

Universidade FUMEC  
Faculdade de Ciências Empresariais  
Programa de Pós-Graduação em Sistemas de Informação e Gestão do  
Conhecimento

# **Machine learning e a evasão escolar - Análise preditiva no suporte à tomada de decisão**

Alex Marques de Souza

Belo Horizonte

2020

Alex Marques de Souza

## **Machine learning e a evasão escolar - Análise preditiva no suporte à tomada de decisão**

Projeto de dissertação de mestrado apresentado ao Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento como parte dos requisitos para a obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento.

**Área de concentração:** Gestão de Sistemas de Informação e do Conhecimento.

**Linha de pesquisa:** Tecnologia e Sistemas de Informação.

Orientador: Prof. Dr. Luiz Cláudio Gomes Maia

Belo Horizonte

2020

**Dados Internacionais de Catalogação na Publicação (CIP)**

S729m Souza, Alex Marques de, 1978-  
Machine learning e a evasão escolar: análise preditiva no  
suporte à tomada de decisão/ Alex Marques de Souza. - Belo  
Horizonte, 2020.  
138f. ; il. ; 29,7 cm

Orientador: Luiz Cláudio Gomes Maia  
Dissertação (Mestrado em Sistemas de Informação e  
Gestão do Conhecimento), Universidade FUMEC, Faculdade  
de Ciências Empresariais, Belo Horizonte, 2020.

1. Evasão escolar. 2. Aprendizado do computador. 3.  
Tecnologia da informação. I. Título. II. Maia, Luiz Cláudio  
Gomes. III. Universidade FUMEC, Faculdade de Ciências  
Empresariais.

CDU: 371.3

Dissertação intitulada “**Machine learning e a evasão escolar - Análise preditiva no suporte à tomada de decisão**” de autoria de Alex Marques de Souza, aprovada pela banca examinadora constituída pelos seguintes professores:



---

Prof. Dr. Luiz Cláudio Gomes Maia – Universidade FUMEC  
(Orientador)



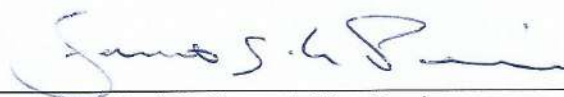
---

Prof. Dr. Rodrigo Moreno Marquês – Universidade FUMEC  
(Examinador Interno)



---

Prof. Dr. Frederico Cesar Mafra Pereira – UFMG  
(Examinador Externo)



---

Prof. Dr. Fernando Silva Parreiras  
Coordenador do Programa de Pós-Graduação em Sistemas de Informação e Gestão do  
Conhecimento da Universidade FUMEC

Belo Horizonte, 17 de fevereiro de 2020.

Eu dedico este trabalho à minha mãe **Eliane da Conceição Marques** (In Memoriam), a minha querida esposa **Elisângela Antunes Silva** e aos meus amados filhos **João Álex Marques Antunes** e **Melissa Marques Antunes**.

# Agradecimentos

Agradeço a minha mãe Eliane, sem ela nada em minha vida teria sido possível.

Agradeço imensamente a minha esposa Elisângela, que me acompanhou ao longo dessa jornada e em tantos momentos foi privada de lazer e das viagens de férias, para que eu pudesse dar continuidade a esse trabalho.

Devo no mínimo um obrigado a todos meus tios, tias e primos, por todo o suporte que me deram em minha infância, por todos puxões de orelha que eu mereci e foram dados.

No meio acadêmico, gostaria de agradecer ao apoio prestado pelo meu orientador Prof. Dr. Luiz Maia, obrigado por me orientar e por estar comigo nessa trajetória.

A Prof. Dr. Frederico Cesar Mafra Pereira e ao Rodrigo Moreno Marques por participarem desse momento.

Professores do Mestrado FUMEC: obrigado pelo aprendizado e por contribuírem para o meu crescimento como pessoa, como estudante e como profissional.

Meus sinceros agradecimentos a FUMEC, por me proporcionar essa oportunidade.

Ao funcionário da secretária Júlio César Teixeira e Silva sempre disposto a ajudar.

Aos amigos Leandro, Presleyson e Rodrigo pela ajuda prestada e por estar comigo ao longo dessa caminhada.

O conhecimento serve para encantar as pessoas, não para humilhá-las.

Mário Sérgio Cortella

# Resumo

A evasão de alunos é um acontecimento mundial e encontra-se presente nas diversas instituições de ensino brasileiras, sejam elas públicas ou privadas. Os Gestores escolares precisam de ferramentas que possibilite gerar conhecimento da instituição de ensino e assim traçar estratégias para lidar com os desafios no ambiente institucional. Uma técnica que vem se mostrando eficiente e capaz de fornecer conhecimento acerca do passado, presente e futuro da instituição, possibilitando aos gestores promoveres ações eficientes de combate à evasão escolar é a Machine Learning - ML. Neste contexto, o presente trabalho utilizou a metodologia CRISP-DM para a aplicação do processo de ML. O objetivo da pesquisa foi testar e/ou aplicar técnicas de ML na busca de possíveis razões para a evasão escolar. Entende-se como contribuição central deste trabalho a geração de modelos de predição que permita apoiar os gestores em ações de combate à evasão. Foram testados 6 algoritmos de classificação. O melhor algoritmo foi *Two-class Boosted Decision Tree* que apresentou um Accuracy de 0,964 e uma curva AUC de 0,994. O segundo foi o *Two-Class Neural Network* que apresentou uma Accuracy de 0,944 e uma curva AUC de 0,988. Assim, a partir deste conhecimento gerado, será possível traçar estratégias de combate à evasão proporcionando uma mudança organizacional significativa na forma de compreender e combater a evasão escolar.

**Palavras-chave:** Aprendizagem de Máquina, Evasão Escolar, Combate à Evasão, CRISP-DM.



# Abstract

Student dropout is a worldwide event and is present in several Brazilian educational institutions, whether public or private. School managers need tools that make it possible to generate knowledge of the educational institution and thus outline strategies to deal with challenges in the institutional environment. Machine Learning is a technique that has been shown to be efficient and capable of providing knowledge about the past, present and future of the institution, enabling managers to promote efficient actions to combat school dropout. In this context, the present work used the CRISP-DM methodology for the application of the ML process. The objective of the research was to test and/or apply ML techniques in search of possible reasons for school dropout. The central contribution of this work is understood to be the generation of prediction models that allow the support of managers in actions to combat evasion. Six classification algorithms were tested. The best algorithm was "Two-class Boosted Decision Tree" which presented an Accuracy of 0.964 and an AUC curve of 0.994. The second was the "Two-Class Neural Network" which presented an Accuracy of 0.944 and an AUC curve of 0.988. Thus, from this knowledge generated, it will be possible to outline strategies to combat dropout by providing a significant organizational change in the way of understanding and combating school dropout.

**Key-words:** Machine Learning, dropout School, Evasion Combating, CRISP-DM.

# Lista de ilustrações

Figura 1 – Jovens de 16 anos com ensino fundamental . . . . .	23
Figura 2 – Jovens de 19 anos com ensino médio . . . . .	23
Figura 3 – Jovens de 16 anos que não concluíram o ensino fundamental . . . . .	24
Figura 4 – Jovens de 19 anos que não concluíram ensino médio . . . . .	24
Figura 5 – Trajetória dos estudante - Brasil 2010 a 2016 . . . . .	25
Figura 6 – Trajetória dos estudantes por categoria administrativa - Brasil 2010 a 2016 . . . . .	25
Figura 7 – Trajetória dos estudantes por modalidade de ensino - Brasil 2010 a 2016	26
Figura 8 – Modelo de abandono . . . . .	28
Figura 9 – Modelo conceitual para o abandono . . . . .	29
Figura 10 – Categorias de aprendizagem de máquina . . . . .	33
Figura 11 – Árvore de decisão . . . . .	36
Figura 12 – Floresta Aleatória . . . . .	37
Figura 13 – Representação gráfica de SVM . . . . .	38
Figura 14 – Modelo de funcionamento das redes neuronais . . . . .	39
Figura 15 – Regressão Logística . . . . .	40
Figura 16 – Relatório anual - 2018 divulgação . . . . .	48
Figura 17 – Total de matrículas por forma de pagamento . . . . .	49
Figura 18 – Matrículas em processo e carga horária concluída . . . . .	50
Figura 19 – Matrículas concluídas por modalidade e tipo de curso . . . . .	51
Figura 20 – Matrículas concluídas por tipo curso . . . . .	52
Figura 21 – Matrículas concluídas por êxito tecnológico . . . . .	53
Figura 22 – Matrículas por sexo e maioridade . . . . .	53
Figura 23 – Aproveitamento por status em 2017 . . . . .	55
Figura 24 – Aproveitamento por status em 2018 . . . . .	55
Figura 25 – Cronograma do sistema acadêmico - SA . . . . .	56
Figura 26 – Lançamento de frequência do aluno - SA . . . . .	56
Figura 27 – Pesquisa curso no sistema acadêmico - SA . . . . .	57
Figura 28 – Cadastro do aluno no sistema acadêmico - SA . . . . .	57
Figura 29 – Fluxo do modelo ML . . . . .	59
Figura 30 – Fluxo do modelo CRISP-DM . . . . .	61
Figura 31 – Ambiente de desenvolvimento do SQL Server . . . . .	64
Figura 32 – Ambiente de inicial do visual studio data tools . . . . .	65
Figura 33 – Machine learning studio - MLS . . . . .	66
Figura 34 – Tela principal do machine learning studio - MLS . . . . .	67
Figura 35 – Experimentos, módulos e base de dados . . . . .	68

Figura 36 – Diagrama azure machine learning studio - MLS . . . . .	68
Figura 37 – Criação do Banco de Dados . . . . .	70
Figura 38 – ETL - Carga das tabelas . . . . .	71
Figura 39 – Script da consulta SQL com todas as tabelas . . . . .	72
Figura 40 – Algoritmo Two-Class Support Vector Machinep . . . . .	84
Figura 41 – Algoritmo Two-Class Support Vector Machinep . . . . .	84
Figura 42 – Algoritmo Two-Class Support Vector Machinep . . . . .	85
Figura 43 – Algoritmo Two-Class logistic Regression . . . . .	87
Figura 44 – Algoritmo Two-Class logistic Regression . . . . .	88
Figura 45 – Algoritmo Two-Class logistic Regression . . . . .	88
Figura 46 – Algoritmo Two-Class Locally-Deep VSM . . . . .	89
Figura 47 – Algoritmo Two-Class Locally-Deep VSM . . . . .	90
Figura 48 – Algoritmo Two-Class Locally-Deep VSM . . . . .	90
Figura 49 – Algoritmo Two-Class Decision Jungle . . . . .	91
Figura 50 – Algoritmo Two-Class Decision Jungle . . . . .	91
Figura 51 – Algoritmo Two-Class Decision Jungle . . . . .	92
Figura 52 – Algoritmo Two-Class Neural Network . . . . .	93
Figura 53 – Algoritmo Two-Class Neural Network . . . . .	93
Figura 54 – Algoritmo Two-Class Neural Network . . . . .	94
Figura 55 – Algoritmo Two-Class Boosted Decision Tree . . . . .	95
Figura 56 – Algoritmo Two-Class Boosted Decision Tree . . . . .	95
Figura 57 – Algoritmo Two-Class Boosted Decision Tree . . . . .	96

# Lista de tabelas

Tabela 1 – Quantidade de publicações nas bases de dados eletrônicos . . . . .	18
Tabela 2 – Principais modelos de abandono . . . . .	29
Tabela 3 – Causas de Evasão segundo alguns autores . . . . .	30
Tabela 4 – Visualizações de páginas por site . . . . .	47
Tabela 5 – Crescimento nas Redes Sociais . . . . .	47
Tabela 6 – Status do aluno no curso . . . . .	54
Tabela 7 – Síntese dos objetivos relacionados com coleta de dados . . . . .	59
Tabela 8 – Quantidade de tabelas por tema . . . . .	70
Tabela 9 – Relação de atributos . . . . .	73
Tabela 10 – Relação dos atributos que foram excluídos . . . . .	75
Tabela 11 – Relação de atributos derivados . . . . .	75
Tabela 12 – Grupo de idade criado a partir do atributo DatNasc . . . . .	77
Tabela 13 – Relação de atributos mesclados . . . . .	78
Tabela 14 – Base final utilizada na geração do modelo ML . . . . .	79
Tabela 15 – Desempenho dos algoritmos . . . . .	83

