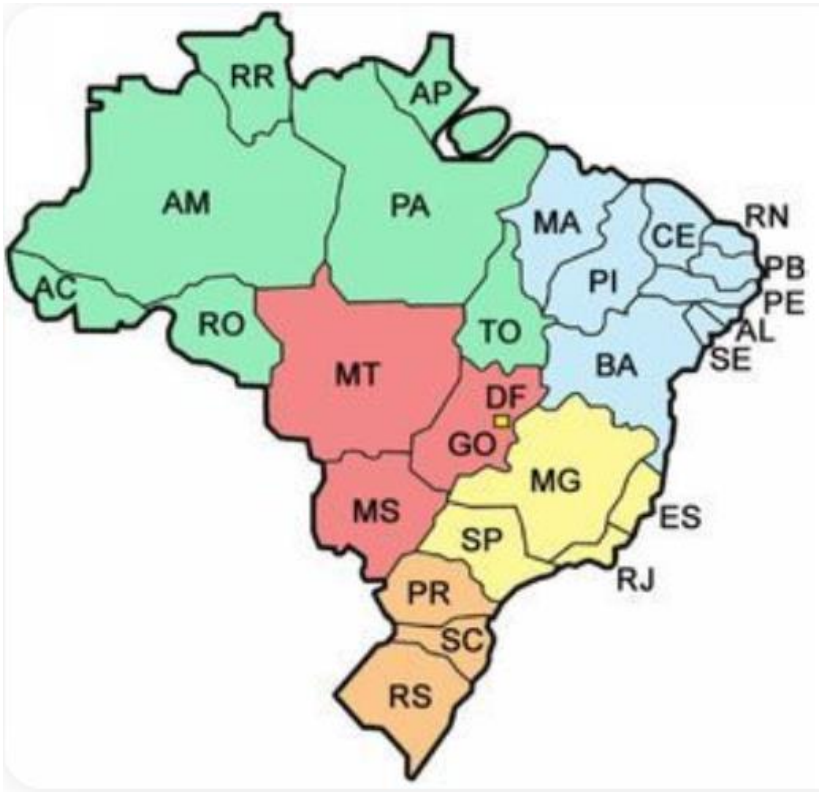
Estas variáveis foram extraídas e tratadas da base de dados fornecida pela IES estudada. Das 17 variáveis fornecidas através de planilha de Excel (Quadro 5), 9 foram selecionadas para serem analisadas pelo simulador. Destas 8 são variáveis de entrada e 1 de saída. As variáveis selecionadas permitem realizar no momento do ingresso do aluno na IES estudada a simulação de predição de evasão em relação a localização geográfica do aluno (cidade, estado e região), curso escolhido, idade do aluno e polo de estudo.

O simulador desenvolvido tem aplicação para a área de educação, ensino superior, na modalidade à distância, para qualquer organização que atua neste segmento. Seus potenciais mais expressivos são para organizações na modalidade a distância que operam em mais de um Estado do Brasil, assim identificando características do fenômeno da evasão em cada área (demográfico social), não somente para compreender o fenômeno, mas também quando confrontado com

outras informações dos seus bancos de dados em relação ao aluno ou estratégias de posicionamento, acadêmicas, financeiras entre outras.

O Brasil possui 26 estados mais o Distrito Federal, 27 unidades da federação. O Brasil é um país continental, com características políticas, sociais, econômicas e etc distintas em certas localidades. Abaixo Figura 8 representando as unidades federativas do Brasil.

Figura 9 - Mapa do Brasil - Estados



Fonte – Pinterest (https://br.pinterest.com/pin/584412489124287258/)

O simulador desenvolvido permite relacionar as variáveis: região, estado, cidade, curso, polo, idade, mês, ano e polo e predizer seu potencial de evasão com base em dados históricos e de aprendizado do algoritmo. O simulador permite inserir mais variáveis quando necessário para o desenvolvimento de novos cenários para o processo decisórios.

Este relacionamento das variáveis, pode indicar não somente se o aluno está ou não propenso a evasão, mas apresentar informações importantes aos gestores das variáveis que impactam no fenômeno da evasão e assim contribuem no aprimoramento do modelo decisório organizacional.

Um exemplo é observar que alunos de uma determinada cidade do Brasil possuem mais propensão a evasão do que alunos em outras cidades em condições similares das variáveis inseridas no simulador.

O conhecimento gerado pode modificar ou aprimorar estratégias comerciais, estratégias de precificação, estratégias de marketing, estratégias pedagógicas entre outras, ou seja, pode potencializar a eficiência e eficácia dos resultados organizacionais em várias dimensões.

Interessante também que o sistema permite realizar simulações de cenários para cada estado do Brasil, isto também tem seu valor na etapa do planejamento organizacional, pois permite direcionar estratégias de captação que atendam os principais objetivos e metas corporativas.

A contribuição maior do simulador e seu propósito é compreender o fenômeno da evasão, seja de forma simulada, seja no momento da matrícula do aluno, seja durante o seu curso e assim mitigar a probabilidade de o aluno evadir aumentando as taxas de alunos concluindo seu curso, seus sonhos e promovendo uma nação mais justa e com potencial de desenvolvimento.

4.2 PACOTE DE SOFTWARE DO *WEKA*

Para o desenvolvimento do simulador de predição, tendo como base o pacote de software do WEKA, utilizou-se um computador com microprocessador intel i5, 64 bits, CPU 2Ghz e 4GB de memória RAM.

O acesso e download do pacote de software do WEKA realizou-se através do site: https://www.cs.waikato.ac.nz/ml/weka/.

De início, para o desenvolvimento do simulador, é importante que as informações estejam organizadas, podem ser através de uma planilha, estrutura de dados ou mesmo banco de dados.

O formato dos arquivos de entrada que o *WEKA* utiliza-se para organizar os dados tem o nome de ARFF (*Attribute-Relation File Format)*. O arquivo a ser gerado precisa conter as informações do domínio do atributo, valores que as instâncias podem representar a classe.

O arquivo ARFF é dividido em duas partes: a primeira com a lista que define os atributos (tipo do atributo/valores), a segunda é composta pelos dados.

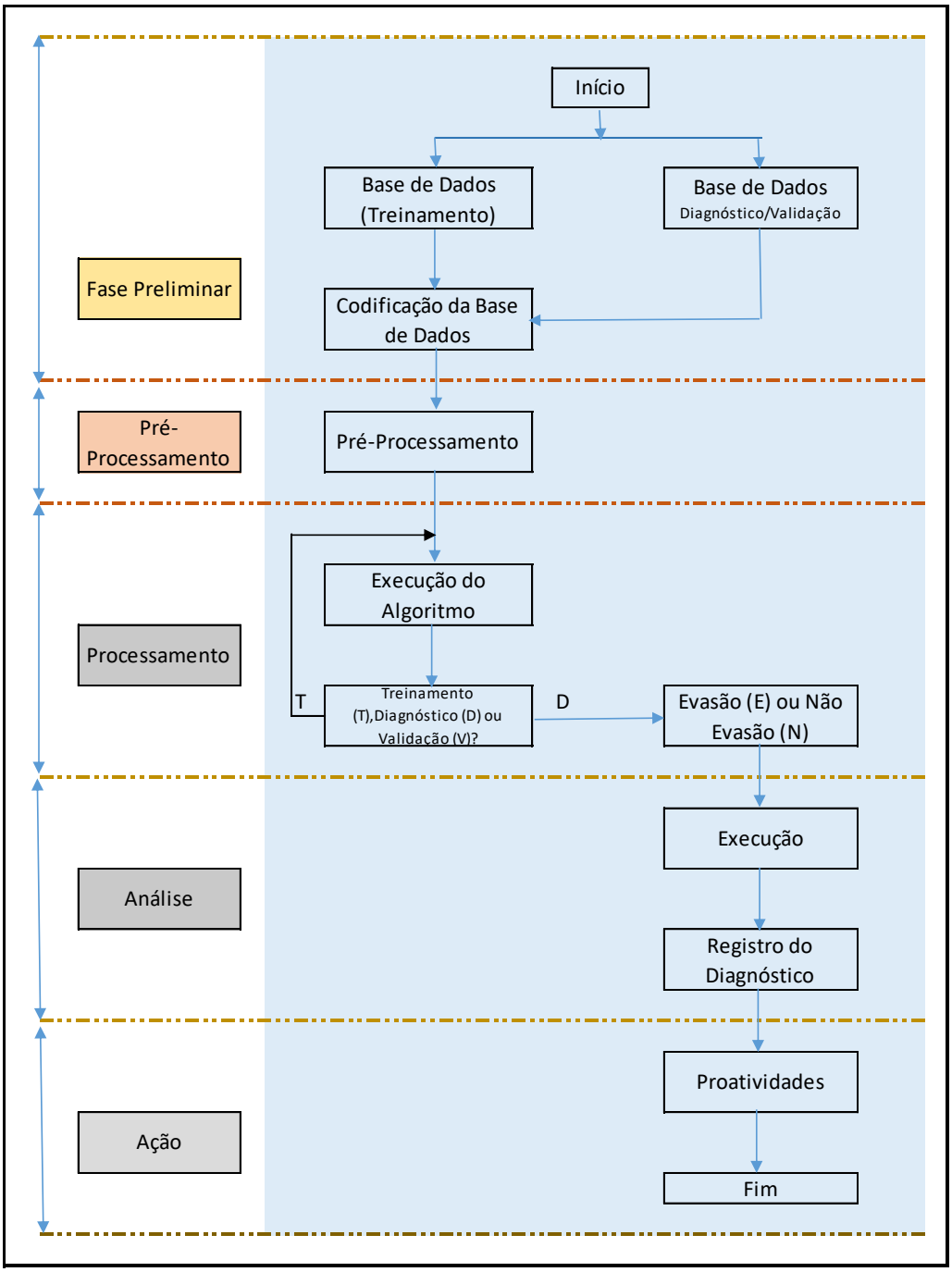
Uma característica do arquivo ARFF são as marcações que precisam ser indicadas para seu correto funcionamento, que são 3: *@atribute, @relation e @data*.

Após a instalação do *WEKA*, não é necessário fazer nenhuma configuração complementar para seu funcionamento.

4.3 ESTRUTURA DO SIMULADOR PREDITIVO

Na Figura 10 observa-se a estrutura e as fases do desenvolvimento do simulador computacional de predição:

FIGURA 10 - FLUXOGRAMA DA ESTRUTURA DE DESENVOLVIMENTO DO SIMULADOR PREDITIVO



Fonte: Adaptado Martinho (2014)

A implementação de um simulador computacional capaz de predizer a evasão é constituída em um processo de 5 fases, sendo:

a) Fase Preliminar

O momento do acesso as informações do estudo de caso da instituição pesquisada. Cabem aqui também a análise documental das informações e suas

validações para uso em um possível modelo para predizer a evasão. Em seguida, realiza-se uma “limpeza” nos dados selecionados e úteis, excluindo-se os dados em não conformidade. Os dados são categorizados em variáveis (atributos) de entrada e de saída no modelo escolhido para análise, assim como os dados escolhidos que irão compor a base de treinamento, diagnóstico (teste) e validação do simulador de predição.

b) Fase Pré-Processamento

As variáveis (atributos) de entrada e saída selecionadas devem ser imputadas no simulador de predição com base no pacote de software *WEKA,* com suas linhas de comando e codificação interpretadas pelo sistema, deixando as informações dentro dos padrões necessários para que o processamento possa ocorrer, interpretando o que são dados de treinamento, diagnóstico (teste) e validação.

c) Fase Processamento

O simulador executa as informações (variáveis/atributos) inseridas nas linhas de comando, tanto na fase do treinamento, diagnóstico e validação e apresenta os resultados conforme a modelagem algorítmica classificatória escolhida.

d) Fase Análise

Com os resultados apresentados pelo simulador o analista ou gestor pode realizar suas interpretações e tomar ações e decisões condizentes com as estratégias realinhadas aos resultados. Importante observar neste momento que os resultados de evasão indicados podem ser analisados e interpretados as variáveis de entrada que mais estão influenciando na evasão.

e) Fase Ação

Uma quinta fase que pode ser elencada, após as análises, chamada de ação. Trata-se de uma revisão ao planejamento estratégico e operacional. Esta revisão possibilitará que a organização reveja seus processos internos e externos e assim opere de tal forma que as variáveis que impactam no aumento da evasão sejam permanentemente analisadas, ou seja, uma organização trabalhando de forma proativa ou preventiva, mitigando evasão e trabalhando com maior eficiência e eficácia nos investimentos realizados.