# Lista de abreviaturas e siglas

AD	<i>Árvore de Decisão</i>
AM	<i>Aprendizagem de Máquina</i>
ANN	<i>Artificial Neural Networks</i>
AP	<i>Aprendizagem Profunda</i>
API	<i>Application Programming Interface</i>
AUC	<i>Area Under the Curve</i>
AVA	<i>Ambiente Virtual Aprendizagem</i>
BDTD	<i>Biblioteca Digital Brasileira de Teses e Dissertações</i>
CAPES	<i>Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior</i>
CART	<i>Classification And Regression Tree</i>
CEEES	<i>Comissão Especial de Evasão do Ensino Superior</i>
CHAID	<i>Chi-square Automatic Interaction Detector</i>
CRISP-DM	<i>Cross-industry standard process for data mining</i>
DATASET	<i>Conjunto de Dados</i>
DLNN	<i>Deep Learning Neural Network</i>
DN	<i>Departamento Nacional</i>
DP	<i>Deep Learning</i>
EDM	<i>Educational Data Mining</i>
EEM	<i>Enhanced Evaluate Model</i>
ENADE	<i>Exame Nacional de Desempenho dos Estudantes</i>
ETL	<i>Extract Transform Load</i>
FIC	<i>Formação Inicial Continuada</i>
FUMEC	<i>Fundação Municipal Educação Comunitária</i>

GTI	<i>Gerência Tecnologia Informação</i>
IA	<i>Inteligência Artificial</i>
IES	<i>Instituição de Ensino Superior</i>
INEP	<i>Instituto Nacional de Estudos e Pesquisas</i>
KDD	<i>Knowledge Discovery in Databases</i>
MBA	<i>Master Business Administration</i>
MEC	<i>Ministério da Educação</i>
ML	<i>Machine Learning</i>
MLP	<i>Multilayer Perceptron</i>
MLS	<i>Machine Learning Studio</i>
MVS	<i>Máquinas de Vetores de Suporte</i>
PSG	<i>Programa de Gratuidade</i>
RNA	<i>Redes Neurais Artificiais</i>
RNP	<i>Rede Neural Profunda</i>
RNS	<i>Rede Neural Simples</i>
ROC	<i>Receiver Operating Characteristic</i>
SA	<i>Sistema Acadêmico</i>
SciELO	<i>Scientific Electronic Library Online</i>
SEMMA	<i>Sampling, Exploring, Modifying, Modelling and Assessment</i>
SNN	<i>Simple Neural Network</i>
SQL	<i>Structured Query Language</i>
SSAS	<i>SQL Server Analysis Services</i>
SSIS	<i>SQL Server Integration Services</i>
SSRS	<i>SQL Server Reporting Services</i>
SVM	<i>Support Vector Machine</i>
TIC	<i>Tecnologias da Informação e Comunicação</i>
WEKA	<i>Waikato Environment for Knowledge Analysis</i>

# Sumário

	<b>Sumário</b> . . . . .	<b>14</b>
<b>1</b>	<b>INTRODUÇÃO</b> . . . . .	<b>16</b>
1.1	<b>Problema de Pesquisa</b> . . . . .	<b>17</b>
1.2	<b>Motivação e Justificativa</b> . . . . .	<b>18</b>
1.3	<b>Lacuna a ser Explorada</b> . . . . .	<b>18</b>
1.4	<b>Objetivos</b> . . . . .	<b>19</b>
1.4.1	Objetivo geral . . . . .	19
1.4.2	Objetivos específicos . . . . .	19
1.5	<b>Contribuições</b> . . . . .	<b>20</b>
1.6	<b>Adequação do Projeto de Mestrado e a Linha de Pesquisa</b> . . . . .	<b>20</b>
1.7	<b>Estrutura da Dissertação</b> . . . . .	<b>21</b>
<b>2</b>	<b>REFERENCIAL TEÓRICO</b> . . . . .	<b>22</b>
2.1	<b>Evasão Escolar</b> . . . . .	<b>22</b>
2.1.1	Tipos de evasão . . . . .	26
2.1.2	Modelos de evasão . . . . .	27
2.1.3	Causas de evasão . . . . .	30
2.1.4	Questões financeiras e a evasão . . . . .	31
2.2	<b>Machine Learning</b> . . . . .	<b>32</b>
2.2.1	Aprendizagem supervisionada . . . . .	34
2.2.1.1	Árvore de decisão . . . . .	34
2.2.1.2	Floresta aleatória . . . . .	36
2.2.1.3	Máquina de vetores de suporte . . . . .	37
2.2.1.4	Redes neurais artificiais . . . . .	38
2.2.1.5	Regressão logística . . . . .	39
<b>3</b>	<b>TRABALHOS RELACIONADOS</b> . . . . .	<b>41</b>
<b>4</b>	<b>CONTEXTUALIZAÇÃO AMBIENTE</b> . . . . .	<b>45</b>
4.1	<b>Tecnologia da Informação na Educação</b> . . . . .	<b>45</b>
4.2	<b>Caracterização da Instituição Objeto da Pesquisa</b> . . . . .	<b>45</b>
4.3	<b>Sistema Acadêmico - SA</b> . . . . .	<b>56</b>
<b>5</b>	<b>METODOLOGIA</b> . . . . .	<b>58</b>
5.1	<b>Caracterização da Pesquisa</b> . . . . .	<b>58</b>
5.2	<b>CRISP-DM</b> . . . . .	<b>60</b>

<b>5.3</b>	<b>Compreensão do Negócio</b>	<b>62</b>
5.3.1	Determinar os objetivos do negócio	62
5.3.2	Avaliar a situação	62
5.3.3	Ferramentas e técnicas	63
<b>5.4</b>	<b>Compreensão dos Dados</b>	<b>69</b>
5.4.1	Coleta de dados	69
5.4.2	Descrição dos dados	71
5.4.3	Exploração dos dados	73
5.4.4	Qualidade dos dados	74
<b>5.5</b>	<b>Preparação dos Dados</b>	<b>74</b>
5.5.1	Seleção de dados	74
5.5.2	Limpeza de dados	75
5.5.3	Construção de dados	75
5.5.4	Integração de dados	78
5.5.5	Formatação de dados	78
<b>5.6</b>	<b>Modelagem</b>	<b>80</b>
<b>5.7</b>	<b>Avaliação</b>	<b>81</b>
<b>5.8</b>	<b>Implementação</b>	<b>81</b>
<b>6</b>	<b>TESTES E RESULTADOS OBTIDOS</b>	<b>83</b>
6.1	Algoritmo Two-Class Support Vector Machine	83
6.2	Algoritmo Two-Class logistic Regression	87
6.3	Algoritmo Two-Class Locally-Deep VSM	88
6.4	Algoritmo Two-Class Decision Jungle	90
6.5	Algoritmo Two-Class Neural Network	92
6.6	Algoritmo Two-Class Boosted Decision Tree	94
<b>7</b>	<b>CONSIDERAÇÕES FINAIS</b>	<b>97</b>
	<b>REFERÊNCIAS</b>	<b>100</b>
	<b>ANEXO A – ANEXOS</b>	<b>105</b>
	<b>ANEXO B – ANEXOS</b>	<b>113</b>
	<b>ANEXO C – ANEXOS</b>	<b>114</b>
	<b>ANEXO D – ANEXOS</b>	<b>118</b>
	<b>ANEXO E – ANEXOS</b>	<b>130</b>
	<b>ANEXO F – ANEXOS</b>	<b>133</b>



# 1 INTRODUÇÃO

Esta pesquisa trata do uso de Aprendizagem de Máquina (AM), em inglês *Machine Learning* (ML), como ferramenta para ajudar os gestores escolares na verificação e combate à evasão escolar. A partir da análise dos dados históricos armazenados em bancos de dados, objetivou-se a construção de modelos analíticos que possam aprender com dados, identificar padrões e ajudar na tomada de decisão.

O uso da tecnologia na educação requer, um olhar mais extenso. É preciso que haja o envolvimento de novas formas de ensinar, aprender condizente com a sociedade tecnológica, que deve se caracterizar pela integração, complexidade e convivência com a diversidade de linguagens e formas de representar o conhecimento(MOURA; ZIVIANI; OLIVEIRA, 2018).

Vive-se em um cenário de extrema competitividade e a informação passou a ser um efetivo organizacional precioso. Com a diminuição da distância entre as inovações tecnológicas, o processo de tomada de decisão e o tempo, tornaram-se um diferencial diante a concorrência(LOPES; MUÝLDER; JUDICE, 2012).

Atualmente, as organizações geram muitos dados e, em razão disso, frequentemente não conseguem processá-los, ou até mesmo interpretá-los, chegando assim a não ter respostas em tempo hábil para tomar decisões (TURBAN; VOLONINO, 2013).

A grande maioria das empresas e organizações da nossa sociedade utiliza sistemas de gestão de informação modernos. Esses sistemas são construídos com base de dados relacionais fundamentalmente vocacionados para armazenar, com altos níveis de eficiência, os resultados das operações recorrentes das organizações (CALDEIRA, 2012).

A instituição escolhida para a utilização dos dados enquadra-se neste contexto citado por (CALDEIRA, 2012). A instituição é suportada por vários sistemas, com base de dados relacionais. Estes sistemas são constituídos muitas vezes por módulos, que estabelecem a base do processo de negócio da instituição. Seus bancos de dados são grandes e estão desenhados e formatados por transações, dificultando a recolha para a análise e tratamento da informação.

Difícilmente existe numa instituição uma plataforma que seja capaz de transformar grandes volumes de dados, das mais diversas áreas, em conhecimentos específicos para os gestores escolares. A recolha da informação ainda é lenta, parcial, suscetível de erros que provoca uma sobrecarga adicional no sistema, um aumento significativo nos chamados abertos ao setor de informática, e não oferece de uma forma rápida, adequada e eficaz, o suporte à tomada de decisão.

ML é uma área da ciência da computação que evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial(IA). Desse modo, estuda meios para que máquinas sejam capazes de tarefas que seriam executadas por pessoas, bem como, é uma programação usada nos computadores, construída por regras anteriormente estipuladas que autorizam que os computadores tomem decisões com base nos dados disponíveis. Nesse sentido, a base do funcionamento são os algoritmos, que são sequências estipuladas e instruções que vão ser seguidas por um computador (BIAMONTE et al., 2017).

Esse estudo torna-se expressivo não apenas para apoiar a análise de dados sobre a evasão, mas também para permitir que os modelos analíticos aqui desenvolvidos possam servir de base para outras questões associadas a dados acadêmicos e administrativos como referência para outras instituições de ensino.

## 1.1 Problema de Pesquisa

A evasão escolar está presente em diversos níveis e modalidades, alcançando desde a educação básica até a superior. Esse assunto vem sendo estudado em diversas pesquisas acadêmicas, nas suas diversas modalidades de ensino. Entender as causas, compreender como esse processo ocorre nas instituições de ensino e conhecer a visão dos gestores educacionais sobre esta problemática, pode auxiliar na assimilação deste fenômeno social e nas ações preventivas para sua redução. A atenuação da evasão reduziria prejuízos sociais, econômicos, políticos, acadêmicos e financeiros a todos os envolvidos no processo educacional, desde o estudante até os órgãos governamentais.

Especialmente nos sistemas acadêmicos, existe um crescente volume de dados que incluem: matrícula, cursos, professores, disciplinas, frequências, notas, entre outras informações. Os problemas de tomada de decisão em instituições de ensino são muitos mais amplos e complexos, envolvendo riscos e incertezas. Para tomar a decisão correta, os gestores precisam de informações confiáveis, expressivas e de forma ativa. É nesse momento que as técnicas de ML podem contribuir para a tomada de decisão.

Nessas circunstâncias, o problema a ser resolvido por esta pesquisa é: **de que forma a aplicação de técnicas de ML pode auxiliar na identificação de grupos de estudantes em risco de evasão escolar?**

## 1.2 Motivação e Justificativa

Observou-se a importância desse tema a partir de um levantamento feito no Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), na Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) e *Scientific Electronic Library Online* (SciELO) no período de 2010 a 2019. A partir desse levantamento tornou-se explícito a ausência de estudos envolvendo “evasão escolar” e “Aprendizagem de Máquina”, apontando a necessidade do presente estudo com perspectivas a sanar a lacuna acadêmica existente.

A tabela 1 demonstra a ausência de estudos envolvendo Técnicas de ML voltados para a área de combate à evasão. Dessa forma, é indispensável o desenvolvimento de modelos analíticos eficientes e eficazes para identificação, predição, avaliação e acompanhamento de estudantes em risco de evasão, possibilitando o planejamento e a adoção de medidas proativas no intuito de minimizar o problema.

Tabela 1 – Quantidade de publicações nas bases de dados eletrônicas

<b>Termos pesquisados</b>	<b>CAPES</b>	<b>BDTD</b>	<b>SciELO</b>
"Evasão escolar"	304	297	61
"Machine Learning"	350.622	954	168
"Aprendizagem de Máquina"	35	308	8
"Data Mining"	144.896	664	329
"Mineração de Dados"	307	715	104
"Evasão escolar"AND "Machine Learning"	0	0	0
"Evasão escolar"AND "Aprendizagem de Máquina"	0	0	0
"Evasão escolar"AND "Data Mining"	1	3	1
"Evasão escolar"AND "Mineração de Dados"	4	1	0
"School Dropout"AND "Machine Learning"	53	0	0
"School Dropout"AND "Data Mining"	35	0	1

Fonte: Elaborado pelo Autor (2019).