4.4 TRATAMENTO DOS DADOS – FASE PRELIMINAR, PRÉ-PROCESSAMENTO E PROCESSAMENTO

As informações recebidas da instituição, que foram encaminhadas em formato de planilha eletrônica (excel), foram tratadas para identificar as inconsistências que poderiam impactar no funcionamento da linguagem de programação do sistema de predição (algoritmo), ou até mesmo na etapa das análises. No Quadro 8 partes da planilha recebida para ilustrar o arquivo:

QUADRO 8 - PLANILHA (PARTE) COM OS DADOS RECEBIDOS DA INSTITUIÇÃO

Data de matrícula	Situação	Código da turma	Nome da turma	Data de início da turma	Turno da turma	Código do pólo	Estado do aluno	Nível do curso	Modalidade do curso	Última data da situaçao ativo do aluno
02/02/2015	CANCELADO	245266	2015/02 GD ENGENHARIA DE PRODUÇÃO	02/02/2015	NÃO APLICÁVEL	944	ES	GRADUAÇÃO	DISTÂNCIA	0
02/02/2015	ATIVO	260149	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NOITE	1966	GO	GRADUAÇÃO	DISTÂNCIA	24/03/2020
02/02/2015	ABANDONO CONFIRMAÇÃO	279677	2015/02 GD ENGENHARIA DE COMPUTAÇÃO	02/02/2015	NÃO APLICÁVEL	1934	MG	GRADUAÇÃO	DISTÂNCIA	01/03/2016
02/02/2015	CANCELADO	262060	2015/02 GD ENGENHARIA DE PRODUÇÃO	02/02/2015	NÃO APLICÁVEL	155	BA	GRADUAÇÃO	DISTÂNCIA	0
02/02/2015	CANCELADO	265024	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NÃO APLICÁVEL	56	PR	GRADUAÇÃO	DISTÂNCIA	0
02/02/2015	ATIVO	279641	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NOITE	107	PR	GRADUAÇÃO	DISTÂNCIA	24/03/2020
02/02/2015	ATIVO	279594	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NOITE	909	RJ	GRADUAÇÃO	DISTÂNCIA	24/03/2020
02/02/2015	ATIVO	268851	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NOITE	783	PA	GRADUAÇÃO	DISTÂNCIA	24/03/2020
02/02/2015	TRANCADO	268445	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NÃO APLICÁVEL	491	RS	GRADUAÇÃO	DISTÂNCIA	22/03/2018
02/02/2015	CANCELADO	261219	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NOITE	127	MA	GRADUAÇÃO	DISTÂNCIA	0
02/02/2015	TRANSFERIDO INTERNO	250216	2015/02 GD ENGENHARIA DE PRODUÇÃO	02/02/2015	NOITE	966	PR	GRADUAÇÃO	DISTÂNCIA	21/08/2016
02/02/2015	CANCELADO	279714	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NÃO APLICÁVEL	214	BA	GRADUAÇÃO	DISTÂNCIA	0
02/02/2015	CANCELADO	279722	2015/02 GD ENGENHARIA DE PRODUÇÃO	02/02/2015	NÃO APLICÁVEL	986	SP	GRADUAÇÃO	DISTÂNCIA	0
02/02/2015	ABANDONO CONFIRMAÇÃO	269851	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NÃO APLICÁVEL	222	PR	GRADUAÇÃO	DISTÂNCIA	22/02/2016
02/02/2015	CANCELADO	279732	2015/02 GD ENGENHARIA ELÉTRICA	02/02/2015	NÃO APLICÁVEL	440	MG	GRADUAÇÃO	DISTÂNCIA	0
02/02/2015	CANCELADO	250216	2015/02 GD ENGENHARIA DE PRODUÇÃO	02/02/2015	NOITE	966	PR	GRADUAÇÃO	DISTÂNCIA	0

Fonte: do Autor

Das informações constantes na planilha, referente aos 25.354 alunos, 3.042 alunos foram desconsiderados, os principais motivos foram: campos com falta de informações, turmas canceladas, alguns alunos constantes na planilha eram da modalidade presencial (este estudo concentrou-se na modalidade a distância), dados inconsistentes como data de nascimento, situação acadêmica, falta de dados, entre outros. Para o estudo foram considerados 22.312 alunos, distribuídos da seguinte forma, conforme Tabela 1:

TABELA 1 - QUANTIDADE DE ALUNOS MATRICULADOS POR CURSO E ANO – À DISTÂNCIA

Nome do Curso	ATIVO	EVADIDO	Total Geral	ATIVO %	EVADIDO %
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO	1071	2259	3330	32,2%	67,8%
BACHARELADO EM ENGENHARIA DE PRODUÇÃO	3125	5796	8921	35,0%	65,0%
BACHARELADO EM ENGENHARIA ELÉTRICA	3801	6260	10061	37,8%	62,2%
Total Geral	**7997**	**14315**	**22312**	**35,8%**	**64,2%**

Fonte: do Autor

Na planilha recebida, dados como nome, endereço específico e outros que caracterizam informações pessoais do aluno não foram encaminhadas para esta pesquisa pela IES.

Dos 17 atributos referentes a cada aluno foi necessário adequar alguns campos para o uso no sistema de predição concebido. As, adequações realizadas foram:

a) Na planilha constava o mês e ano da entrada do aluno em um único campo da tabela, foi necessário desmembrar esta informação em duas colunas para o sistema de classificação (algoritmo), ficando uma coluna com mês de entrada e outra com o ano de entrada;
b) Na coluna situação do aluno, constavam classificações de cada aluno como Ativo, Cancelado, Trancado, Transferência interna, Falecido e Formado. Foi adequado o campo para que somente ATIVO e EVADIDO constassem como dados válido;
c) O campo data de nascimento, que estava no formato de data foi convertido para idade nominal do aluno (número absoluto);
d) Foi criado um campo para especificar a região do Brasil de moradia do aluno (1- Sul, 2- Sudeste, 3-Norte, 4- Centro Oeste e 5- Nordeste) para contribuir no processo de predição do sistema;
e) Foi criado o campo meses evadidos, na planilha constava no formato data a informação e foi convertida em meses absolutos de evasão.
f) Os cursos que estavam com seu nome na forma completa na tabela foram adequados para sigla: Engenharia de Produção = EP; Engenharia da Computação = EC e Engenharia Elétrica = EE.
g) O sequenciamento das colunas na planilha também foi alterado, para permitir que a última coluna da planilha fosse a situação do aluno (ativo ou evadido), que é interpretado pelo simulador preditivo como a classe a ser analisada.

Ao final deste processo, a planilha tratada ficou com 9 atributos também no formato de uma planilha de excel com os 22.312 alunos, observando-se que 8 destes atributos serão os preditores (de entrada) do simulador de predição. Abaixo modelo na Tabela 2:

TABELA 2 - PLANILHA FINAL AJUSTADA (ATRIBUTOS FINAIS-PREDITORES) (EXEMPLO)

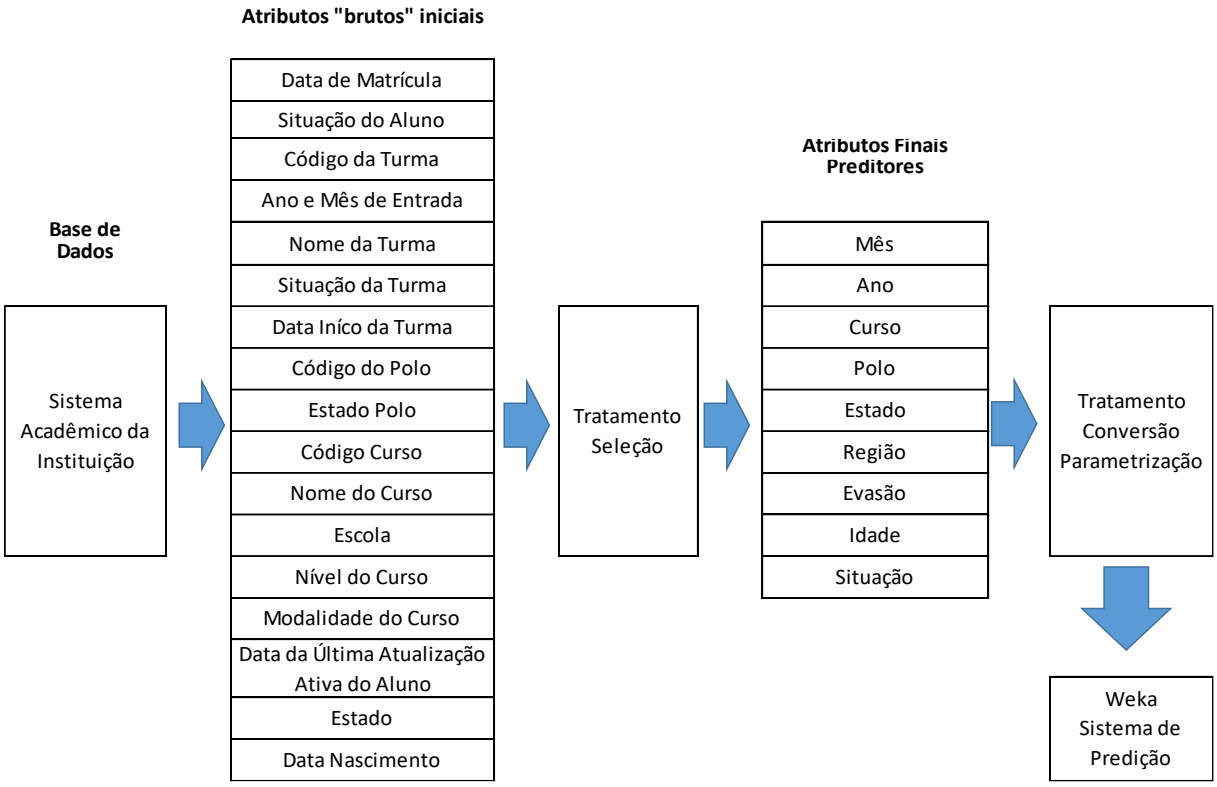
Mês	Ano	Curso	Idade	Polo	Evasão	Estado	Região	Situação
2	2015	EP	36	2045	6	AP	5	EVADIDO
2	2015	EP	42	127	28	MA	3	EVADIDO
2	2015	EP	33	127	24	MA	3	EVADIDO
2	2015	EP	27	68	7	PR	1	EVADIDO

Fonte: do Autor

Os outros atributos recebidos foram desconsiderados para este estudo.

A Figura 11 apresenta a macro visão do processo de tratamento das informações recebidas da Instituição de Ensino para inserção no simulador de predição baseado na ferramenta *WEKA*:

FIGURA 11 - VISÃO MACRO DO PROCESSO PRELIMINAR DO SIMULADOR DE PREDIÇÃO



Fonte: do Autor.

Após a planilha ajustada nos atributos (variáveis) preditores, realizou-se a conversão do arquivo xls para a extensão CSV (*comma separated values*), modelo de extensão de arquivo para preparar os dados para inserir no simulador. A conversão foi realizada com recursos do próprio Excel. O arquivo na sequência foi aberto no software Word para editar/substitiur as vírgulas, espaços e inserir o cabeçalho com as linhas de programação necessárias para a operação do pacote de software do *WEKA* (modelo ARFF). Na Figura 12 observa-se o formato final deste arquivo antes da sua inserção no sistema *WEKA*:

FIGURA 12 – ARQUIVO FINAL - NOME: EVASÃO

```
@relation evasão

@attribute mês {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
@attribute Ano {2015, 2016, 2017, 2018, 2019, 2020}
@attribute Curso {EP, EC, EE}
@attribute Idade {16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29,
30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47,
48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80}
@attribute Polo real
@attribute Estado {RS, SC, PR, SP, ES, MG, RJ, AC, AL, AM, AP, BA, CE,
DF, GO, MA, MS, MT, PA, PB, PE, PI, RN, RO, RR, SE, TO}
@attribute Região {1, 2, 3, 4, 5}
@attribute Evasão {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
52, 53, 54, 55, 56, 57, 58, 59, 60, 61}
@attribute Situação {ATIVO, EVADIDO}

@data
2,2015,EP,27,491,RS,1,61,ATIVO
2,2015,EP,41,817,MG,2,61,ATIVO
2,2015,EP,36,2045,AP,5,6,EVADIDO
2,2015,EP,42,127,MA,3,28,EVADIDO
2,2015,EP,33,127,MA,3,24,EVADIDO
```

Fonte: do Autor.