Com isso, o fundamento desse trabalho é contribuir na busca de razões para a evasão escolar por meio da utilização de técnicas de ML. Os sistemas gerenciais comuns não atendem esse quesito e a instituição, objeto de estudo, tem atualmente dificuldade de conseguir identificar alunos em risco de evasão por intermédio dos relatórios gerados pelo SA.

## 1.3 Lacuna a ser Explorada

O foco desse estudo foi buscar informações de tecnologias relacionadas à ML que possam assim permitir a elucidação dos possíveis motivos da evasão escolar tendo como base os registros históricos do banco de dados do sistema acadêmico da instituição.

Tendo isto em mente, a lacuna examinada foi uma identificação de técnicas ML que possam ajudar os gestores escolares a encontrar alunos em risco de evasão. Pretendeu-se, portanto, a partir desse cenário científico e tecnológico, avançar nos estudos e apresentar direcionamentos que auxiliem nesse propósito.

O estudo permitiu a comunicação com conceitos interdisciplinares como a recuperação da informação, a organização e a representação da informação. Com isso, as áreas de Ciência da Computação, Ciência da Informação, Aprendizagem de Máquina e Inteligência Artificial foram parte do objeto de estudo.

## 1.4 Objetivos

Nesta seção, são apresentados os objetivos, geral e específicos, que orientaram a construção dessa pesquisa.

### 1.4.1 Objetivo geral

O presente trabalho tem como objetivo testar e aplicar técnicas de ML na busca de possíveis razões para a evasão escolar.

### 1.4.2 Objetivos específicos

Para atingir esse objetivo principal, faz-se necessário obter alguns objetivos mais específicos, que são:

1. Realizar um levantamento bibliográfico para conhecer as possíveis causas da evasão e os modelos e teorias relacionadas à questão;
2. Verificar com o público alvo (usuários chave) os registros (variáveis) dos alunos quanto às informações do curso, data de matrícula, situação, horário do curso;
3. Identificar quais dados e informações, conforme levantamento bibliográfico e pesquisa realizada com o público alvo, estão presentes no banco de dados e que precisam ser considerados para elaboração do modelo;
4. Selecionar, relacionar e correlacionar as características mais relevantes para compor o banco de dados das amostras;
5. Preparar os conjuntos de dados para treinamento e teste do modelo;
6. Treinar, testar e validar o modelo;
7. Analisar os resultados obtidos com o modelo proposto;

## 1.5 Contribuições

Este projeto é muito relevante para o negócio das instituições de ensino como um todo. Na realidade até a presente data, já ocorreram algumas iniciativas nesse sentido, mas atualmente as soluções relativas são ainda limitadas. Assim, esse projeto pretende provocar um forte impacto no tratamento dos dados escolares existentes, em informação útil ao negócio da educação.

A introdução de ferramentas analíticas permitirá uma gestão muito mais eficiente, através da análise dos principais indicadores de negócio, assim como na identificação de padrões e tendências de suporte ao negócio da educação.

A partir da pesquisa pretende-se encontrar alunos em risco de evasão, possibilitando que os gestores a tomem uma atitude planejando reverter o acontecimento.

Além disso, o estudo poderá ser utilizado como suporte para outros estudos, gerando possibilidade de comparação com outros trabalhos e reflexões para outras instituições.

## 1.6 Adequação do Projeto de Mestrado e a Linha de Pesquisa

A presente pesquisa está aderente ao Programa de Mestrado em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC, com uma questão interdisciplinar entre Administração (Gerenciamento de Projetos) e Ciência da Computação (com Inteligência Artificial e processamento de linguagem natural). O estudo retrata técnicas de ML aplicadas a uma instituição de educação com o intuito de descobrir alunos em risco de evasão.

O presente trabalho compreende os estudos de técnicas de ML visando apresentar metodologias que auxiliem nas descobertas de alunos em grupo de risco de evasão. O projeto está classificado na linha de pesquisa Tecnologia e Sistemas de Informação na trilha T2 - Cognição, Aprendizado de Máquina e Recuperação da Informação. Com isso, espera-se contribuir com estudos já existentes.

Em alinhamento ao programa, o projeto irá contribuir para análise de técnicas, métodos e boas práticas utilizadas AI e ML. Será feita uma revisão da literatura permitindo a exploração de conceitos interdisciplinares com várias áreas do conhecimento. Trata-se de um assunto que estabelece relação com ramos de conhecimentos correlatos como a recuperação da informação, atingindo as áreas da Ciência da Informação, Ciência da Computação, a Inteligência Artificial entre outras.

## 1.7 Estrutura da Dissertação

Além desse primeiro capítulo que abordou o tema, cenário, problema, justificativas, objetivos, contribuições e a aderência ao Programa de Pós-graduação, a dissertação contém outros 7 capítulos.

O segundo capítulo aborda o referencial teórico dos construtos envolvidos no problema de pesquisa. Relata sobre a evasão escolar, ML e suas técnicas.

O terceiro capítulo apresenta os trabalhos relacionados.

O quarto capítulo contempla informações relativas à contextualização do ambiente.

O quinto capítulo apresenta a metodologia utilizada para o desenvolvimento do projeto.

O sexto capítulo apresenta os resultados obtidos, seguido das considerações finais, as referências e os apêndices.

## 2 REFERENCIAL TEÓRICO

### 2.1 Evasão Escolar

De um modo geral a evasão pode ser definida como a não conclusão de uma unidade educacional, ou seja, de qualquer modalidade de educação, que guie o alunado a um entendimento habilitado (FIALHO et al., 2014).

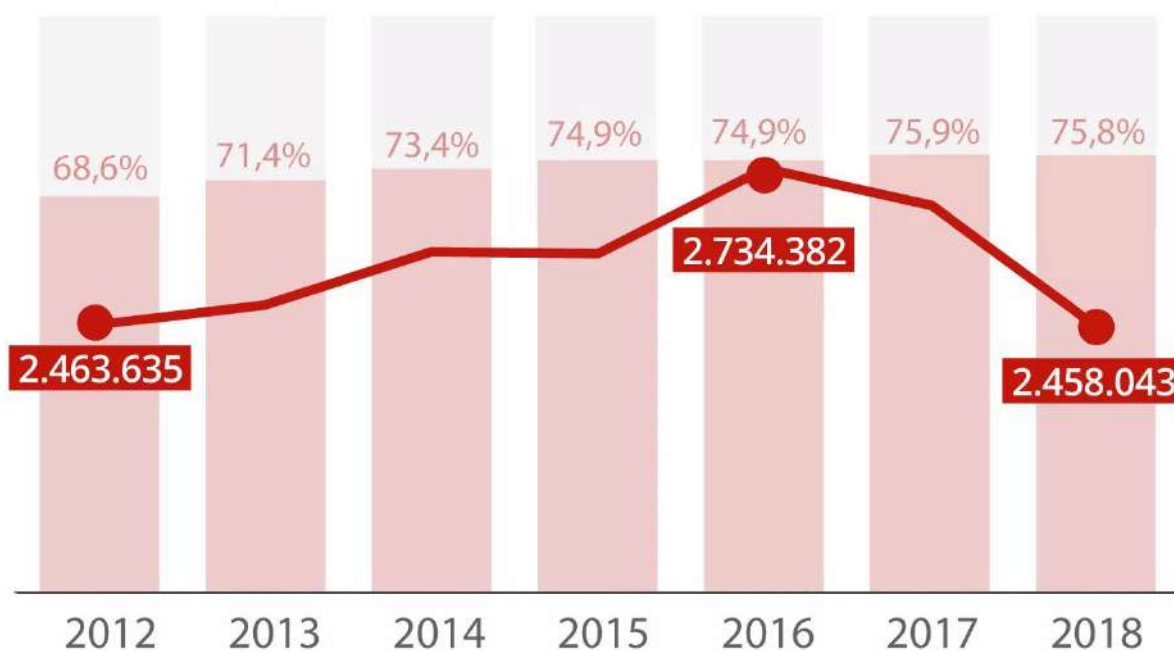
O que se percebe na literatura internacional é que parece não existir um consentimento referente à percepção do termo evasão escolar, especialmente nas Instituições de Ensino Superior (IES). Tinto (1975), que é uma das referências no tema, delinea abandono escolar como: a dinâmica de o aluno deixar a (IES) e de modo algum obter o diploma. No Brasil, uma das primeiras iniciativas tomada para se medir a evasão escolar foi praticada pela Comissão Especial de Evasão do Ensino Superior (CEEES), em 1996, constituindo-se ainda uma das mais aplicadas para a efetuação dos cálculos dos déficits de alunos.

Vários autores como Fritsch, Vitelli e Rocha (2016) citam que a evasão escolar se associa com a perda de alunos que começam seus cursos, mas que não conseguem finalizar seus estudos. De acordo com o que se conceitua como “evadido” ocorrem grandes discordâncias nos resultados associados com as taxas de evasão. Vale recordar que essas taxas envolvem um lugar isolado na discussão das políticas públicas, sobretudo, quando se tenta fazer uma análise da eficácia e eficiência da sua aplicação. Dessa maneira uma elucidação sobre diversos tipos e inúmeros métodos de apuração da evasão são imprescindíveis para se ter um conhecimento mais preciso sobre o tipo de evasão em destaque, findando enganos e confrontos entre conteúdos distintos.

No Brasil, escolas públicas e privadas vêm enfrentando um grande problema: o número crescente de evasão escolar. Seguem algumas informações que mostram o problema da evasão escolar em todos os níveis no Brasil.

A figura 1 mostra que o número de jovens com 16 anos que concluiu o ensino fundamental foi diminuindo entre 2016 a 2018.

Figura 1 – Jovens de 16 anos com ensino fundamental



Fonte: IBGE/Pnad Contínua/Todos Pela Educação

No caso do ensino médio, ou seja, entre os jovens de 19 anos que concluíram os estudos, percebemos uma melhora entre os anos de 2012 a 2018 conforme a figura 2.

Figura 2 – Jovens de 19 anos com ensino médio



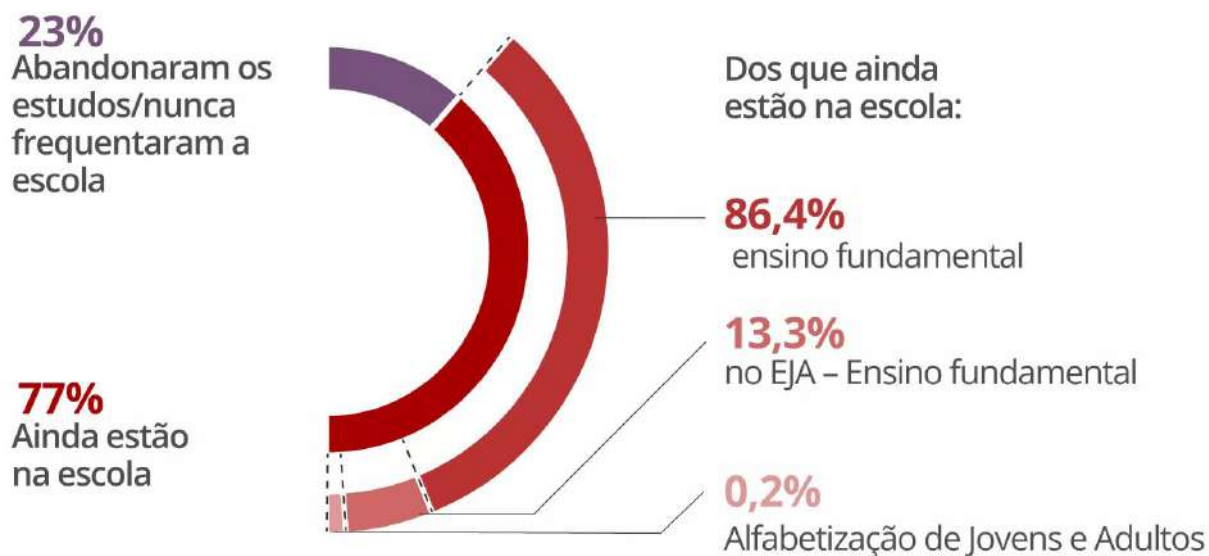
Fonte: IBGE/Pnad Contínua/Todos Pela Educação

Já a figura 3 mostra a realidade dos jovens de 16 anos que ainda não concluíram o ensino fundamental. Entre esses jovens 23% abandonaram os estudos ou nunca frequen-



taram a escola e 77% ainda estão na escola e não finalizaram o ensino fundamental.

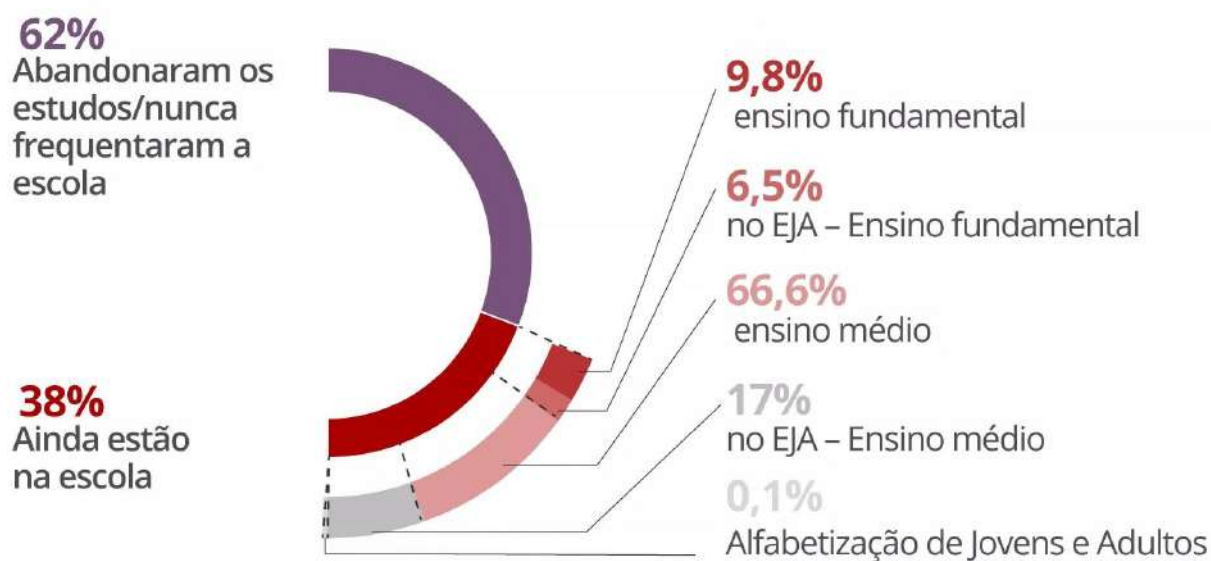
Figura 3 – Jovens de 16 anos que não concluíram o ensino fundamental



Fonte: IBGE/Pnad Contínua/Todos Pela Educação

Entre os alunos de 19 anos que não concluíram o ensino médio o número é bem maior. Entre esses jovens 62% abandonaram os estudos ou nunca frequentaram a escola conforme nos mostra a figura 4. Outro aspecto preocupante é que 16,3% dos 38% que ainda estão na escola, fazem o ensino fundamental.

Figura 4 – Jovens de 19 anos que não concluíram ensino médio

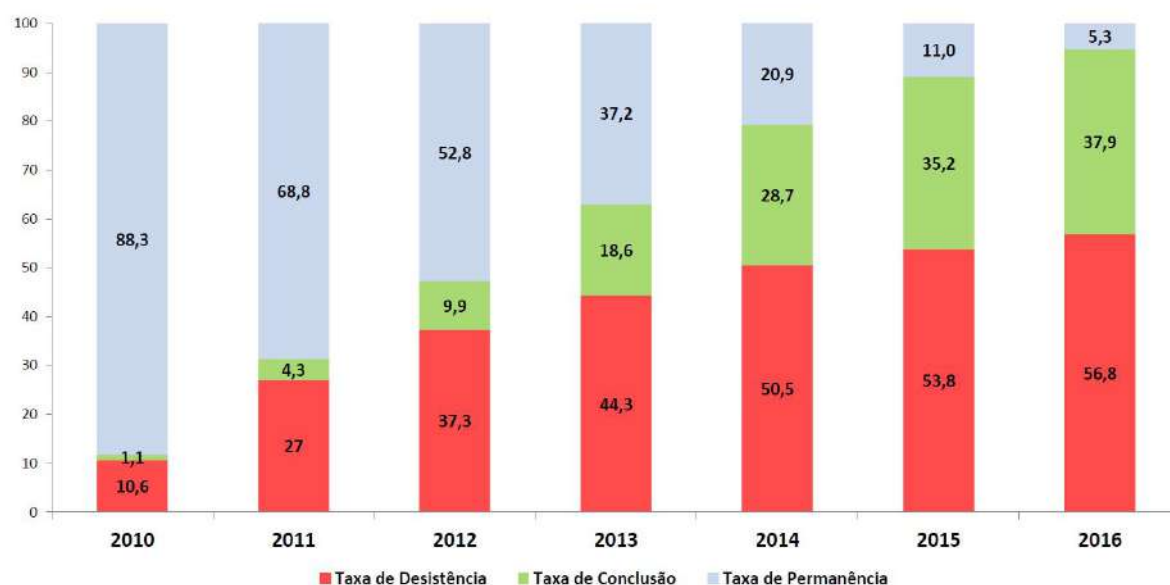


Fonte: IBGE/Pnad Contínua/Todos Pela Educação

O relatório publicado pelo Instituto Nacional de Estudos e Pesquisas (INEP) em

2019 (INEP, 2019) aponta a trajetória dos estudantes das IES no Brasil. Esses indicadores, (i) Taxa de Desistência, (ii) de Conclusão, e (iii) de Permanência são do curso de graduação calculados a partir do acompanhamento da trajetória dos alunos ingressantes em um determinado ano. A figura 5 mostra a evolução dos indicadores de trajetória dos estudantes de 2010 a 2016 no Brasil.

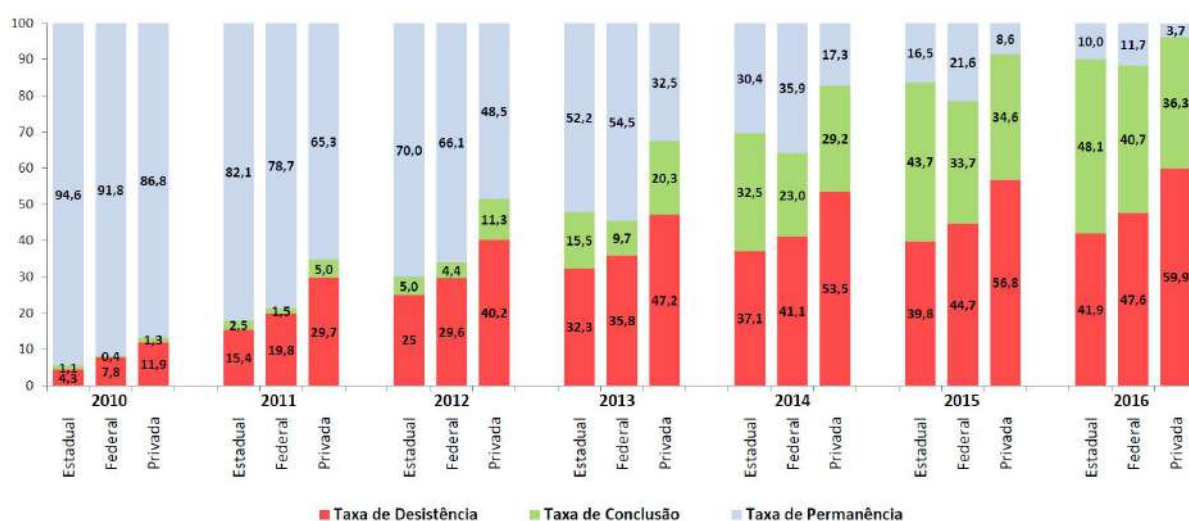
Figura 5 – Trajetória dos estudante - Brasil 2010 a 2016



Fonte: Autor adaptado (INEP, 2019).

Já na figura 6 tem-se uma visão desses indicadores por instituições estaduais, federais e privadas. Verifica-se que ao longo dos anos as instituições privadas e federais tiveram uma taxa de desistência maior.

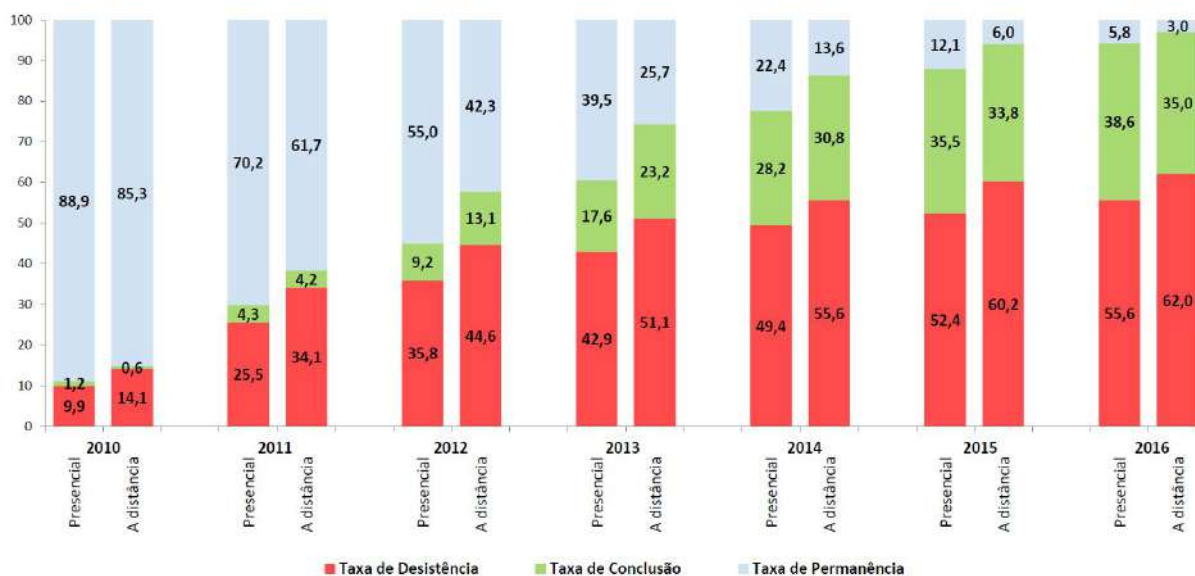
Figura 6 – Trajetória dos estudantes por categoria administrativa - Brasil 2010 a 2016



Fonte: Autor adaptado (INEP, 2019).

A figura 7 mostra os indicadores por modalidade presencial e a distância. Observe que nos cursos a distância a desistência se manteve um pouco maior ao longo dos anos.

Figura 7 – Trajetória dos estudantes por modalidade de ensino - Brasil 2010 a 2016



Fonte: Autor adaptado (INEP, 2019).

### 2.1.1 Tipos de evasão

No Brasil, um marco no estudo da evasão foi o trabalho preliminar da Comissão Especial de Estudos Sobre a Evasão (BRASIL,1996), subordinada à Secretaria de Educação Superior – MEC, instituída em 1995, com objetivo de progredir as reflexões acerca da evasão. Nesse estudo a comissão, apresentou a necessidade de determinar a evasão do curso, da instituição e do sistema (PRESTES; FIALHO, 2018).

- **Evasão do Curso:** a evasão do curso corresponde muitas vezes às mudanças entre áreas similares, ou seja, mobilidade acadêmica.
- **Evasão da Instituição:** já a evasão da instituição se refere à saída do curso e da instituição para outra IES.
- **Evasão do Sistema:** a evasão do sistema já é algo mais sério, pois além do aluno sair do curso e da instituição ele desiste de estudar de vez. Além disso, ainda existe a saída voluntária, cujo cancelamento do curso acontece a pedido do aluno, e a saída involuntária que acontece por intervenção da IES por diferentes razões possíveis (TINTO; CULLEN, 1973). Apesar de a saída involuntária ter uma característica de “expulsão”, ela entra no cálculo das taxas de evasão da mesma forma como o abandono voluntário.

Outro ponto a ser considerado é o período em que o aluno evade do curso, podendo se diferir entre evasão imediata, acontece logo no primeiro ano dos estudos, e a evasão tardia, que é o resultado de um processo gradativo (SANTOS et al., 1994).