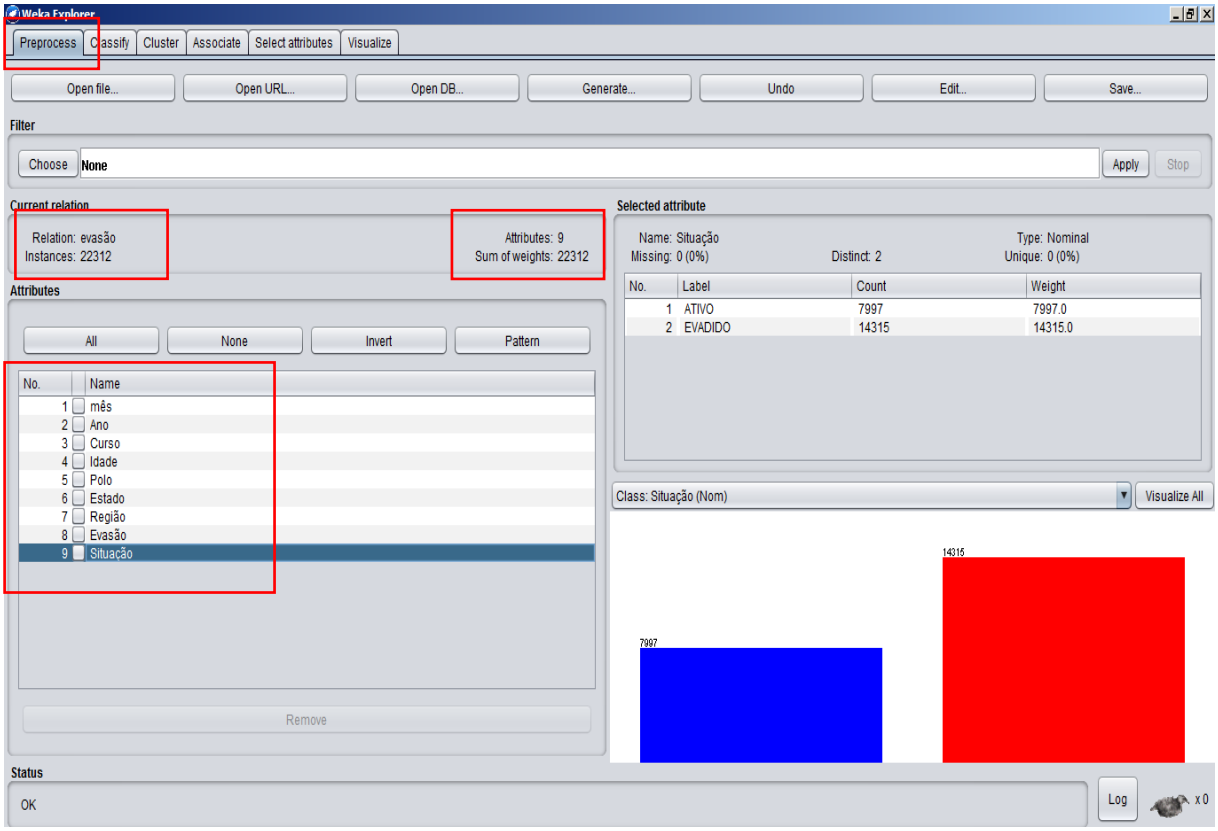
Na Figura 12 observa-se o modelo de arquivo e linhas de comando para o correto funcionamento do simulador de predição. São 3 os cabeçalhos de programação de comandos necessários, sendo:

a) Declaração de Relação: o comando**:** *@relation* <nome do arquivo>, nesta pesquisa o nome do arquivo foi Evasão;
b) Declaração de Atributos: Os atributos são declarados através de uma sequência ordenada de *@attributes***.** Cada atributo no conjunto de dados deve possuir sua própria declaração usando @attribute que identifica unicamente o nome deste atributo e o tipo de dado. A ordem em que são declarados indicam a ordem em que aparecem no conjunto de dados;
c) Declaração do Conjunto de Dados: Os dados são as instâncias e são declarados um por linha e deve-se separar os atributos com vírgula.

4.5 TRATAMENTO DOS DADOS - ANÁLISE E AÇÃO

Na Figura 13, após a leitura do arquivo Evasão o simulador computacional realizou as classificações iniciais dos atributos e instâncias (dados dos alunos), a figura apresenta umas das telas possíveis de serem analisadas.

FIGURA 13 - TELA SIMULADOR DE PREDIÇÃO - ARQUIVO EVASÃO

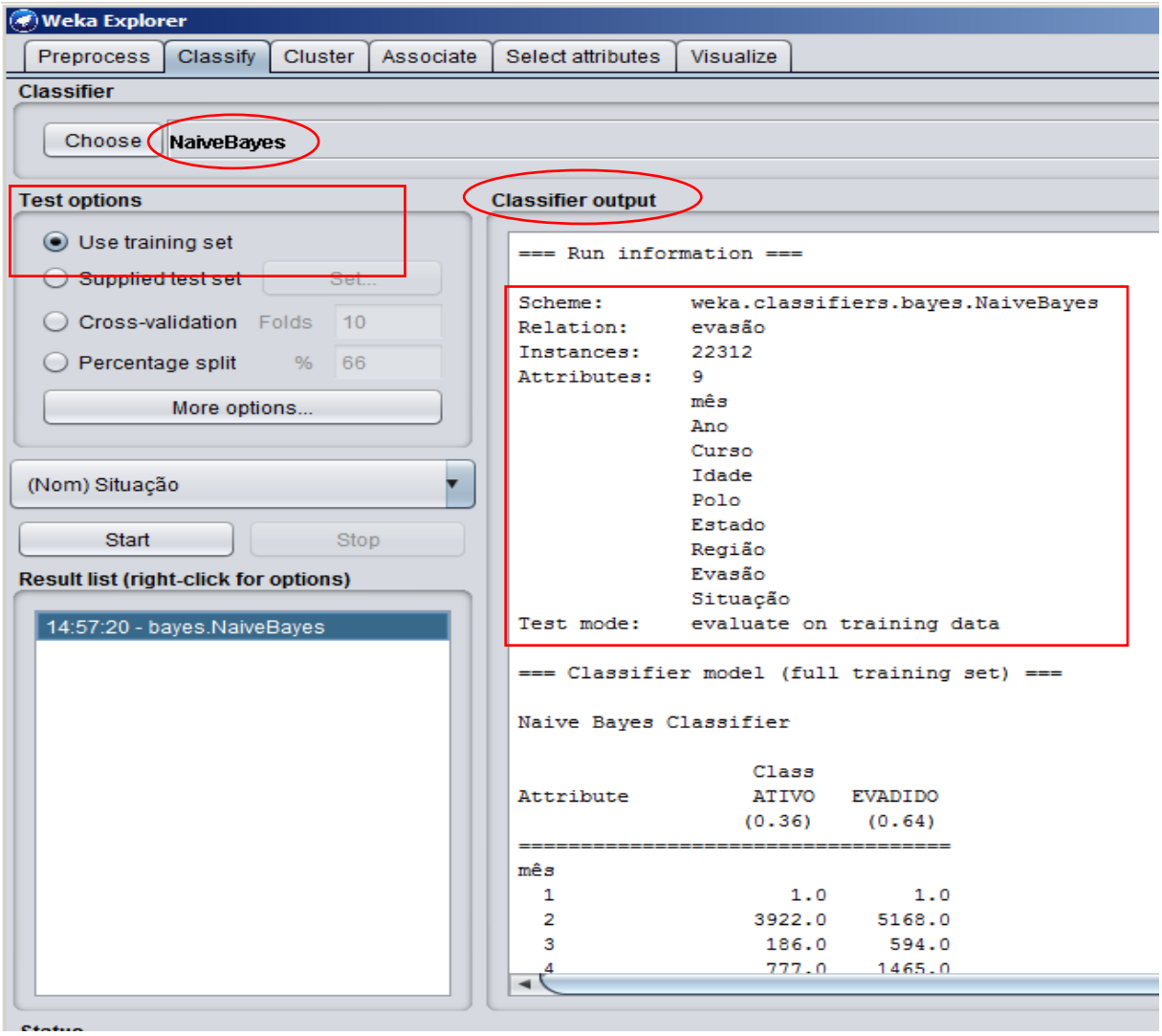


Fonte: do Autor.

Com as informações inseridas no simulador e a parametrização dos dados de entrada inicia-se o pré-processamento no simulador de predição para a validação dos dados inseridos da planilha do Excel.

Neste momento da execução do simulador ocorre a parametrização do Algoritmo Classificador com a escolha do classificador *Naives Bayes.* Na Figura 14 a imagem da tela da configuração do algoritmo:

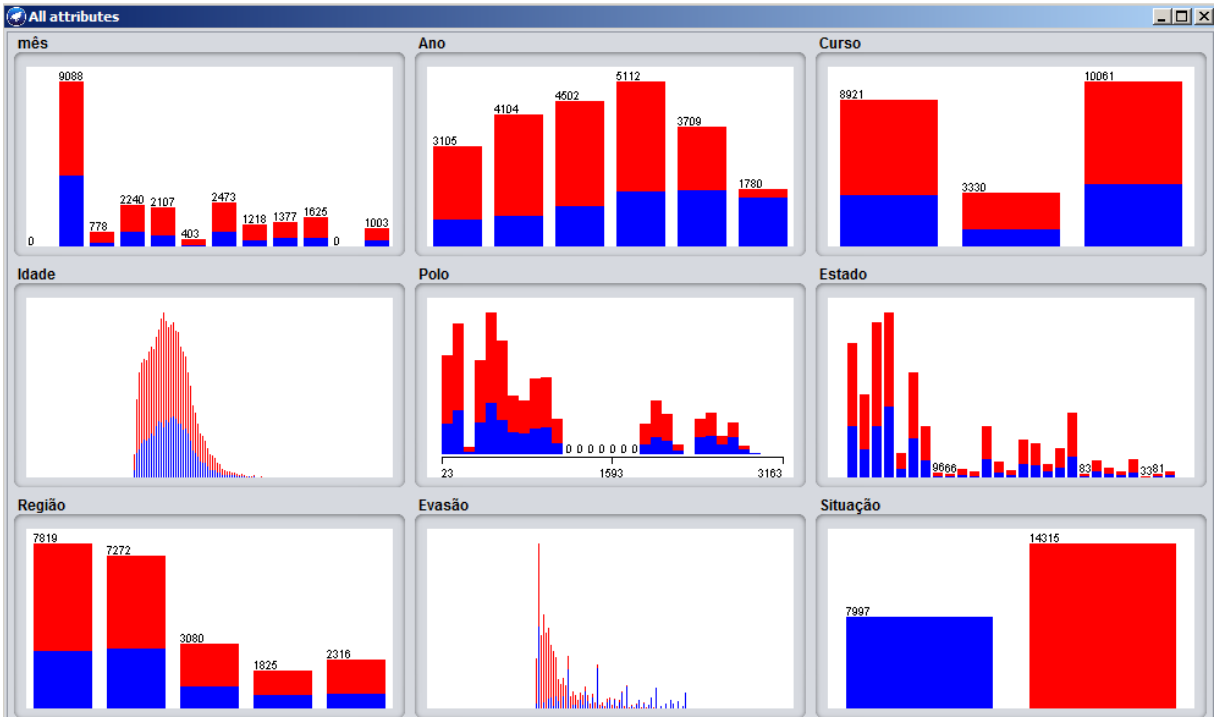
FIGURA 14 - TELA SIMULADOR DE PREDIÇÃO ALGORITMO CLASSIFICADOR



Fonte: do Autor.

Nesta mesma tela do simulador, são mostradas as execuções do algoritmo e os resultados correspondentes do simulador de predição, indicando os alunos que irão evadir e os alunos que continuarão na instituição. Na Figura 15 a representação das análises de forma gráfica que o simulador de predição permite utilizar:

FIGURA 15 - TELA SIMULADOR DE PREDIÇÃO ANÁLISE GRÁFICA



Fonte: do Autor.

Os recursos gráficos que o simulador de predição permite, facilita análises complementares e auxilia os gestores no processo decisório.

O simulador permite criar uma matriz com todos os atributos de entrada e realizar cruzamento entre os mesmos, criando possibilidades de análise e recursos gráficos.

5 ANÁLISE DOS RESULTADOS

Neste capítulo, evidenciam-se os resultados obtidos com a aplicação do simulador computacional de predição de evasão com as bases de dados da IES estudada.

A análise dos resultados desta pesquisa ocorreu em 2 etapas:

a) Primeira: analisando os resultados dos 8 atributos tratados na etapa de pré-processamento;
b) Segunda: analisando os resultados da predição do simulador concebido;

5.1 ANÁLISE DOS RESULTADOS – VARIÁVEIS DE ENTRADA

Com o tratamento das informações recebidas pela instituição e organizadas nos 9 atributos de análise observados no Quadro 9, foi possível identificar algumas análises e cenários de resultados.

A instituição pesquisada possibilitou acesso a dados dos anos de 2015 a 2020 dos cursos de Engenharia Elétrica, Engenharia da Computação e Engenharia da Produção que a mesma oferta todos os anos. Houve um crescimento significativo na quantidade de alunos matriculados entre 2015 e 2018, com redução de matrículas em 2019 conforme Tabela 3. Em 2020, como o arquivo utilizado para esta pesquisa recebeu-se em março, só existem matrículas dos meses de janeiro e fevereiro para o ano de 2020.

TABELA 3 - ALUNOS ATIVOS E EVADIDOS POR ANO (2015 A 2020) QUANT. E %

Ano de Entrada	ATIVO	EVADIDO	Total Geral	% EVADIDO
2015	836	2269	3105	73,1%
2016	964	3140	4104	76,5%
2017	1242	3260	4502	72,4%
2018	1688	3424	5112	67,0%
2019	1747	1962	3709	52,9%
2020	1520	260	1780	14,6%
Total Geral	**7997**	**14315**	**22312**	**64,2%**

Fonte: do Autor

A taxa média da instituição neste período de evasão foi de 64,2%.

Conforme Tabela 4 o curso de Engenharia Elétrica tem mais representatividade numérica de alunos evadidos em relação aos outros cursos.

TABELA 4 - QUANTIDADE DE ALUNOS ATIVOS E EVADIDOS POR ANO E CURSO – 2015 A 2020

Curso / Situação Aluno	2015	2016	2017	2018	2019	2020	Total Geral
ATIVO	**836**	**964**	**1242**	**1688**	**1747**	**1520**	**7997**
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO	50	91	136	223	273	298	1071
BACHARELADO EM ENGENHARIA DE PRODUÇÃO	289	347	463	712	688	626	3125
BACHARELADO EM ENGENHARIA ELÉTRICA	497	526	643	753	786	596	3801
EVADIDO	**2269**	**3140**	**3260**	**3424**	**1962**	**260**	**14315**
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO	211	445	526	678	350	49	2259
BACHARELADO EM ENGENHARIA DE PRODUÇÃO	920	1247	1367	1363	780	119	5796
BACHARELADO EM ENGENHARIA ELÉTRICA	1138	1448	1367	1383	832	92	6260
Total Geral	**3105**	**4104**	**4502**	**5112**	**3709**	**1780**	**22312**

Fonte: do Autor

Mas ao observar-se a Tabela 5, a mesma informação em percentual percebe-se que o curso de Engenharia da Computação tem 67,8% dos seus alunos matriculados no decorrer de 2015 a 2020 evadidos. Estes percentuais de evasão estão acima dos valores médios indicados pelo Censo da Educação Superior de 2018 do INEP, que é próximo de 60%. As outras Engenharias também apresentam valores superiores a 60% de evasão entre 2015 e 2018.

TABELA 5 - QUANTIDADE DE ALUNOS ATIVOS E EVADIDOS POR ANO E CURSO (2015 A 2020) EM %

Curso / Situação Aluno	2015	2016	2017	2018	2019	2020	Total Geral
ATIVO	**836**	**964**	**1242**	**1688**	**1747**	**1520**	**7997**
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO	19,2%	17,0%	20,5%	24,8%	43,8%	85,9%	32,2%
BACHARELADO EM ENGENHARIA DE PRODUÇÃO	23,9%	21,8%	25,3%	34,3%	46,9%	84,0%	35,0%
BACHARELADO EM ENGENHARIA ELÉTRICA	30,4%	26,6%	32,0%	35,3%	48,6%	86,6%	37,8%
EVADIDO	**2269**	**3140**	**3260**	**3424**	**1962**	**260**	**14315**
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO	80,8%	83,0%	79,5%	75,2%	56,2%	14,1%	67,8%
BACHARELADO EM ENGENHARIA DE PRODUÇÃO	76,1%	78,2%	74,7%	65,7%	53,1%	16,0%	65,0%
BACHARELADO EM ENGENHARIA ELÉTRICA	69,6%	73,4%	68,0%	64,7%	51,4%	13,4%	62,2%
Total Geral	**3105**	**4104**	**4502**	**5112**	**3709**	**1780**	**22312**

Fonte: do Autor

Outro ponto importante para análise dos resultados da evasão, em cada curso por ano é interpretar o modelo adotado na planilha, que foi adequado e tratado nesta pesquisa para o modelo do MEC da taxa de desistência, ou seja, você acompanha a taxa de evasão dos alunos matriculados em determinado ano durante os anos do curso. Assim ao observar-se pontualmente o curso de Engenharia da Computação no ano de 2015 do percentual de alunos evadidos, observa-se uma taxa de 80,8%, então, para cada 10 alunos que se matricularam em 2015 em Engenharia da Computação, 8 desistiram no decorrer dos anos do curso. Uma taxa muito superior do censo do MEC. Nos resultados desta tabela o ponto mais crítico da evasão está no ano de 2016 no curso de Engenharia da Computação, 83% de evasão.

Quando se analisam os resultados da evasão nos cursos pesquisados por Região Geográfica no Brasil, percebe-se na Tabela 6 a representatividade de alunos por região, ativos e evadidos. Existe uma concentração de alunos matriculados nas regiões sul e sudeste.

TABELA 6 - QUANTIDADE DE ALUNOS ATIVOS E EVADIDOS POR REGIÃO (2015 A 2020)

	Regiões do Brasil					
Nome do Curso/Situação Aluno	Sul	Sudeste	Norte	Centro Oeste	Nordeste	Total Geral
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO	**1196**	**1135**	**428**	**237**	**334**	**3330**
ATIVO	387	378	132	76	98	1071
EVADIDO	809	757	296	161	236	2259
BACHARELADO EM ENGENHARIA DE PRODUÇÃO	**3478**	**3071**	**1004**	**634**	**734**	**8921**
ATIVO	1159	1195	320	227	224	3125
EVADIDO	2319	1876	684	407	510	5796
BACHARELADO EM ENGENHARIA ELÉTRICA	**3145**	**3066**	**1648**	**954**	**1248**	**10061**
ATIVO	1186	1288	614	342	371	3801
EVADIDO	1959	1778	1034	612	877	6260
Total Geral	**7819**	**7272**	**3080**	**1825**	**2316**	**22312**

Fonte: do Autor

Ao analisar os resultados dos mesmos dados em percentual conforme a Tabela 7, as taxas de evasão mais representativas concentram-se na região nordeste para os 3 cursos pesquisados. A região que apresenta as menores taxas é a região sudeste.

TABELA 7 - QUANTIDADE DE ALUNOS ATIVOS E EVADIDOS POR REGIÃO (2015 A 2020) EM %

	Regiões do Brasil					
Nome do Curso/Situação Aluno	Sul	Sudeste	Norte	Centro Oeste	Nordeste	Total Geral
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO	**1196**	**1135**	**428**	**237**	**334**	**3330**
ATIVO	32,4%	33,3%	30,8%	32,1%	29,3%	32,2%
EVADIDO	67,6%	66,7%	69,2%	67,9%	70,7%	67,8%
BACHARELADO EM ENGENHARIA DE PRODUÇÃO	**3478**	**3071**	**1004**	**634**	**734**	**8921**
ATIVO	33,3%	38,9%	31,9%	35,8%	30,5%	35,0%
EVADIDO	66,7%	61,1%	68,1%	64,2%	69,5%	65,0%
BACHARELADO EM ENGENHARIA ELÉTRICA	**3145**	**3066**	**1648**	**954**	**1248**	**10061**
ATIVO	37,7%	42,0%	37,3%	35,8%	29,7%	37,8%
EVADIDO	62,3%	58,0%	62,7%	64,2%	70,3%	62,2%
Total Geral	**7819**	**7272**	**3080**	**1825**	**2316**	**22312**

Fonte: do Autor

Ao refinar as informações das regiões do Brasil por Estados observa-se um ponto de atenção da evasão. Na Tabela 8 observa-se que os Estados com maior taxa de evasão são: Roraima, Amapá, Alagoas e Rondônia.

TABELA 8 - ALUNOS ATIVOS E EVADIDOS POR ESTADO (2015 A 2020)
QUANT. E %

ESTADOS	ATIVO	EVADIDO	Total Geral	% EVADIDO
AC	28	68	96	70,8%
AL	16	50	66	75,8%
AM	49	123	172	71,5%
AP	33	106	139	76,3%
BA	392	694	1086	63,9%
CE	112	240	352	68,2%
DF	46	112	158	70,9%
ES	187	325	512	63,5%
GO	280	502	782	64,2%
MA	251	466	717	65,0%
MG	834	1376	2210	62,3%
MS	123	160	283	56,5%
MT	196	406	602	67,4%
PA	427	936	1363	68,7%
PB	23	60	83	72,3%
PE	121	231	352	65,6%
PI	72	135	207	65,2%
PR	1069	2194	3263	67,2%
RJ	359	725	1084	66,9%
RN	53	83	136	61,0%
RO	107	281	388	72,4%
RR	6	27	33	81,8%
RS	1068	1736	2804	61,9%
SC	595	1157	1752	66,0%
SE	26	55	81	67,9%
SP	1481	1985	3466	57,3%
TO	43	82	125	65,6%
Total Geral	**7997**	**14315**	**22312**	**64,2%**

Fonte: do Autor

Destaca-se nas análises de resultados a visão por idade do aluno evadido. A instituição pesquisada tem alunos matriculados entre 16 e 80 anos de idade nos cursos de engenharia. Na Tabela 9 observa-se que alunos com idade de 29 anos são os que mais evadem (taxa e quantidade) em cursos de engenharia na modalidade à distância na faixa de 16 a 52 anos. Na faixa de 30 a 40 anos concentram-se as menores taxas com destaque para alunos com 39 anos (menor evasão).

TABELA 9 - ALUNOS ATIVOS E EVADIDOS POR IDADE – DE 16 A 52 ANOS 2015 A 2020 – QUANT. E %

Idade Aluno	ATIVO	EVADIDO	TOTAL GERAL	EVASÃO POR IDADE %
16	1	3	4	75,0%
17	43	100	143	69,9%
18	156	337	493	68,4%
19	175	484	659	73,4%
20	216	513	729	70,4%
21	242	507	749	67,7%
22	235	510	745	68,5%
23	248	552	800	69,0%
24	276	546	822	66,4%
25	263	546	809	67,5%
26	325	566	891	63,5%
27	353	580	933	62,2%
28	348	655	1003	65,3%
29	319	725	1044	69,4%
30	365	623	988	63,1%
31	347	607	954	63,6%
32	380	590	970	60,8%
33	384	592	976	60,7%
34	370	554	924	60,0%
35	343	578	921	62,8%
36	341	488	829	58,9%
37	303	495	798	62,0%
38	312	457	769	59,4%
39	276	388	664	58,4%
40	229	347	576	60,2%
41	172	288	460	62,6%
42	155	257	412	62,4%
43	133	211	344	61,3%
44	99	177	276	64,1%
45	89	169	258	65,5%
46	91	137	228	60,1%
47	70	114	184	62,0%
48	47	100	147	68,0%
49	43	91	134	67,9%
50	45	74	119	62,2%
51	28	59	87	67,8%
52	28	51	79	64,6%

Fonte: do Autor

Acima de 52 anos a quantidade de alunos matriculados é menor e as taxas de evasão apresentam situações de variação extrema, conforme Tabela 10:

TABELA 10 - ALUNOS ATIVOS E EVADIDOS POR IDADE – DE 53 A 80 ANOS 2015 A 2020 – QUANT. E %

Idade Aluno	ATIVO	EVADIDO	TOTAL GERAL	EVASÃO POR IDADE %
53	17	38	55	69,1%
54	18	32	50	64,0%
55	14	25	39	64,1%
56	16	23	39	59,0%
57	16	17	33	51,5%
58	17	19	36	52,8%
59	10	14	24	58,3%
60	6	11	17	64,7%
61	5	15	20	75,0%
62	7	9	16	56,3%
63	1	5	6	83,3%
64	5	2	7	28,6%
65		8	8	100,0%
66	6	6	12	50,0%
67	2	3	5	60,0%
68	1	1	2	50,0%
69	2	4	6	66,7%
70		1	1	100,0%
71	1	1	2	50,0%
72	1		1	0,0%
73	1	2	3	66,7%
74		2	2	100,0%
75		2	2	100,0%
76	1	1	2	50,0%
77		1	1	100,0%
79		1	1	100,0%
80		1	1	100,0%
Total Geral	**7997**	**14315**	**22312**	**64,2%**

Fonte: do Autor

Uma análise inovadora desta pesquisa sobre evasão, a quantidade de alunos evadidos por mês da duração do curso, não observada em outras pesquisas. Os cursos de engenharias da instituição pesquisada possuem duração de 60 meses até sua conclusão pelo aluno (em regime normal de conclusão pelo aluno).

Na Tabela 11 observa-se que existe uma grande concentração da evasão nos cursos de engenharia nos primeiros 12 meses do curso, ou seja, 52,8% do total de alunos matriculados evadem, e nos 3 primeiros meses, a evasão é de 24,8%.