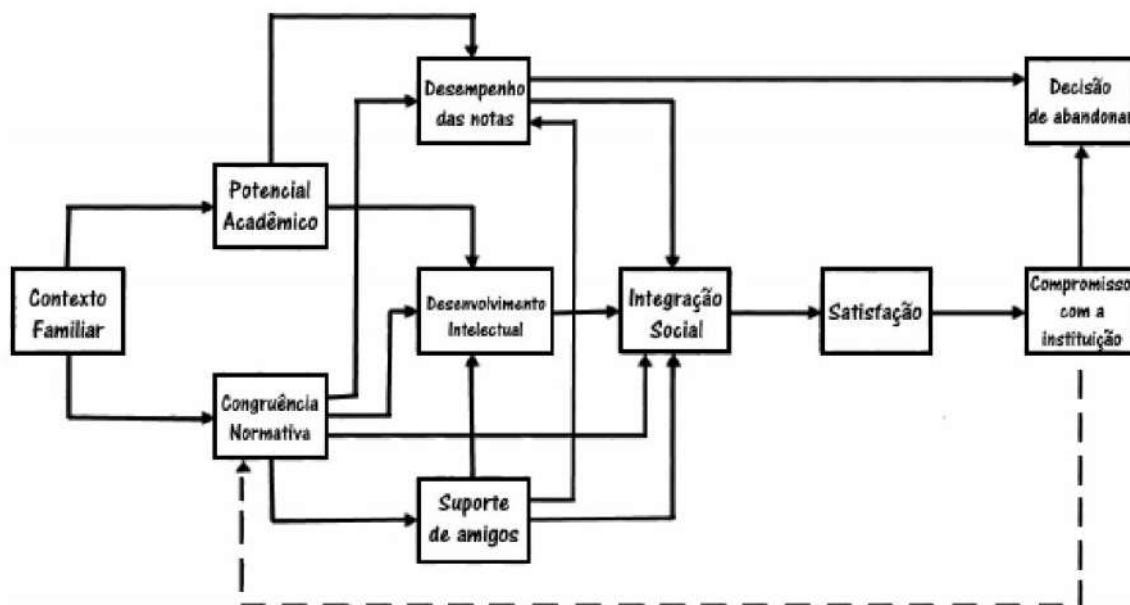
### 2.1.2 Modelos de evasão

Para adotar medidas capazes de minimizar os processos de evasão, as instituições educativas precisam adotar algumas medidas importantes além de conhecerem as causas que motivaram essa ocorrência e suas diferentes nuances. Existem, na literatura, diversas causas que se inter cruzam (PRESTES; FIALHO, 2018). As derivadas dos fatores financeiros, que estão relacionadas com as situações familiares e socioculturais e, em alguns casos, com o trabalho. E as de natureza acadêmica, como as trajetórias de escolaridade, problemas com a didática de metodologias de ensino e, até, com o estado emocional do aluno.

A natureza dos motivos está alusiva aos aspectos psicológicos e individuais, sendo que também esses aspectos causam o abandono. Assim, a teia de relações que contribui para causar evasão é entrelaçada, dificultando destacar aquelas que realmente preponderam na decisão do aluno. Os modelos de Spady (1971) e Tinto (1975) são considerados como as primeiras tentativas de integrar a pluralidade desses fatores em um modelo causal coerente. Ambos se balizam na teoria de suicídio de (WOLFF; DURKHEIM, 1960) e explicam o fenômeno da evasão como resultado de uma integração acadêmico-social insuficiente, o que significa que há um desajuste entre o indivíduo e a instituição.

Nesse entendimento, Spady (1970) fornece uma importante referência para se concentrar a atenção na interação entre características do aluno e as influências, expectativas e solicitações impostas por várias fontes no ambiente universitário. Além do contexto familiar, Spady (1970) utilizou, em seu modelo, as seguintes variáveis independentes: potencial acadêmico, desempenho acadêmico, congruência normativa, suporte em amigos e desenvolvimento intelectual. A figura 8 abaixo apresenta o modelo.

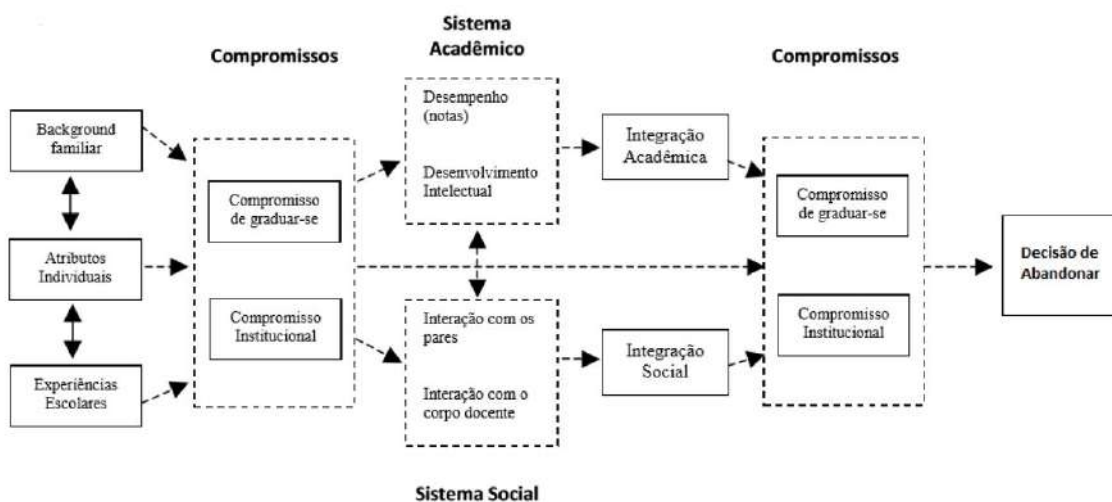
Figura 8 – Modelo de abandono



Fonte: American Journal of Sociology - (SPADY, 1970).

Tinto (1975) insinua que os fatores do ambiente externo à faculdade que podem impactar no processo de abandono, podem ser melhor observados graças a mudança na percepção que o estudante tem sobre seu objetivo de concluir a graduação. Assim, é importante o acompanhamento longitudinal do seu comprometimento com esse objetivo. Também argumenta que um estudante tende a sair da faculdade quando percebe que outra forma alternativa de investimento de tempo, de energia e de recurso lhe trará maior benefício, em relação ao custo de manter-se ao longo do tempo na instituição. Como exemplo o autor cita que uma redução na oferta de empregos disponíveis pode levar os indivíduos a perceberem uma menor probabilidade de que as energias investidas atualmente na instituição renderão retornos aceitáveis no futuro. Na figura 9 o primeiro modelo de Vicent Tinto, desenvolvido em 1975.

Figura 9 – Modelo conceitual para o abandono



Fonte: (TINTO, 1975).

Em suma, o levantamento bibliográfico demonstrou que os principais modelos de estudo do processo de evasão/permanência têm origem nos Estados Unidos, onde há um expressivo acervo de pesquisas realizadas. Na tabela 2 é apresentada uma síntese dos principais modelos desenvolvidos.

Tabela 2 – Principais modelos de abandono

Modelo	Variáveis	Autor
Explicativo do processo de abandono	Contexto familiar; Potencial acadêmico; Congruência normativa; Recursos humanos; Desenvolvimento intelectual; Suporte de amigos; Integração Social; Satisfação; Compromisso com uma instituição;	(SPADY, 1970)
Longitudinal do processo de abandono	Contexto familiar; Atributos individuais; Experiências escolares; Compromisso de graduar-se; Compromisso institucional; Desempenho acadêmico; Desenvolvimento intelectual; Interação com os pares; Interação com os docentes; Integração acadêmica; Integração social;	(SPADY, 1975)
Conceitual de desgaste	Características do estudante antes do ensino superior; Fatores institucionais; Contato com informal corpo docente; Outras experiências universitárias; Resultados educacionais;	(PASCARELLA, 1980)

<p>Conceitual de desgaste do aluno não tradicional</p>	<p>Variáveis de fundo: idade, status de inscrição, residência, objetivos educacionais, ensino médio do ensino médio, etnia, gênero e educação dos pais.</p> <p>Variáveis acadêmicas: habilidades de estudo e hábitos de estudo, apoio acadêmico, absenteísmo, maior certeza, disponibilidade do curso.</p> <p>Variáveis ambientais: finanças, trabalho de incentivo externo, responsabilidades da família, oportunidade de transferência.</p> <p>Variáveis de integração social.</p> <p>Resultados acadêmicos (desempenho acadêmico global).</p> <p>Resultados psicológicos (prêmio prático e desenvolvimento pessoal, satisfação, comprometimento com os objetivos, estresse).</p> <p>Intenção de sair.</p>	<p>(BEAN, METZNER, 1985)</p>
--	--	------------------------------

Fonte: Elaborado pelo Autor (2019).

### 2.1.3 Causas de evasão

Reforçando que no Brasil as escolas públicas e as privadas vêm enfrentando um grande problema do crescente número de evasão escolar. Seguem alguns temas relacionados e seus respectivos autores, que pesquisaram sobre o problema da evasão escolar. Pois, essa ainda é uma problemática atual e que preocupa muitos profissionais na área de educação e nas instituições de ensino como um todo.

Na procura pelas causas do fracasso escolar alguns estudos já mostraram vários fatores relacionados à evasão. A tabela 3 mostra algumas causas de evasão.

Tabela 3 – Causas de Evasão segundo alguns autores

Causas	Autores
Desmotivação	Coelho (2001), Frankola (2001) Neves (2006), Ramminger (2006)
Falta de companheiros presenciais	Frankola (2001), Neves (2006), Longo(2009)
Falta de tempo	Abraed (2008), Almeida (2008), Comarella (2009), ~Censo (2014), Neves (2006), Pacheco (2007), Ramminger (2006)
Falta de disciplina	Coelho (2001)
Problemas familiares	Almeida (2008), Ramminger (2006)
Questão financeira	Abraed (2008), Ramminger (2006)



Acham que curso seria mais fácil	Abraed (2008), Almeida (2008), Ramminger (2006)
Dificuldade de acesso ao computador e internet	Almeida (2008), Pacheco (2007)
Falta de preparo do professor	Sihler e Ferreira (2011)
Alta rotatividade de tutores	Almeida (2008)
Falta de feedback do tutor	Almeida (2008)
Falta de adaptação à EAD	Abraed (2008), Censo (2014), Longo (2009), Prensky (2001)

Fonte: Elaborado pelo Autor (2019).

Tendo em vista a dificuldade que envolve o fenômeno evasão, principalmente em função dos diferentes métodos utilizados para identificar as causas, dos diferentes objetos de estudos, o anexo A apresenta um quadro com o levantamento dos principais trabalhos estudados sobre variáveis que estão relacionadas com a evasão.

#### 2.1.4 Questões financeiras e a evasão

Poucos são, entretanto, aqueles que direcionam o seu foco para as perdas financeiras das instituições educacionais devido à evasão. Um dos poucos estudos nessa linha é o proposto por [Mendonça \(2012\)](#), realizado na Universidade do Rio de Janeiro em 2012. O estudo apresenta prejuízos institucionais, mostrando que o abandono do aluno tem um custo muito alto para a educação no Brasil.

Para [Johann et al. \(2012\)](#), os estorvos no acesso e permanência escolar têm sido características básicas do sistema educacional brasileiro, sendo necessário detectar as raízes deste problema a fim de encontrar possíveis soluções.

Segundo [Lobo \(2012\)](#) diminuir a evasão escolar custa seis vezes menos do que trazer um novo estudante até a instituição de ensino. Combater a evasão é o modo mais eficaz de aumentar o número de matrículas e, ainda, mostrar a realidade do processo, tendo em vista que fazer o aluno concluir seu curso, com qualidade, significa que a escola atingiu seu objetivo.

No caso da Universidade de Brasília, os danos financeiros causados pela evasão, em 2015, ocasionaram um prejuízo estimado em 95,6 milhões de Reais ([PINHEIRO, 2015](#)). Esses exemplos, mesmo pontuais, servem de amostra para um fenômeno que vem ocorrendo em diferentes outras instituições de ensino superior do país, considerando a abrangência do fenômeno. Sobre isso, comenta ([PEREIRA, 2014](#)):

Os números da evasão também se refletem em alto custo financeiro, de acordo com o estudo feito por Solano Portela, diretor Educacional da Universidade



Mackenzie, que estimam os custos da evasão para uma instituição. Segundo ele, se a evasão é de 25%, com uma mensalidade média de R\$ 500,00, só a perda anual de receita para cada mil alunos é de R\$ 375.000,00. Uma instituição com 20 mil alunos chegaria a perder com a evasão R\$ 7.500.000,00 a cada ano.

Outro ponto de análise da evasão sendo regida por questões financeiras é a desigualdade social existente no Brasil. De acordo com os estudos de (KÜCKELHAUS; SANTOS; LUZ, 2018),

[...] a renda tem papel fundamental não só por proporcionar aos mais ricos as melhores condições de estudo (escolas privadas, cursinhos), mas também por possibilitar ao aluno maior oportunidade de escolha da carreira que melhor se adequa às suas aptidões, favorecendo assim a permanência da desigualdade.

Como visto, são muitos os fatores relacionados à evasão escolar. E uma técnica que pode ser utilizada na descoberta desse fenômeno é ML. Os algoritmos de ML são capazes de identificar o perfil desses estudantes a partir de dados de outros estudantes. Na próxima seção serão analisados alguns desses algoritmos.