TABELA 11 - ALUNOS ATIVOS E EVADIDOS POR MÊS DE EVASÃO 2015 A 2020 – QUAT. E % - DOS MESES DE 0 A 30

Mês da Evasão	EVADIDO Quant.	Total Geral - Quant.	EVADIDO Acumulado Mês a Mês - Quant.	EVADIDO Mês Sobre Total de Alunos (%)	EVADIDO Acumulado Mês a Mês - Sobre Total Evadido (%)	EVADIDO Acumulado Mês a Mês - Sobre Total de Alunos (%)
0	797	885	797	3,6%	5,6%	3,6%
1	1456	2888	2253	6,5%	15,7%	10,1%
2	1293	1293	3546	5,8%	24,8%	15,9%
3	1542	1654	5088	6,9%	35,5%	22,8%
4	1332	1332	6420	6,0%	44,8%	28,8%
5	1236	1405	7656	5,5%	53,5%	34,3%
6	926	1115	8582	4,2%	60,0%	38,5%
7	960	994	9542	4,3%	66,7%	42,8%
8	677	894	10219	3,0%	71,4%	45,8%
9	389	508	10608	1,7%	74,1%	47,5%
10	437	437	11045	2,0%	77,2%	49,5%
11	315	533	11360	1,4%	79,4%	50,9%
12	416	416	11776	1,9%	82,3%	52,8%
13	230	919	12006	1,0%	83,9%	53,8%
14	206	206	12212	0,9%	85,3%	54,7%
15	243	305	12455	1,1%	87,0%	55,8%
16	166	233	12621	0,7%	88,2%	56,6%
17	147	148	12768	0,7%	89,2%	57,2%
18	131	271	12899	0,6%	90,1%	57,8%
19	163	237	13062	0,7%	91,2%	58,5%
20	89	395	13151	0,4%	91,9%	58,9%
21	101	237	13252	0,5%	92,6%	59,4%
22	90	90	13342	0,4%	93,2%	59,8%
23	74	269	13416	0,3%	93,7%	60,1%
24	105	105	13521	0,5%	94,5%	60,6%
25	57	765	13578	0,3%	94,9%	60,9%
26	72	72	13650	0,3%	95,4%	61,2%
27	65	112	13715	0,3%	95,8%	61,5%
28	53	54	13768	0,2%	96,2%	61,7%
29	59	166	13827	0,3%	96,6%	62,0%
30	41	200	13868	0,2%	96,9%	62,2%

Fonte: do Autor

No terceiro mês concentra-se a maior evasão em um único mês, taxa de 6,9%.

Do 31º mês em diante observa-se uma baixa evasão, indicando uma estabilidade conforme Tabela 12.

TABELA 12 - ALUNOS ATIVOS E EVADIDOS POR MÊS DE EVASÃO 2015 A 2020 – QUANT. E % - DOS MESES DE 31 A 60

Mês da Evasão	EVADIDO Quant.	Total Geral - Quant.	EVADIDO Acumulado Mês a Mês - Quant.	EVADIDO Mês Sobre Total de Alunos (%)	EVADIDO Acumulado Mês a Mês - Sobre Total Evadido (%)	EVADIDO Acumulado Mês a Mês - Sobre Total de Alunos (%)
31	57	57	13925	0,3%	97,3%	62,4%
32	27	181	13952	0,1%	97,5%	62,5%
33	38	38	13990	0,2%	97,7%	62,7%
34	40	152	14030	0,2%	98,0%	62,9%
35	26	304	14056	0,1%	98,2%	63,0%
36	42	42	14098	0,2%	98,5%	63,2%
37	14	397	14112	0,1%	98,6%	63,2%
38	31	31	14143	0,1%	98,8%	63,4%
39	21	61	14164	0,1%	98,9%	63,5%
40	13	17	14177	0,1%	99,0%	63,5%
41	18	81	14195	0,1%	99,2%	63,6%
42	8	154	14203	0,0%	99,2%	63,7%
43	21	21	14224	0,1%	99,4%	63,8%
44	19	108	14243	0,1%	99,5%	63,8%
45	19	20	14262	0,1%	99,6%	63,9%
46	3	67	14265	0,0%	99,7%	63,9%
47	10	195	14275	0,0%	99,7%	64,0%
48	11	12	14286	0,0%	99,8%	64,0%
49	4	375	14290	0,0%	99,8%	64,0%
50	6	6	14296	0,0%	99,9%	64,1%
51	9	58	14305	0,0%	99,9%	64,1%
52	4	4	14309	0,0%	100,0%	64,1%
53	6	84	14315	0,0%	100,0%	64,2%
55		149	14315	0,0%	100,0%	64,2%
56		94	14315	0,0%	100,0%	64,2%
57		2	14315	0,0%	100,0%	64,2%
58		137	14315	0,0%	100,0%	64,2%
60		51	14315	0,0%	100,0%	64,2%
61		276	14315	0,0%	100,0%	64,2%
Total Geral	**14315**	**22312**				

Fonte: do Autor

Ao analisar os resultados apresentados, a contribuição desta visão direciona os esforços do processo decisório e as estratégias da organização. Esta pesquisa não analisou dados acadêmicos de desempenho dos alunos, que podem contribuir ou não para a evasão, mas para estudos futuros podem ser um importante estudo para vincular as informações desta pesquisa e assim resultados mais específicos do processo de causa e efeito da evasão.

Outro ponto interessante dos resultados, está relacionado ao momento que o aluno entra na instituição pelo processo seletivo, o mês do seu ingresso. Na Tabela 13 observa-se a quantidade de alunos matriculados por mês de entrada e suas taxas de evasão;

TABELA 13 - ALUNOS ATIVOS E EVADIDOS POR MÊS DE ENTRADA 2015 A 2020 – QUANT. E %

Mês de Entrada	ATIVO	EVADIDO	% Evadidos	Total Geral
2	3921	5167	56,9%	9088
3	185	593	76,2%	778
4	776	1464	65,4%	2240
5	578	1529	72,6%	2107
6	90	313	77,7%	403
7	841	1632	66,0%	2473
8	317	901	74,0%	1218
9	489	888	64,5%	1377
10	489	1136	69,9%	1625
12	311	692	69,0%	1003
Total Geral	**7997**	**14315**	**64,2%**	**22312**

Fonte: do Autor

O primeiro processo seletivo do ano, que tem a entrada no segundo mês de cada ano, é a entrada com menor evasão do ano, assim como é a entrada bem mais representativa na quantidade de alunos matriculados. No processo seletivo do mês de junho de cada ano a Engenharia da Produção atinge 83,3% de taxa de evasão em análise por curso. Este resultado contribui no processo decisório dos recursos direcionados para as campanhas de captação durante o ano.

5.2 ANÁLISE DOS RESULTADOS – SIMULADOR PREDITIVO

Considerando todos os dados dos 22.312 alunos inseridos no simulador de predição, com o classificador algorítmico *Neive Bayes,* na fase de treinamento do sistema, indicou o desempenho do algoritmo em 87,1% (indicador de acurácia), ou seja, classificou corretamente 19.432 alunos. Indicador importante para os envolvidos no processo decisório. Para chegar neste percentual os seguintes cálculos e passos foram realizados:

Primeiro encontrou-se o resultado através do sistema da Matriz de Confusão (*confussion matrix*), por esta matriz é possível visualizar as classificações corretas e incorretas, que nesta pesquisa apresentaram os seguintes resultados na Figura 16 extraída do simulador de predição:

FIGURA 16 - MATRIZ DE CONFUSÃO – EXTRAÍDA DO SIMULADOR DE EVASÃO

```
=== Confusion Matrix ===

    a     b   <-- classified as
 6572  1425 |     a = ATIVO
 1455 12860 |     b = EVADIDO
```

Fonte: do Autor

A matriz foi readequada para compreensão no Quadro 9 para representar os *status* positivo e negativo dos resultados:

QUADRO 9 - MATRIZ DE CONFUSÃO

	Ativos (a)	Evadidos (b)
Ativos (a)	6572 **Verdadeiro Positivo (VP)**	1425 **Falso Negativo (FN)**
Evadidos (b)	1455 **Falso Positivo (FP)**	12860 **Verdadeiro Negativo (VN)**

Fonte: do Autor

A interpretação dos resultados do estudo de caso e dos dados encontrados deve ser realizado da seguinte forma:

a) Verdadeiros Positivos - VP: quantidade de alunos ATIVOS classificados corretamente como ATIVOS pelo algoritmo de classificação, na figura a quantidade de alunos classificados foi de 6.572;
b) Falsos Positivos - FP: quantidade de alunos ATIVOS classificados incorretamente como EVADIDOS, na figura representado por 1.455 alunos;
c) Falsos Negativos - FN: quantidade de alunos EVADIDOS classificados incorretamente como ATIVOS, nos resultados apresentados foram 1.425 alunos;

d) Verdadeiro Negativos - VN: quantidade de alunos EVADIDOS classificados corretamente como EVADIDOS, nos resultados apresentados foram 12.860 alunos.

Desta forma é possível compreender o detalhamento dos 2.880 alunos classificados de forma equivocada dos 22.312 alunos totais no simulador de predição.

O ***Kappa*** é o indicador de desempenho proposto por Jacob Cohen em 1960, um método estatístico, mede a concordância de dois conjuntos de dados, nesta pesquisa os ATIVOS e EVADIDOS, calculado pela seguinte fórmula:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

Onde:

p0 é a taxa de aceitação relativa (taxa de acerto ou erro)

pe é a taxa hipotética de aceitação

Obs: quando a concordância é total entre os dois conjuntos de dados o k =1,0

Para calcularmos **p0** basta dividirmos o número de concordâncias (VP e VN) com a quantidade total de alunos:

Neste estudo de caso **p0**, a taxa de aceitação relativa é representada da seguinte forma:

p0 = VP + VN = 6572 + 12.860 = 19.432 dividido por 22.312 (total de alunos)

= 0,870921 (87,1%)

ou seja, o resultado encontrado até esta etapa é o desempenho da classificação do algoritmo de classificação *Naive Bayes* desenvolvido no simulador com recurso do pacote de softwares do *WEKA.*

Para continuação do cálculo do ***Kappa*** realizou-se o cálculo do **Pe**, que é a probabilidade de concordância randômica:

a) **Pe=VP + FN / FP + VN** = (6.572 + 1.425) / (1.455 + 12.860) = 7997/14.315 **= 0,558644**

Calculando o valor final do ***Kappa***:

$$\mathbf{K} = \frac{0.870921 - 0.558644}{1 - 0.558644} = \mathbf{0.707539}$$

Landis e Koch (1977) desenvolveram um classificador do desempenho do indicador ***Kappa***, na Figura 17 as faixas de classificação do desempenho:

FIGURA 17 - TABELA DE DESEMPENHO DO INDICADOR KAPPA

Valor do coeficiente Kappa	Interpretação
<0	Sem concordância
0-0.19	Concordância pobre
0.20-0.39	Concordância fraca
0.40-0.59	Concordância moderada
0.60-0.79	Concordância substanciual
0.80-1.00	Concordância quase perfeita

Fonte: adaptado Landis e Koch (1977)

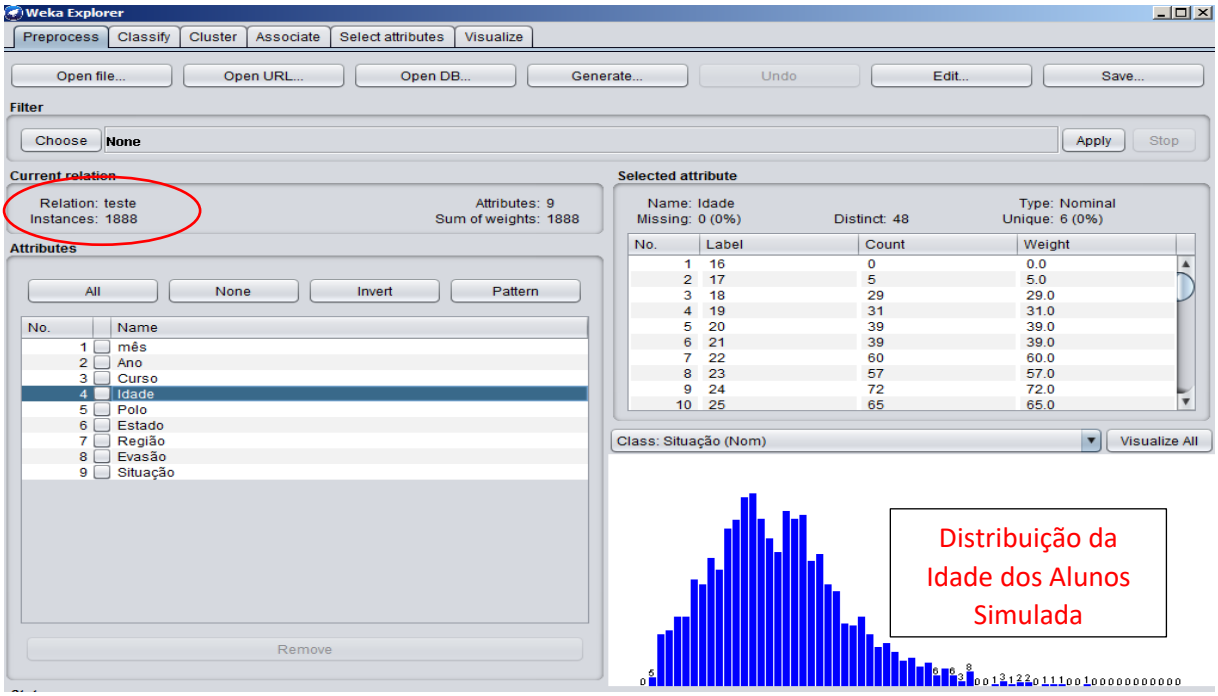
Com o resultado do *Kappa* no estudo de caso em **0,71** a interpretação é de que é um simulador de predição de concordância substancial para uso. Desta forma, com a utilização das 8 variáveis de entrada definidas para o estudo de caso é possível a utilização do simulador para as fases seguintes de análise.