## 2.2 Machine Learning

Com o uso progressivo das ferramentas tecnológicas, as organizações são diariamente responsáveis pela produção de enormes conjuntos de dados. E esses dados contêm informações importantes para a atividade das mesmas.

O processamento desses grandes conjuntos de dados necessita de auxílio de ferramentas da estatística e matemática. Pode-se considerar ML como uma área da ciência da computação que atua sobre diversos métodos e possui dois principais objetivos, que é a capacidade de aprendizagem e o desempenho preditivo. Efetivar em um computador a capacidade de aprendizagem tem sido um dos maiores desafios atualmente (VASCONCELOS, 2017). Em conformidade com (MICHALSKI, 1983) o processo de aprendizagem computacional possui um importante propósito de estudo da ML, assim como a capacidade de previsão da ML.

Enquanto a IA é a ciência mais ampla em simular as habilidades humanas, ML pode ser definida como um método de análise de dados que automatiza a construção de modelos analíticos, ou seja, é um ramo da IA baseado na ideia que os sistemas podem aprender com dados, identificar padrões e tomar decisões com intervenção humana mínima. Quanto mais dados forem inseridos, mais experiência o computador obtém, o que torna o seu desempenho melhor (CIELEN; MEYSMAN; ALI, 2016).

ML compreende diversos métodos que partilham os mesmos objetivos: a capacidade preditiva do modelo e automatizar o processo de treino com a base de dados inserida. A ML vem maximizar o desempenho preditivo dos modelos por meio de ferramentas estatísticas (ALPAYDIN, 2004). A ML trabalha com modelos direcionados para a previsão "out of sample", isto é, embora os modelos sejam construídos com uma determinada amostra de dados pretende-se que tenham um bom desempenho quando efetuam previsões sobre novos dados inseridos (SOUZA et al., 2016).

ML pode ser descrito como o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento de forma automática (MITCHELL et al., 2006).

De acordo com Silva e Zhao (2016), a ML está relacionada ao estudo e desenvolvimento de algoritmos capazes de fazer com que computadores aprendam, sem terem sido explicitamente programados. As técnicas de aprendizado de máquina podem ser utilizadas em diversas áreas. Para tal, é necessário traduzir o problema a ser tratado para o domínio da ML que, em geral, requer um conjunto de características como entrada e produz como saída um critério de agrupamento ou classificação.

ML é uma área da ciência da computação que pesquisa os meios para que as máquinas consigam fazer tarefas que seriam executadas por pessoas. É uma programação usada nos computadores, construída por regras anteriormente estipuladas que autorizam que os computadores tomem decisões com base nos dados disponíveis. De acordo com essas programações, a máquina tem uma habilidade para tomar decisões que podem resolver problemas (FENG et al., 2018).

A figura 10 ilustra como os algoritmos de ML normalmente são agrupados segundo três paradigmas apesar dessa abordagem diferir um pouco de acordo com alguns autores:

Figura 10 – Categorias de aprendizagem de máquina



O autor adaptado de (MONARD; BARANAUSKAS, 2003).

- **Aprendizado supervisionada:** Essa categoria é mais utilizada em ML, pois lida com algoritmos que precisam de exemplos rotulados para treinar os modelos.
- **Aprendizado Não Supervisionada:** Essa categoria utiliza algoritmos que não precisam de nenhum tipo de rótulo e sem nenhuma interação humana.
- **Aprendizado Semi-supervisionado:** Essa categoria mescla as características do aprendizado supervisionado e do aprendizado não supervisionado. Essa abordagem é útil em casos como grandes bases de dados.

Nesse trabalho serão utilizados os algoritmos de aprendizagem supervisionada, visto que precisa-se fazer uma modelagem preditiva para a evasão escolar.

### 2.2.1 Aprendizagem supervisionada

O aprendizado supervisionado está associado aos modelos preditivos, e as suas tarefas mais comuns são a classificação e a regressão (WITTEN et al., 2016). Os modelos preditivos frequentemente aplicam funções de aprendizado supervisionado para estimar valores desconhecidos ou futuros de variáveis dependentes em função das características das variáveis independentes relacionadas. Modelos preditivos têm o objetivo específico que nos permite prever os valores desconhecidos de variáveis de interesse a partir de valores conhecidos de outras variáveis. O formato da previsão pode ser pensado como um mapeamento de aprendizagem a partir de um conjunto de entrada como um vetor de medições e uma saída como um escalar (HAN; PEI; KAMBER, 2011).

- **Classificação:** É uma subcategoria de aprendizagem supervisionada que tem como proposta analisar a entrada e atribuir um rótulo a ela. Geralmente são usados quando as previsões são de natureza distinta, ou seja, um simples 0 ou 1. Seu resultado é um valor discreto, como por exemplo, se um aluno evadiu ou não.
- **Regressão:** É outra subcategoria de aprendizagem supervisionada que tem como proposta modelar a dependência do rótulo de dados para determinar como o rótulo será alterado à medida que os valores dos recursos forem variando. É usado quando o valor que está sendo previsto diverge de um simples 0 ou 1.

A principal diferença entre ambos os modelos preditivos é que, enquanto a classificação prevê rótulos categóricos (discretos, não ordenados) para os dados, a regressão estabelece modelos de funções com valores contínuos.

#### 2.2.1.1 Árvore de decisão

Um algoritmo do tipo árvore de decisão (AD) é uma ferramenta preditiva aplicável na resolução de problemas de regressão e classificação em diversas áreas. Normalmente

um algoritmo de árvore de decisão é construídos de forma a dividir uma base de dados em diferentes condições. Pode-se dizer que esse método é prático e bastante utilizado na aprendizagem supervisionada e o seu trajeto desde a “raiz” até à “folha” corresponde a uma regra de classificação.

Géron (2019) enfatiza que as árvores de decisão são os algoritmos de ML mais versáteis, que podem executar tarefas de classificação e regressão. Eles são algoritmos poderosos, capazes de ajustar conjuntos de dados complexos. Em uma análise de decisão, seu algoritmo pode ser usado para representar visualmente e explicitamente as decisões tomadas.

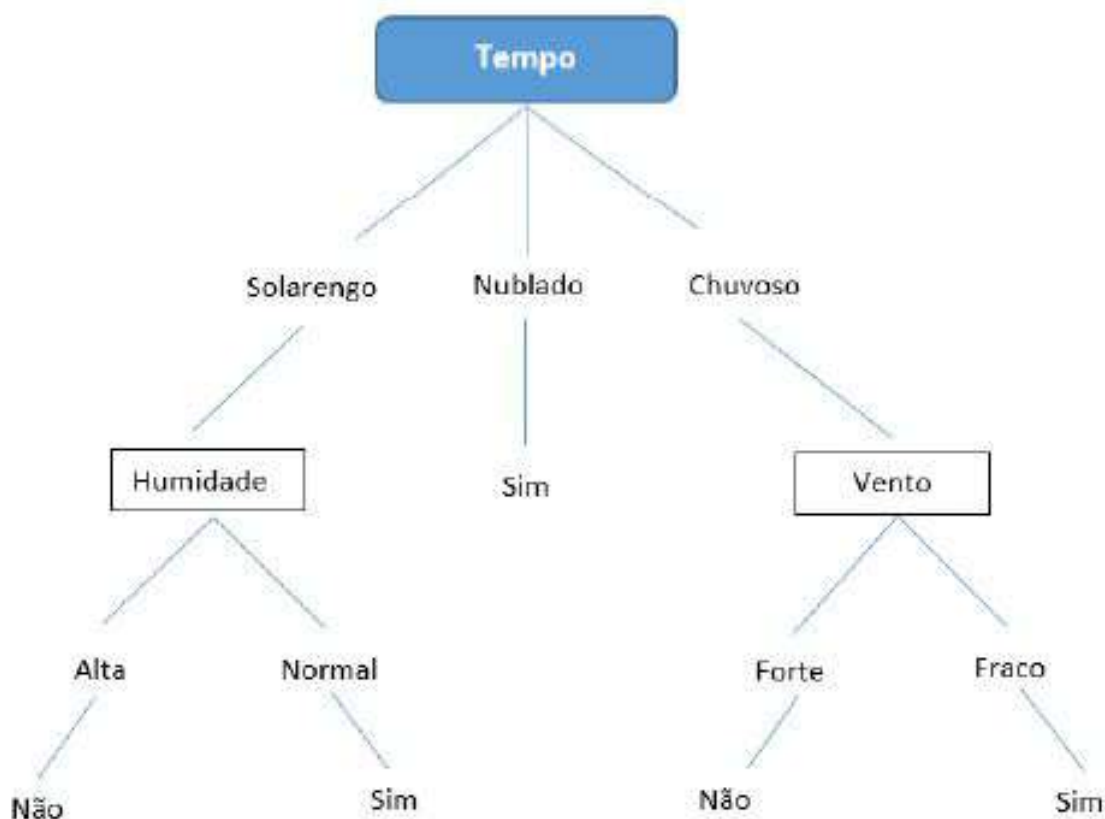
Um algoritmo de árvore de decisão consiste em nós em forma de uma árvore, esses nós, a partir da raiz, avaliam uma regra predefinida, uma condição ou operação matemática. As folhas da árvore representam o valor previsto. Assim, uma das características mais vantajosas do algoritmo de árvore de decisão é a sua transparência, o que significa que a árvore resultante pode ser lida e analisada facilmente por um ser humano (RAMEZANKHANI et al., 2014).

O objetivo é criar um modelo que possa prever o valor de uma variável de destino, aprendendo as regras de decisão simples inferidas a partir dos dados apresentados. Os algoritmos de AD são criados através de duas etapas:

- **Entropia:** É o grau ou quantidade de incerteza na aleatoriedade dos elementos, em outras palavras, é uma medida de impureza;
- **Ganho de Informação:** Mede a mudança relativa na entropia em relação ao atributo independente e busca estimar as informações contidas em cada atributo (KINGSFORD; SALZBERG, 2008).

A figura 11 ilustra um exemplo simples de uma árvore de decisão.

Figura 11 – Árvore de decisão



Fonte: (MITCHELL; LEARNING, 1997).

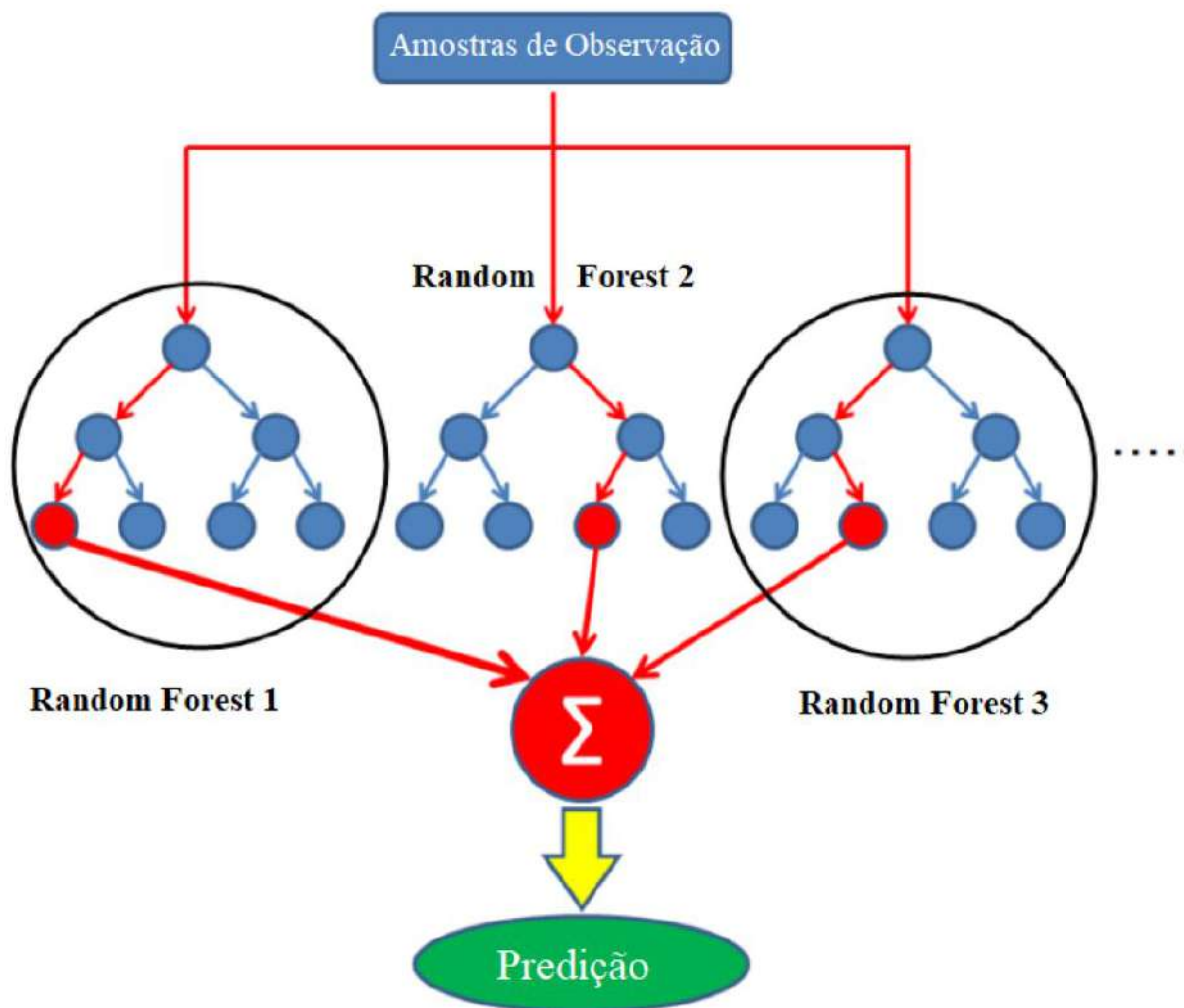
#### 2.2.1.2 Floresta aleatória

Um algoritmo do tipo Floresta Aleatória ou em inglês Radom Forest, é uma evolução do algoritmo de árvore de decisão. Engloba a construção de várias árvores de decisão e, em seguida, usa a combinação de sua saída para melhorar a capacidade de generalização do algoritmo.