5.3 ANÁLISE DOS RESULTADOS – BASE TESTE (SIMULAÇÃO)

Nesta fase da pesquisa, elaborou-se um arquivo para poder simular a eficiência e execução de predição do simulador desenvolvido, este arquivo foi elaborado com 1.888 alunos novos hipotéticos (número médio dos ingressantes em cada ano).

O arquivo recebeu o nome de Teste. Para a simulação todos os alunos foram considerados como alunos ATIVOS, e, assim, pode-se verificar quantos e quais alunos o sistema considera como grupo de risco de evasão. Na Figura 18 observa-se a imagem do simulador de predição com os 1.888 novos alunos e no gráfico a distribuição da população por idade (exemplo de possibilidades gráficas):

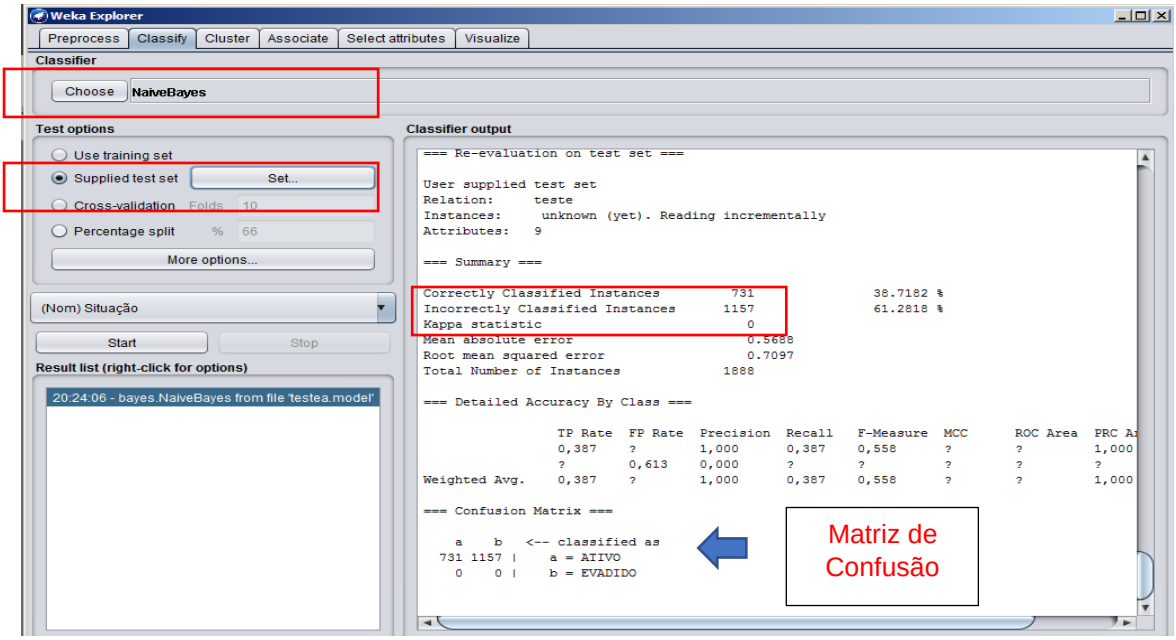
FIGURA 18 - TELA DA BASE DE TESTES DO SIMULADOR DE PREDIÇÃO



Fonte: do Autor

Para a realização do simulador de predição com novos dados (instâncias), utiliza-se como base do modelo de predição o arquivo de aprendizado utilizado durante os anos de 2015 a 2020 com os 22.312 alunos (base de treinamento do algoritmo), ou seja, a base histórica. Na Figura 19 a ilustração do simulador com a opção de realização da base de teste/validação:

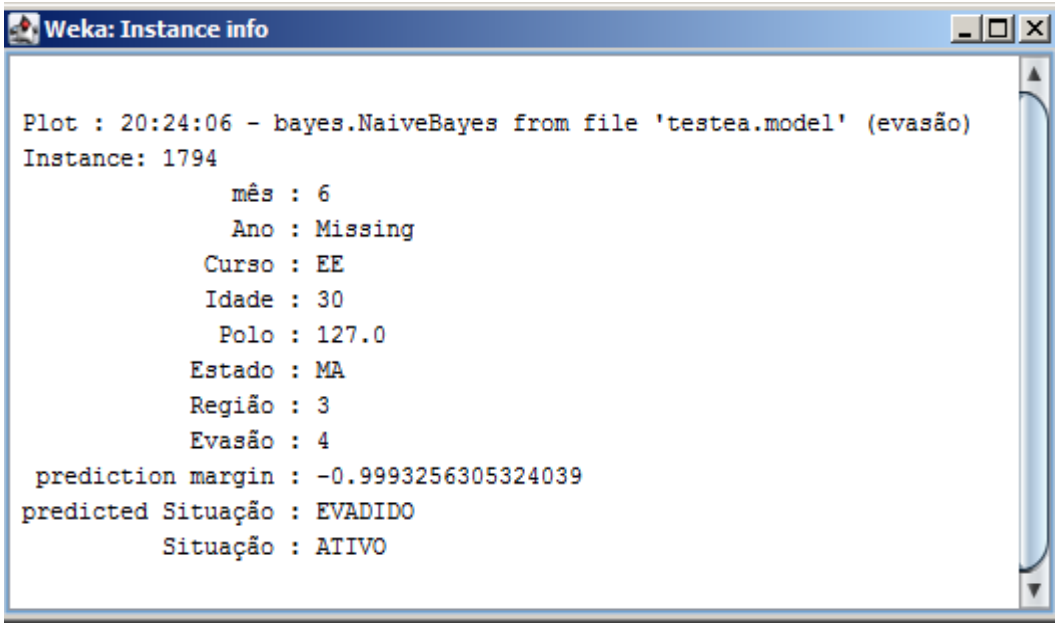
FIGURA 19 - TELA DA BASE DE TESTES DO SIMULADOR DE PREDIÇÃO



Fonte: do Autor

Os resultados encontrados demonstram que nesta nova base simulada, de 1.888 novos alunos, 1.157 estão no grupo de risco de evasão, ao qual a instituição precisa monitorar a partir deste momento. A taxa futura de evasão caso a instituição não realize alguma intervenção neste grupo será de 61,28%. O simulador indica que 731 alunos se manterão ativos. Na Figura 20 observa-se o exemplo de um aluno que o simulador indica como de risco de evadir:

FIGURA 20 - TELA DA BASE DE TESTES DO SIMULADOR DE PREDIÇÃO

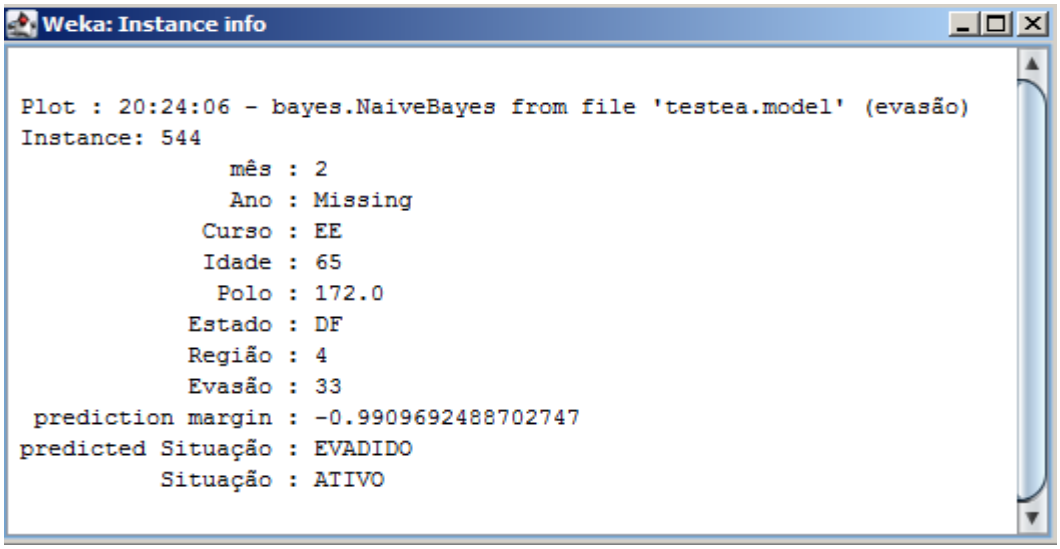


Fonte: do Autor

A linha de nome *Instance*, identifica um aluno codificado com o número 1794, observa-se que o parâmetro de entrada indica que este aluno está ATIVO (linha: Situação), mas o simulador de predição *(predicted situação)* pelo modelo de classificação escolhido (algoritmo), classifica o aluno no grupo de risco, ou seja, ele provavelmente irá evadir-se. Neste exemplo o aluno entrou na instituição em junho, escolheu o curso de Engenharia Elétrica (EE), tem 30 anos de idade (uma das idades com maior taxa de evasão na base de treinamento), é do Estado do Maranhão (que apresenta também altas taxas de evasão pela análise por estados).

Outro exemplo na Figura 21:

FIGURA 21 - TELA DA BASE DE TESTES DO SIMULADOR DE PREDIÇÃO

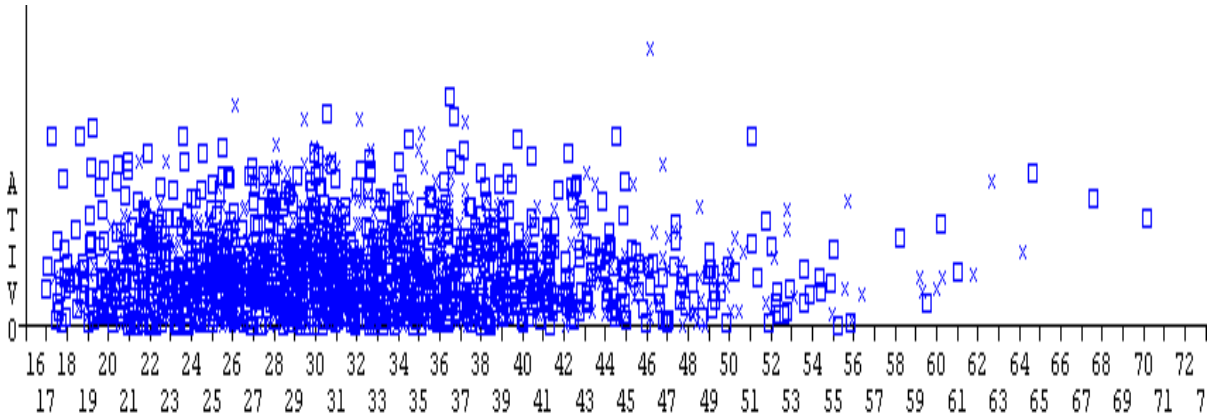


Fonte: do Autor

Neste exemplo também um aluno que foi classificado no grupo de risco, ingressou na instituição em fevereiro (mês com menores taxas de evasão), escolheu o curso de Engenharia Elétrica (com menor taxa de evasão), mas matriculou-se no Distrito Federal e tem 65 anos de idade, atributos com altas taxas de evasão.

O simulador possui diversos recursos gráficos que possibilitam ver a “massa” de dados dos alunos e os grupos de risco, na Figura 22 um exemplo dos alunos ativos reais e os classificados pelo sistema como grupo de risco pelo atributo IDADE (quando relacionado com as outras 7 variáveis de entrada):

FIGURA 22 - ALUNOS ATIVOS E GRUPO DE RISCO - BASE TESTE POR IDADE

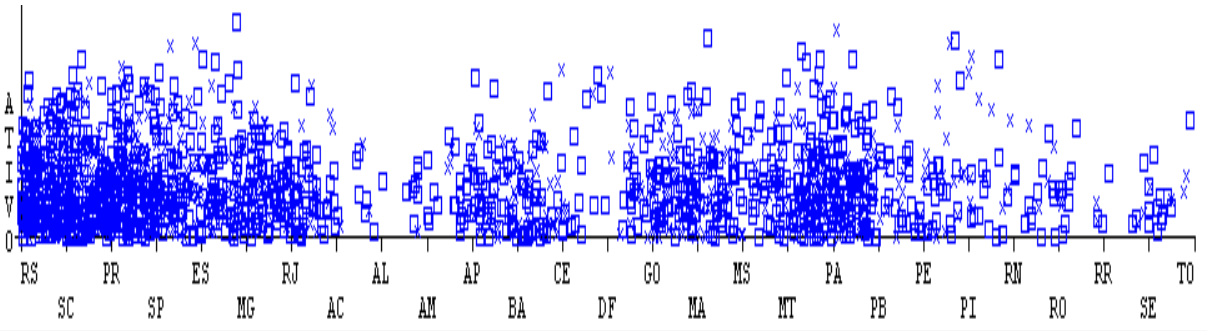


Fonte: do Autor

Os pequenos **"X"** indicados na figura são os alunos classificados corretamente como alunos ativos e os pequenos **"quadrados"** indicam os alunos que estão no grupo de risco da evasão por IDADE. Observa-se que para as idades maiores a probabilidade de evasão é superior, como indicado na base de treinamento do sistema de predição.

Na Figura 23 a análise por Estado:

FIGURA 23 - ALUNOS ATIVOS E GRUPO DE RISCO (BASE DE TESTE) – POR ESTADO

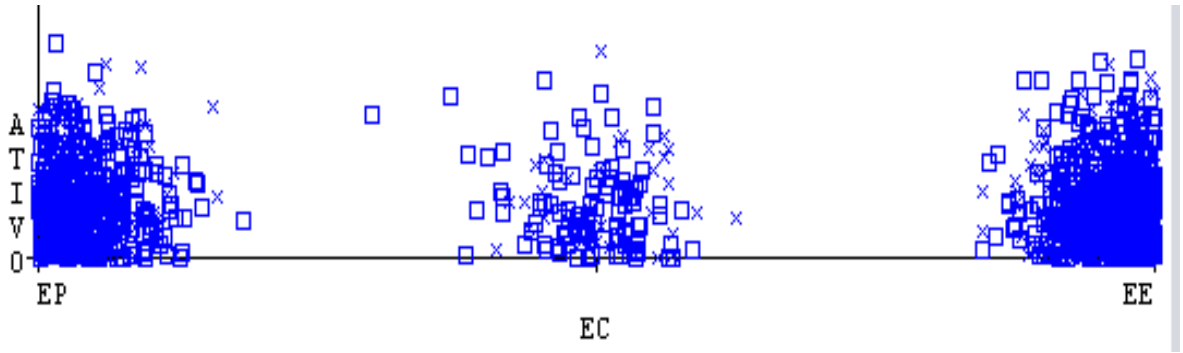


Fonte: do Autor

Ao analisar-se os resultados por Estado do Brasil no gráfico, observar-se que na base de teste/validação, a representação de maior taxa de evasão concentra-se nos Estados da região Norte e Nordeste, como indicado na base de treinamento. Ponto de atenção para a instituição participante do estudo de caso.

Na Figura 24 observa-se um exemplo da base de teste por curso escolhido:

FIGURA 24 - ALUNOS ATIVOS E GRUPO DE RISCO (BASE DE TESTE) – POR CURSO



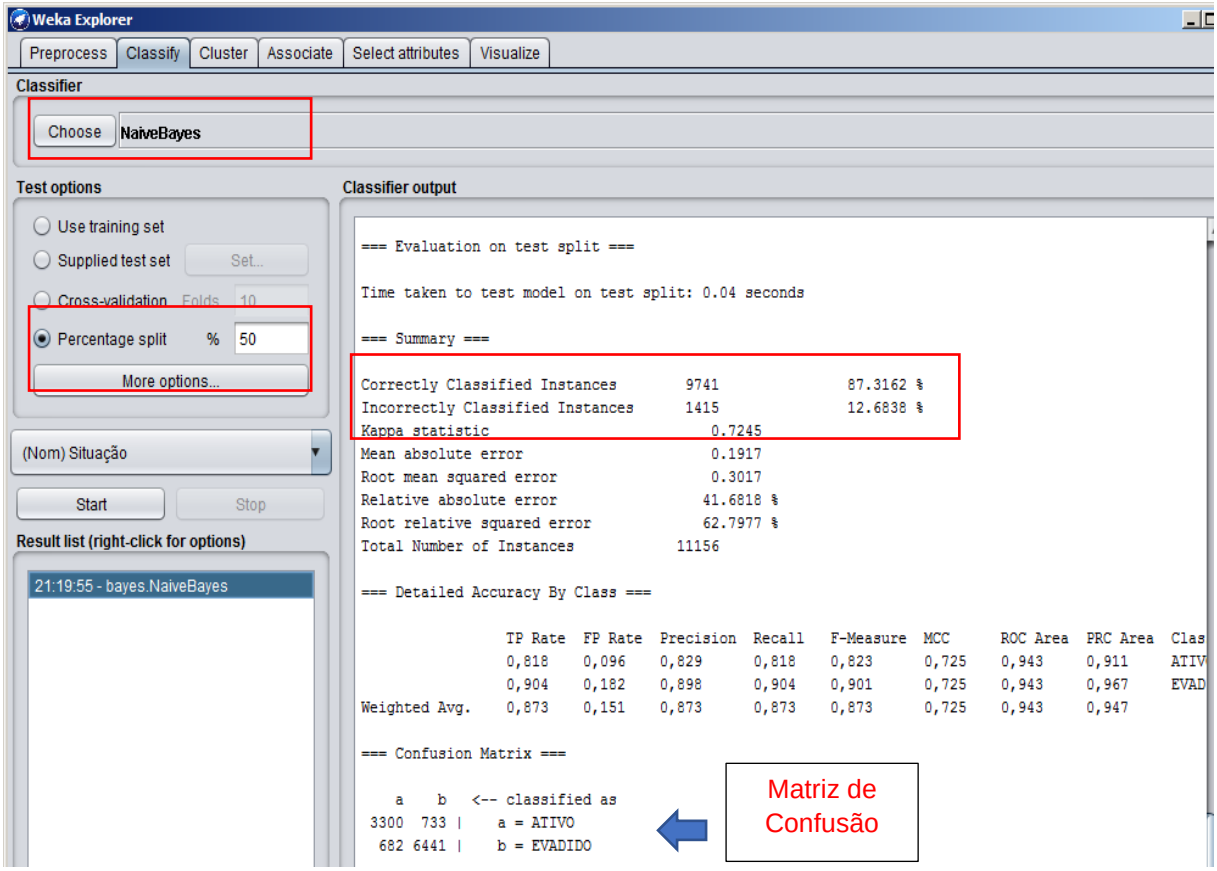
Fonte: do Autor

Ao analisar-se os resultados, observa-se o grupo de risco mais intenso no curso de Engenharia da Computação (EC) em relação aos cursos de Engenharia Elétrica (EE) e Produção (EP).

5.4 ANÁLISE DOS RESULTADOS – BASE VALIDAÇÃO

Outro recurso que o sistema possui, é verificar o tamanho da base de dados que poderíamos utilizar como base de “treinamento” do simulador de predição, e a base de “teste”, dentro de uma acurácia e um indicador *Kappa* “aceitável”, desta forma foi realizada várias simulações na base dos 22.312 alunos, na Figura 25 observa-se esta simulação no simulador de predição:

FIGURA 25 - TELA SIMULADOR DE PREDIÇÃO - 50% TREINAMENTO E 50% TESTE (EXEMPLO)



Fonte: do Autor

Na Tabela 14 uma simulação utilizando-se a base dos 22.312 alunos como sendo a própria base de treinamento e teste e seu nível de *Kappa:*

TABELA 14 - TABELA DE RESULTADOS - % DE TREINAMENTO E TESTE - INDICADORES

% Treinamento	% Teste	% Acurácia	% Kappa
10	90	85,8	0,69
20	80	86,7	0,71
30	70	86,6	0,71
40	60	87,2	0,72
50	50	87,3	0,72
60	40	87,1	0,72
70	30	87,1	0,72
80	20	87,3	0,73
90	10	86,9	0,72

Fonte: do Autor

O simulador permite programar na base de dados o percentual a ser utilizado, como base de treinamento e um percentual da base para teste. Na tabela 14 observa-se que para uma base de treinamento de 10% e base de teste de 90% a acuraria da

predição é de 85,8%, para 50% de base de treinamento e 50% base de teste a acurácia apresenta 87,3%. Com o algoritmo escolhido e os atributos de entrada selecionados percebe-se que mesmo com uma base de treinamento menor e possível ter uma acurácia significativa.

6 CONSIDERAÇÕES FINAIS

Muitas são as conjecturas levantadas quando o assunto é evasão na educação superior. Supõem-se descontentamento com o curso escolhido, falta de recursos financeiros para saldar as mensalidades, dificuldades em conciliar horas de estudo e de trabalho, entre outras hipóteses. Seja qual for o problema, há que se refletir sobre como solucioná-lo, ou ao menos, apontar caminhos que apontem possibilidades de resolução.

Essa dissertação nasce de um desconforto do pesquisador frente a esta situação, pois atua em cargos de gestão a muitos anos e reconhece ser este um tema que aflige tanto a instituição, que perde o aluno, quanto o próprio aluno, que perde chances de crescimento, desenvolvimento ou aprimoramento pessoal e profissional. Por isto a proposição do problema de pesquisa: Como conceber um simulador computacional preditivo, produto desta dissertação, utilizando mineração de dados, para predizer os alunos em risco de evasão escolar dos cursos de engenharia na modalidade de ensino à distância de uma Instituição de Educação Superior (IES)?

Um simulador, pelos olhos do pesquisador, possibilita predizer acontecimentos futuros, contribuindo na reflexão sobre o conjunto de variáveis que envolvem a evasão discente dos graduandos, no caso deste trabalho, das engenharias, pois, concede aos gestores, dados que permitem uma melhor tomada de decisão no combate a evasão.

Concentrando-se neste problema e em busca de respostas, delineou-se como objetivo geral da dissertação, propor um modelo computacional utilizando mineração de dados, para predizer os alunos em risco de evasão escolar dos Cursos de Engenharias na modalidade de educação à distância e contribuir para o processo decisório das IES na mitigação deste fenômeno.

Para dar conta desta tarefa, inicialmente buscou-se atender ao primeiro objetivo específico explicitado na introdução deste trabalho: Realizar um levantamento bibliográfico sobre a evasão discente e suas possíveis causas, assim como caracterizar o problema da evasão na educação superior na modalidade à distância.

Tal levantamento foi realizado no capítulo 2 desta dissertação, que abordou especialmente os fatores que promovem e reduzem a evasão, assim como as principais tendências com relação ao tema. Neste capítulo, Tontini e Walter (2014),

Sepúlvida (2018) e Cunha e Morosini (2014), contribuíram com conceitos e aspectos inerentes a discussão, apontando que muitas são as razões pelas quais a evasão merece destaque em pesquisas, na busca por mostrar caminhos que favoreçam a sua redução. Laguardia e Portela (2009) e Gottardo et al. (2012) contribuem demonstrando possibilidades de estratégias para o combate a evasão

Ainda quanto a este primeiro objetivo explorou-se o conceito de mineração de dados como a aplicação de técnica de investigação de grandes quantidades de dados que permitem a descoberta de novos padrões e relações que não seriam encontrados facilmente sem a ajuda de uma estratégia específica. Também se abordou os modelos de predição e algoritmos de classificação que conduziram o pesquisador ao WEKA, um sistema que possui um conjunto de algoritmos de aprendizado de máquina para a execução de tarefas de mineração de dados, sendo este software o eleito pelo pesquisador para o delineamento do produto desta dissertação

O segundo objetivo, analisar as informações do banco de dados recebidas da instituição de ensino pesquisada dos alunos dos cursos de graduação em Engenharia Elétrica, da Computação, da Produção na modalidade à distância, identificando as variáveis fornecidas, foi contemplado no tópico – Coleta de dados – em Metodologia.

A coleta dos dados, que ocorreu depois da análise documental das informações encontradas nas bases de dados dos acadêmicos dos cursos selecionados, possibilitou distinguir atributos que pudessem ser usados em um modelo de predição de evasão, simulando no momento do ingresso do aluno na instituição se o mesmo possui ou não risco de evadir.

A concepção do simulador, item que corresponde ao terceiro objetivo específico, qual seja: projetar o simulador de predição em uma arquitetura computacional capaz de identificar os alunos em risco de evasão utilizando como base o pacote de software de apoio do *WEKA*[3] *(Waikato Environment for Knowledge Analysis)*, foi sistematizada no capítulo 4 que apresentou uma solução de um simulador computacional de predição de evasão com os recursos do pacote de software do *WEKA*, desde sua concepção, desenvolvimento e avaliação, para desta

[3] *O WEKA,* conforme *Waikato* (2017) é um sistema que possui um conjunto de algoritmos de aprendizado de máquina para a execução de tarefas de mineração de dados.

forma contribuir no processo decisório das organizações educacionais e agregar valor na compreensão do fenômeno da evasão.

O último objetivo específico, realizar simulações e experimentos para medir a acurácia dos resultados encontrados e identificar as variáveis no estudo de caso que impactam na evasão, foi contemplado no capítulo 5, com o tratamento das informações recebidas pela instituição e organizadas nos atributos de análise, foi possível identificar alguns ensaios e cenários de resultados.

Compreender o fenômeno da evasão na educação superior e mitiga-lo, contribui na conclusão e realização de sonhos que se desfazem pelo caminho da evasão.