Uma vez que as árvores tenham sido treinadas, elas recebem um peso dependendo do seu desempenho de previsão, e a classe resultante é calculada combinando todos os resultados multiplicado pelo seu peso. A classe com maior probabilidade ponderada é decidida como o valor de saída para o algoritmo. Este método fornece uma maior complexidade e permite uma maior precisão de previsão (GÉRON, 2019).

Ele pode ser usado para resolver problemas de regressão e classificação. Em problemas de regressão, a variável dependente é contínua. Em problemas de classificação, a variável dependente é categórica. A figura 12 apresenta um modelo de floresta aleatória.

Figura 12 – Floresta Aleatória



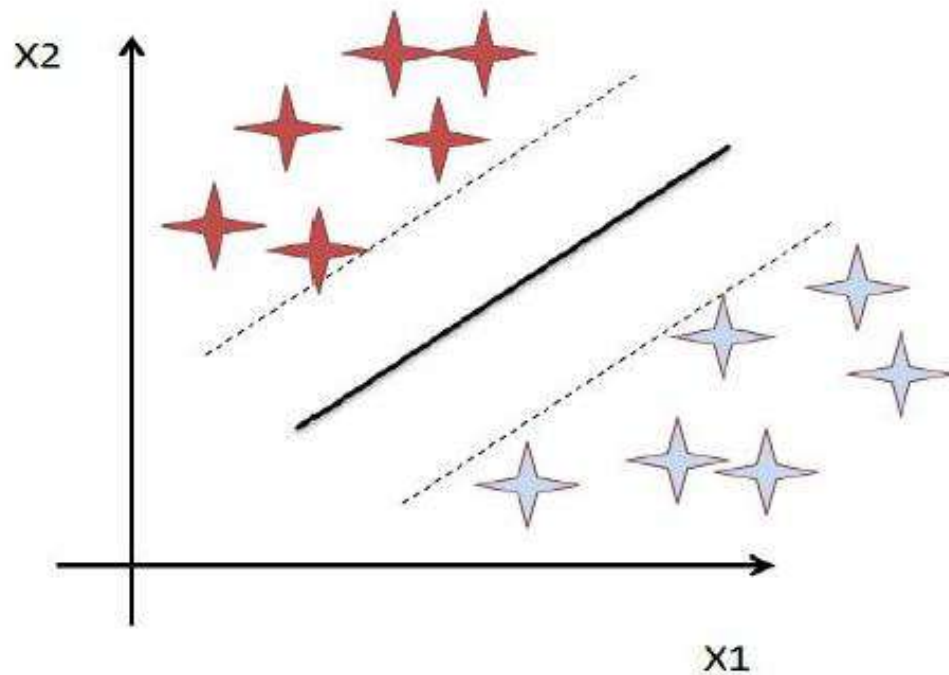
Fonte: (JAMES et al., 2013).

### 2.2.1.3 Máquina de vetores de suporte

Máquinas de Vetores de Suporte (MVS) ou *Support Vector Machine* (SVM) do inglês, é uma técnica inicialmente idealizada por Cortes e Vapnik (1995) que segue a ideia de que os dados podem ser separados entre as classes por meio de uma reta, plano ou hiperplano, de acordo com o seu número de dimensões.

O SVM classifica um vetor de entrada em classes de saída conhecidas. Começa com vários pontos de dados de duas classes e obtém o hiperplano ideal que maximiza a separação das duas classes. Para dados não linearmente separáveis, utiliza o método Kernel para transformar o espaço de entrada num espaço de recursos de alta dimensão, onde um hiperplano ideal linearmente separável pode ser construído conforme figura 13. Exemplos de funções Kernel incluem a função linear, função polinomial, função de base radial e função sigmoid (CHANG; LIN, 2001).

Figura 13 – Representação gráfica de SVM



Fonte: (MITCHELL; LEARNING, 1997).

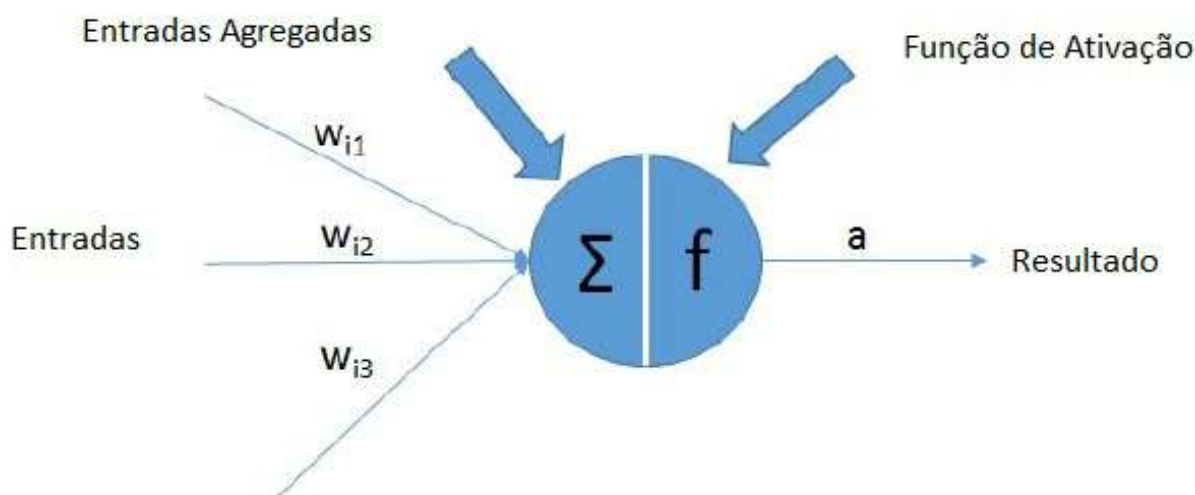
#### 2.2.1.4 Redes neurais artificiais

Redes Neurais Artificiais (RNA), ou *Artificial Neural Networks* (ANN) do inglês, são algoritmos que surgiram inspirados pelo modelo biológico do cérebro humano. De forma análoga, ANNs são compostas por um conjunto de unidades, ou neurônios, interconectados entre si por arestas, chamadas de sinapses. As sinapses possuem pesos, que são ajustados de forma a minimizar o erro resultante na saída da rede. É neste ajuste que ocorre o aprendizado.

As ANNs mais simples são compostas por um único neurônio, conhecido como Perceptron. Conforme ilustrado na Figura 14, o Perceptron recebe como entrada um vetor com os valores dos atributos  $x = [x_1; x_2; \dots; x_n]$  que se ligam ao neurônio pelo conjunto de sinapses com os respectivos pesos  $w = [w_1; w_2; \dots; w_n]$ . Por sua vez, o neurônio computa a saída de acordo com a Equação 2-2 [53]. Para um problema de classificação binária, por exemplo, a saída do Perceptron representa uma das duas classes.



Figura 14 – Modelo de funcionamento das redes neuronais



Fonte: (BERRY; LINOFF, 2004).

#### 2.2.1.5 Regressão logística

A regressão logística é o segundo mais importante algoritmo de aprendizado de máquina, sendo muito semelhante ao algoritmo de regressão linear. A maior diferença entre eles está na maneira em que geralmente são utilizados, enquanto o algoritmo de regressão linear é usado para a predição/previsão de valores os algoritmos de regressão logística são mais usados para a classificação de tarefas.

Assim, [Alghamdi et al. \(2017\)](#) descreve que a regressão logística é um classificador estatístico linear que fornece a probabilidade para cada classe de saída como uma equação dos valores de entrada. É adequado para estudar situações em que existe um conjunto de variáveis explicativas que se correlacionam com uma variável de resposta. É utilizada quando a variável dependente é categórica, normalmente assumindo valores binários (0 ou 1). Ao ser utilizada em problemas de classificação, ela determina a probabilidade de ocorrência de um fato específico com base nos valores das variáveis de entrada.

Para que esse algoritmo possa ser utilizado, a variável resposta do conjunto de treinamento é modelada por meio da distribuição binomial, que tem como parâmetro,  $\mathbf{p}$ , a probabilidade de ocorrência de uma classe específica. Vale destacar que a probabilidade estimada de determinado evento deve estar limitada ao intervalo  $[0,1]$  e apresentar relação direta com os preditores de observação da base de dados analisada ([KUHN; JOHNSON, 2013](#)).

Desta forma, para qualquer variável de entrada  $\mathbf{X}$ , a função logística garantirá que a saída seja um valor compreendido entre 0 e 1. A [Figura 15](#) detalha essa explicação:



Figura 15 – Regressão Logística

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

Fonte: (JAMES et al., 2013).

## 3 TRABALHOS RELACIONADOS

Segundo Pal (2012) é possível fazer previsões de evasão antes mesmo dos estudantes iniciarem os seus cursos, ou seja, o estudante está matriculado, mas ainda não iniciou suas aulas. O autor testa quatro algoritmos de classificação usando dados socioeconômicos e pré-universidade, como desempenho dos alunos no ensino médio. A precisão dos modelos varia de 67.7% a 85.7%.

Na pesquisa realizada por Karamouzis et al. (2008), os autores focaram em prever a taxa de aprovação dos estudantes nos dois primeiros anos dos cursos de graduação. Para isso utilizaram uma rede neural *Multilayer Perceptron* (MLP) com três camadas. Os autores misturaram variáveis demográficas e acadêmicas criando um conjunto de dados (DataSet) com as seguintes variáveis: etnia, sexo, intenção de se matricular na faculdade, idade, escola de origem no ensino médio, o estudante necessita de atendimentos especiais/deficiência?, o estudante necessita de apoio/reforço acadêmico?, CEP, endereço, idade com que terminou o ensino médio, tempo de dedicação para a faculdade e você é um estudante bem sucedido?. O estudo mostra que tiveram uma taxa média de acerto de 72%.

Jadrić et al. (2010), fizeram uma pesquisa com estudantes no início da graduação do curso de Economia. Foram considerados apenas os estudantes que estavam nos primeiros dois anos. O objetivo do trabalho foi testar e comparar os seguintes modelos de classificação: Árvore de Decisão, Regressão Logística e Redes Neurais. Foi utilizada a metodologia SEMMA (*Sampling, Exploring, Modifying, Modelling and Assessment*). Para a construção do dataset os autores optaram por utilizar as variáveis relacionadas com a inscrição do candidato e atributos referentes ao processo de estudos. Os resultados mostraram que o modelo de rede neural evidenciou um melhor desempenho em relação aos outros, indicando que 36% dos estudantes poderão evadir. Além disso, a pesquisa apontou e separou as causas da evasão e delineou o perfil típico do estudante propenso a desistir da faculdade.

Em seu estudo, Manhães, Cruz e Zimbrão (2014), exploraram apenas atributos extraídos de registros acadêmicos dos alunos. Foram usados cinco algoritmos de classificação e dados de seis cursos da Universidade Federal do Rio de Janeiro. Essa abordagem os levou a obter uma precisão de pelo menos 87.0% para cada curso e uma taxa de verdadeiros positivos que varia de 66.08%.