A pesquisa, através da Mineração de Dados Educacionais, conseguiu usar técnicas e modelos de predição com uso de algoritmo de classificação *Naive Bayes*, apresentar o conhecimento necessário para a melhor eficiência e eficácia das organizações em seus processos decisórios operacionais e estratégicos, e, minimizar a complexidade de gestão característica da área de educação.

Uma proposta de solução tecnológica na área da educação, precisa compreender a complexidade do segmento e assim garantir certo grau de precisão em seus resultados, para a assertividade dos caminhos a serem percorridos por todos os envolvidos neste segmento. Esta pesquisa se propôs e demonstrou uma solução, através de um simulador de predição de evasão com grau aceitável de acurácia na predição da evasão em um conjunto de atributos analisados. O uso adequado de soluções tecnológicas na área da educação, de fato pode contribuir com o segmento e com os seus propósitos.

A abordagem desta pesquisa na educação superior, na modalidade à distância, em cursos de Engenharia, demonstrou o desafio de compreender os altos níveis de evasão em cursos que contribuem de forma significativa para o desenvolvimento de uma sociedade e país.

O foco desta pesquisa, em 3 cursos de engenharia em um estudo de caso, no período analisado de 2015 a 2020, com 22.312 alunos, utilizando-se de 8 atributos sócio demográficos de entrada para análise da evasão, com o auxílio de um simulador de predição, utilizando-se das modelagens de mineração de dados com auxílio do

pacote de software do *WEKA,* com o algoritmo *Naive Bayes*, apresentou uma acurácia em seus resultados de 87,1% no simulador desenvolvido.

Os resultados apresentados na etapa do pré-processamento no simulador de predição, analisando-se as variáveis de entrada, demonstrou situações pontuais que merecem atenção, análise e ação para a mitigação da evasão na instituição pesquisada. Em destaque a idade dos alunos matriculados nos cursos de engenharia, pois algumas idades estão mais propensas a evasão, como a idade de 29 anos, assim como alunos oriundos de alguns Estados do Brasil que se classificam no grupo de risco, como Roraima, Amapá, Alagoas e Rondônia. Outro ponto relevante e inovador desta pesquisa foi a análise do mês de evasão do aluno, indicando uma concentração nos primeiros 12 meses do curso. O mês de entrada escolhido pelo aluno para iniciar seus estudos na engenharia também indicou a propensão ou não a evasão, sendo o mês de entrada em junho o com a maior taxa de evasão, 77,3%.

A comparação com outros estudos, permitiu observar que esta pesquisa está alinhada em alguns aspectos e resultados, como: uso da mineração de dados e a tecnologia aplicada a educação, modelos de predição, acurácia nos resultados, assertividade na predição e atributos preditores.

Nestas comparações, estudos mais específicos no ensino superior na modalidade à distância são recentes, observado pelas datas dos estudos levantados na pesquisa bibliográfica desta pesquisa. É uma modalidade de ensino com enormes desafios e taxas de evasão crescentes, que precisa de novas pesquisas, com o cruzamento simultâneo de muitos atributos e uso da tecnologia, para cada vez mais identificar através de modelos de predição o fenômeno da evasão e contribuir na realização de sonhos.

Estudos futuros podem ser desenvolvidos, com o uso do simulador de predição proposto, incluindo também variáveis de desempenho acadêmico de alunos que sejam classificados como de risco ou não com as variáveis de ingresso, ou seja, acompanhar a vida acadêmica do aluno no decorrer do curso e observar novas análises que permitam complementar o processo decisório.

REFERÊNCIAS

ABED, cendo.ead.br. **Relatório analítico de aprendizagem a distância no Brasil**. São Paulo. Pearson Education do Brasil. 2013.

BAKER, R., et al. ***Data mining for education. International encyclopedia of education*** 7, 3 (2010), 112–118.

BAKER, Ryan SJD; ISOTANI, Seiji; CARVALHO, Adriana. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação, v. 19, n. 02, p. 03, 2011.

BAKER, R. S. ***Educational data mining: An advance for intelligent systems in education****. IEEE Intelligent systems* 29, 3 (2014), 78–82.

BAKER, R. S.; Yacef, K. **The state of educational data mining in 2009**: A review and future visions. JEDM| Journal of Educational Data Mining 1, 1 (2009), 3–17.

BARROSO, Marta F & FALCÃO, Eliane BM. **Evasão universitária: o caso do Instituto de Física da UFRJ.** IX Encontro Nacional de Pesquisa em Ensino de Física, v. 9, p. 1-14, 2004.

BOGDAN, R. C.; BIKLEN, S. K. **Investigação qualitativa em educação: uma introdução à teoria e aos métodos.** Portugal: Porto Editora, 1994.

BOUCKAERT, R., EIBE, F., HALL, M., *et al.*, 2010, ***WEKA*** **Manual for Version 3-6-4**. 2010. Disponível em: http://ufpr.dl.sourceforge.net/project/weka/documentation/3.6.x/WekaManual-3-6-4.pdf . Acesso em: 15 ago. 2020.

BRASIL. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**. Relatório da Comissão Especial de Estudos sobre Evasão nas Universidades Públicas Brasileiras. Brasília, DF: ANDIFES/ABRUEM/SESu/MEC, 1997.

BRASIL (2014). Plano Nacional de Educação 2014-2024 (Lei 13.005/2014). Acessado em: http://www.planalto.gov. br/ccivil_03/_Ato2011-2014/2014/Lei/L13005.htm.

BUENO, José Lino Oliveira. **A evasão de alunos**. Paidéia, Ribeirão Preto, n. 5, p. 9-16, ago. 1993.

CARVALHO, L. A. V., ***Datamining – A mineração de Dados no Marketing****, Medicina, Economia, Engenharia e Administração*. Rio de Janeiro: Editora Ciência Moderna Ltda.2005.

CUNHA, E. R.; MOROSINI, M. C. **Evasão na educação superior: uma temática em discussão.** Revista Cocar, v. 7, n. 14, p. 82-89, 2014.

DICIO, **Dicionário online de português**. Disponível em: https://www.dicio.com.br/evasao/.

FARRER, H. et al. **Algoritmos Estruturados**. 2ª ed. Rio de Janeiro : Guanabara Koogan, 1989.

FAYYAD, Usama et al.**The KDD Process for Extracting Useful Knowledge from Volumes of Data, In: Communications of the ACM**. New York, v. 39, n. 11, p.27-39.1996.

FONSECA, J. J. S. **Metodologia da pesquisa científica**. Fortaleza: UEC, 2002.

FURLAN, Matheus Batista **Algoritmos e técnicas para mineração de dados**/ Matheus Batista Furlan. – Assis, 2018.

GOLDSCHMIDT, R.; Passos, E. **Data Mining**. Elsevier Brasil, 2015

GOTTARDO, E. et al. Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais. In: **Anais do Simpósio Brasileiro de Informática na Educação.** 2012.

GOTTARDO, E. et al**.** Avaliação de Desempenho de Estudantes em Cursos de Educação a Distância Utilizando Mineração de Dados. In: **Anais do Workshop de Desafios da Computação Aplicada à Educação**. 2012. p. 30-39.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, *et al.*, 2009, **The WEKA Data Mining Software: An Update**. SIGKDD Explorations, Volume 11, Issue 1. 10-18, 2009.

HAN, J., KAMBER, M., 2006, **Data Mining Concepts and Techniques**. Morgan Kauffmann Publishers, Second Edition, 2006.

HUME, David. **Tratado da natureza humana**. 2.ed. São Paulo. Unesp. 2009.

KEMBER, D. et al. ***Student progress in distance education: identification of explanatory constructs***. British Journal of Educational Psychology, v. 62, p. 285-298, 1992.

LAGUARDIA, J.; PORTELA, M. **Evasão na educação a distância Dropout in distance education.** ETD-Educação Temática Digital, v. 11, n. 1, p. 349-379, 2009.

Landis, J.R. and Koch, G.G. (1977) ***The Measurement of Observer Agreement for Categorical Data. Biometrics***, 33, 159-174.

LOBO, M. B. D. C. M. **Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções**. *Revista da ABMES*, Brasília, v. 1, n. 25, 2012.Disponível em :http://www.institutolobo.org.br/imagens/pdf/artigos/art_087.pdf. Acesso em: 03 ago. 2020.

LIMA, F., ZAGO, N. **Evasão no ensino superior: tendências e resultados de pesquisa**. 2016. 15f. Artigo. ANPED SUL XI. Curitiba. Paraná

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G. **Predição do Desempenho Acadêmico de Graduandos Utilizando Mineração de Dados Educacionais**. Tese de doutorado, 2015b.

MARTINHO, V. R. C. **Sistema inteligente para predição do grupo de risco de evasão discente**. 2014. 145 f. Tese (Doutorado em Engenharia Elétrica) - Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Engenharia de Ilha Solteira, 2014.

MECHIE, D.; SPIEGELHALTER, D.; TAYLOR, C. **Machine Learning, Neural and Statistical Classifications.** Ellis Horwood, 1994.

MINTZBERG, Henry. **Criando organizações eficazes**. 2. ed. São Paulo: Atlas, 2003.

MORAN, José Manuel, MASETTO, Marcos, BEHRENS, Marilda. **Novas tecnologias e mediação pedagógica .**6.ed.SãoPaulo:Papirus,2003.

NEY, Otávio Abrantes de Sá. **Sistemas de informação acadêmica para o controle da evasão**. 2010. 145 f. Dissertação (Mestrado em Engenharia de Produção) - Programa de PósGraduação em Engenharia de Produção, Universidade Federal da Paraíba, João Pessoa, 2010.

PAIVA, R., BITTENCOURT, I.I., SILVA, A.P., ISOTANI, S., JAQUES, P., 2014. ***A Systematic Approach for Providing Personalized Pedagogical Recommendations Based on Educational Data Mining*.** In: International Conference on Intelligent Tutoring Systems, 2014, Honolulu. Lecture Notes in Computer Science, 2014. p. 362-367.

PARENTE, Nórlia N. **As condições de acesso e permanência dos estudantes do curso de licenciatura em Física do IFCE**, Campus De Sobral. 2014. 166f. Dissertação (Mestrado Profissional em Políticas Públicas e Gestão da Educação Superior) - Universidade Federal do Ceará. Fortaleza, 2014

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 3, n. 1, p. 12–27, 2013.

SALIBA, W. L. C. **Técnicas de Programação – Uma Abordagem Estruturada**. São Paulo: Makron, McGraw-Hill, 1992.

SANTOS, Priscila K. **Abandono na Educação Superior: um estudo do tipo Estado do Conhecimento.** Educação Por Escrito. Porto Alegre, v. 5, n. 2, p. 240-255, julho/dezembro, 2014.

SEPÚLVIDA, W.R. **Predição de Evasão na Educação a Distância como Subsídio a Tomada de Decisão.** Dissertação. Universidade Católica de Brasília. 2016.

SILVA, D. **Modelo para predição de risco de evasão na educação a distância utilizando técnicas de mineração de dados**. Dissertação. Universidade Federal Fluminense. Niteroi. 2019

SILVA FILHO, R.L.L; MOTEJUNAS P.R; HIPÓLITO O., LOBO M.**A Evasão no Ensino Superior Brasileiro**. Instituto Lobo de Desenvolvimento da Educação. São Paulo. Cadernos de Pesquisa, v. 37, n. 132, set./dez. 2007

SOBRAL,J.J.Veiga. Educação. 2018. Disponível em: https://revistaeducacao.com.br/2018/09/03/causaeseusefeitos/. Acesso em 30/07/2020.

TAN, P.-N.; Steinbach, M.; Kumar, V. **Introdução ao datamining: mineração de dados.** Ciência Moderna, 2009.

TINTO, V. ***Dropout from higher education: a theoretical synthesis of recent research****. Review of Educational Research,* Washington, v. 45, n. 1, p. 89-125, 1975.

TONTINI, Gérson; WALTER, Silvana Anita. **Pode-se identificar a propensão e reduzir a evasão de alunos?: ações estratégicas e resultados táticos para instituições de ensino superior**. Avaliação, Campinas; Sorocaba, v. 19, n. 1, p. 89-110, mar. 2014.

ZHANG, H. (2004). **The optimality of naive Bayes**. *AA*, v. 1, n. 2, p. 3.

Waikato, T. U. (2017). **Weka: Data mining software in Java, https://www.cs.waikato.ac.nz/ml/ weka/**. Acessado: jun.20.

WEKA (2017). ***Weka 3 - Data Mining with Open Source Machine Learning Software in Java****. htt*p://www.cs.waikato.ac.nz/ml/weka/, [acessado em jun.20].

YIN, R. K. **Estudo de Caso - 5.Ed.: Planejamento e Métodos**. Bookman Editora, 2015. São Paulo.

YANG, Y. (1994). *Expert Network:* ***Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval****. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* , SIGIR '94. Springer-Verlag New York, Inc. http://dl.acm.org/citation.cfm?id=188490.188496, [acessado em ago 20].