O estudo feito por Mustafa et al. (2012) teve como objetivo a previsão de evasão discente através de um modelo dinâmico. Os modelos utilizados foram: Árvores de Classificação e Regressão (CART) e (CHAIDA). Para a construção do dataset, foram utilizadas

as seguintes variáveis: sexo, situação financeira, período do curso que evadiu, se possui alguma deficiência. Os resultados mostraram que a utilização das Árvores de Classificação, apenas com os dados da inscrição dos candidatos não são ideias para se identificar a evasão. As taxas de precisão das Árvores de Classificação não foram significativas, sendo 38,1% com a Árvore CHAID e 28,57% com a Árvore CART. Os autores sugerem acrescentar outros fatores como: idade, etnia, situação de trabalho, ambiente de estudo e tipo de educação para melhorar o desempenho.

[Oladokun et al. \(2008\)](#), aplicaram a rede neural MLP com duas camadas escondidas e cinco neurônios por camada. Para a construção do dataset, os autores utilizaram informações de cinco gerações de graduandos. As variáveis que compuseram o vetor de entrada da rede foram: média final do ensino médio, resultados nas disciplinas de matemática, linguagem, física e química, notas em outras ciências, idade do aluno no momento da inscrição, tempo decorrido entre formação no ensino médio e admissão da universidade, escolaridade dos pais, escola de origem do ensino médio, tipo de ensino médio, local da universidade, local da residência e sexo. Os resultados indicam que a rede neural utilizada atingiu uma precisão global média de 74% de acertos.

Motivados pelo auto índice de evasão, os autores [Dekker et al. \(2009\)](#) elaboraram um estudo experimental na busca de um método de predição da evasão entre os calouros do curso de engenharia. Os autores utilizaram os seguintes modelos baseados nos classificadores WEKA: Árvores de Decisão e um classificador Bayesiano. Para a construção do dataset foram consideradas variáveis relacionadas ao desempenho no ensino médio e dados universitários. Os resultados mostraram que um dos modelos de Árvore de Decisão aplicados atingiu uma precisão de 68% quando analisou somente os dados pré-universitários. Ao verificar o conjunto completo dos dados o método obteve precisão entre 75% e 80% na identificação da evasão.

Segundo [Fu et al. \(2012\)](#), em sua pesquisa foram aplicados questionários com o intuito de compor um dataset com características da personalidade dos estudantes. O objetivo do estudo foi aplicar a regressão de vetores de suporte para construir um modelo de previsão de desempenho. Os dados foram coletados por meio de questionários. O modelo proposto obteve uma precisão próxima de 80% na previsão de desempenho dos estudantes.

[Márquez-Vera et al. \(2013\)](#), explorou a evasão no ensino médio em uma cidade mexicana. Em sua pesquisa utilizou algoritmos de classificação populares e propõe um algoritmo genético que utiliza cost-sensitive learning e técnicas de balanceamento de classes. Utilizam a Accuracy, verdadeiros positivos e verdadeiros negativos para fazer a avaliação e chegam a atingir valores de 93.4%, 94.0% e 88.3%, respectivamente.

O trabalho desenvolvido por [Lykourentzou et al. \(2009\)](#) utilizou as seguintes técnicas de ML: rede neural feed-forward, máquina de vetor de suporte (SVM) e rede neural ARTMAP-Fuzzy simplificada. Para a construção do dataset foram utilizadas variáveis

demográficas e informações detalhadas das atividades dos estudantes e de seu progresso. Na análise dos resultados, a técnica que obteve melhor desempenho quanto à rapidez, sensibilidade e precisão na previsão de estudantes propensos à evasão foi o esquema de decisão. Entre a aplicação dos métodos aplicados individualmente e os esquemas de combinações a taxa global de precisão alcançada foi de 40% a 50% quando utilizaram somente as características demográficas e entre 75% e 85% quando utilizaram todos os atributos dos estudantes.

Utilizando dados de mais de 11 mil estudantes de três cursos de uma instituição de ensino superior de Brasília, [Balaniuk et al. \(2011\)](#), propõe o uso de três algoritmos de classificação para classificar os alunos em "EVASÃO" e "GRADUADO". Para treinar os modelos, foram utilizados como entrada tanto atributos com informações socioeconômicas quanto acadêmicas dos alunos. Por fim, concluiu-se que é possível identificar estudantes com alto risco de evasão com Accuracy de até 80.6%.

[Coelho et al. \(2016\)](#) utiliza a metodologia CRISP-DM juntamente com técnicas de *Educational Data Mining* (EDM) na identificação descritiva e preditiva dos padrões de interação de um ambiente virtual de aprendizagem (AVA). Os resultados demonstraram um bom potencial de predição com cerca de 87% de assertividade a partir de Árvores de Decisão, do desempenho dos alunos quando analisadas as variáveis associadas à interação com os módulos do Moodle utilizados pela Enap, em seu AVA.

A pesquisa desenvolvida por [Hoffmann et al. \(2016\)](#) utiliza Knowledge Discovery in Databases (KDD) e a metodologia CRISP-DM no curso de Zootecnia, partindo de informações sobre os alunos e seus desempenhos acadêmicos e aplicando técnicas de mineração de dados para a previsão. Os resultados demonstram que a abordagem proposta é factível e eficiente. Os experimentos, em dados reais, alcançaram uma acurácia de 98% na previsão, e mais de 70% de sucesso na previsão de alunos que abandonaram o curso.

[Júnior \(2018\)](#) procurou desde a coleta dos dados até a aplicação dos modelos nos experimentos e avaliação dos resultados, todo o trabalho foi conduzido seguindo uma mescla das etapas dos processos de mineração de dados: KDD e CRISP-DM. Os resultados obtidos mostraram uma adequação destes processos ao problema de mineração de dados educacional relacionado à tarefa de classificação de desempenho acadêmico dos alunos submetidos ao ENEM. A técnica de regressão logística produziu índices de propensão de sucesso dos alunos e atingiu resultados superiores a 0,84 e 0,51 para as métricas AUC-ROC e KS2-MAX, respectivamente. Concluiu-se que a abordagem apresentada teve um ótimo resultado na modelagem e na validação de políticas públicas.

Segundo [Botchkarev \(2018b\)](#) o Azure Machine Learning Studio(MLS) tem o potencial de acelerar os experimentos de aprendizado de máquina, oferecendo um recurso integrado conveniente e poderoso ambiente de desenvolvimento. Sua pesquisa relata os resultados de um esforço para criar um módulo *Enhanced Evaluate Model* (EEM) que

facilita e acelera o desenvolvimento e a avaliação de experimentos do Azure. O EEM combina várias métricas de desempenho, permitindo a avaliação de vários lados dos modelos de regressão.

Segundo [Botchkarev \(2018a\)](#) a capacidade de previsão custos na área hospitalar é fundamental para o gerenciamento financeiro e o planejamento orçamentário da assistência médica eficiente. Vários algoritmos de ML de regressão são eficazes para previsões de custos de serviços de saúde. Em seu trabalho criou-se um experimento no MLS para avaliação rápida de vários tipos de modelos de regressão. O modelo foi publicado e pode ser usado por pesquisadores e profissionais para reproduzir os resultados deste estudo, realizar experimentos com seus próprios conjuntos de dados ou adicionar mais modelos de regressão à estrutura.

Como visto, os trabalhos relacionando mostram resultados positivos com os algoritmos de classificação. Alguns trabalhos, como [Coelho et al. \(2016\)](#), obtiveram uma precisão de 87% com o algoritmo de classificação Árvore de Decisão. [Márquez-Vera et al. \(2013\)](#) utilizou algoritmos de classificação e chegou a atingir Accuracy de 94%. Portanto decidiu-se utilizar algoritmos de classificação para verificar se esses algoritmos ainda continuam precisos.

## 4 CONTEXTUALIZAÇÃO AMBIENTE

A contextualização do ambiente apresenta dados sobre a tecnologia da informação na educação e na instituição objeto de estudo, visto que a análise surgiu baseando-se na necessidade de uma instituição de ensino para se entender os possíveis motivos da evasão dos alunos.

### 4.1 Tecnologia da Informação na Educação

Segundo [Veen e Vrakking \(2009\)](#), uma das finalidades da educação ao longo dos tempos foi a de preparar os indivíduos para exercerem diversos papéis na sociedade. No entanto, os autores são categóricos em afirmar que esta primazia tem vindo a decair ao longo das últimas décadas. O que pode estar ligado ao advento das Tecnologias da Informação e Comunicação (TIC), e à necessidade da aprendizagem ao longo da vida em ambientes informais até então nunca pensados. De fato, tal como sugerido na seção anterior, uma vez que as tecnologias estão permanentemente em mudança, a aprendizagem ao longo da vida é consequência natural do momento social e tecnológico em que se vive, a ponto de poder chamar a sociedade de "sociedade de aprendizagem" ([LENCASTRE, 2009](#)).

Para [Coutinho e Lisbôa \(2011\)](#), a utilização das TIC garante a difusão de novas estratégias de veiculação da informação, bem como novos modelos de comunicação, abrindo um leque de possibilidades de mudanças comportamentais e atitudinais ao ser humano em relação aos processos educacionais.

Para [Veen e Vrakking \(2009\)](#) a solução passa pelo uso das tecnologias como parceiras do processo de construção do saber e pela formação de professores. As TIC permitem aproximar pessoas de diferentes origens socioeconômicas, propiciando o aparecimento de espaços para troca de informações e partilha de conhecimentos. Isto se torna um desafio para a escola, pois ensinar em plena Era Digital contribui para criar “oportunidades nunca antes vistas para tornar o ensino uma profissão apaixonante e motivadora, que faça diferença para a sociedade futura. Tais oportunidades relacionam-se a novos papéis, novos conteúdos e métodos de ensino e aprendizagem.” ([VEEN; VRAKING, 2009](#)).

### 4.2 Caracterização da Instituição Objeto da Pesquisa

A pesquisa foi realizada em uma instituição de ensino de Belo Horizonte que é referência em educação profissional na área de comércio de bens, serviços e turismo. A

instituição oferece tanto em modalidade presencial, quanto a distância, cursos livres e técnicos, além de graduação e Master Business Administration - MBA, esse último listado, está no ranking dos melhores cursos do jornal "O Estado de São Paulo". O seu portfólio incluem opções nos segmentos de saúde, segurança, meio ambiente, gastronomia, hospedagem, turismo, produção de alimentos, gestão, comércio, idioma, artes, beleza, design, moda, saúde, asseio, conservação e zeladoria, informação, comunicação e educacional.

Com 40 unidades em todo o Estado, a instituição se destaca por sua metodologia e tradição. Outros diferenciais são os ambientes de práticas pedagógicas, como os Salões de Beleza-Escola, os Restaurantes-Escola, a Pousada-Escola Tiradentes e o Hotel-Escola, primeiro na categoria da América Latina. A instituição também leva seu know-how para Minas Gerais afora por meio de suas 12 carretas-escola que reproduzem nossos ambientes pedagógicos. Além disso, promove atividades de extensão, como palestras e oficinas.

No Ensino Superior, a Faculdade se destaca pela qualidade. A graduação tecnológica em Gestão da Qualidade teve a nota 4 no ENADE, do Ministério da Educação. Esse resultado leva em conta a escala de 1 a 5. O mesmo destaque também foi dado à graduação tecnológica em Gastronomia e ao curso de Administração. Já a graduação tecnológica em Gestão da Qualidade teve o melhor desempenho entre as faculdades privadas da cidade. Em outra avaliação nacional, realizada pelo Guia do Estudante, que avalia cursos de bacharelados em todo Brasil, os cursos de Administração e Ciências Contábeis conquistaram a nota 4. Para complementar essas informações, seguem alguns dados do relatório anual de 2018.

A instituição se empenha em promover a inclusão social. Internamente, o Setor de Educação Inclusiva investe em recursos didáticos e, quando necessário, em metodologias diferenciadas para atender alunos, independentemente de classe social, gênero, orientação sexual, etnia, idade, entre outros aspectos. Além disso, por meio da Rede de Carreiras, proporciona a divulgação gratuita de vagas de emprego e currículos de candidatos e outras ações para aquecer o mercado de trabalho. Já com o Programa de Gratuidade (PSG) oferece cursos em várias áreas. Para apresentar uma noção ainda mais ampla em números, foi apresentada uma breve análise do seu relatório anual de 2018.

Em 2018, a instituição apostou em um novo site. Mais moderno, com novas funcionalidades e navegabilidade mais rápida e eficiente. O novo portal também é responsivo (adaptável aos smartphones, tablets e demais dispositivos móveis). Também foi feita reformulação de conteúdo, com linguagem mais adequada à web, que favorece o ranking nos buscadores da internet. Mesmo com pouco tempo de medição, as métricas já mostram avanços grandes na interação com o usuário.

Ferramentas de marketing tais como: Design Thinking, Duplo Diamante, Matriz de Alinhamento, Mapa de Stakeholders e Highlights, além de pesquisas com alunos da Instituição, foram utilizadas para entender as necessidades, dores e desejos dos clientes,

traduzidos nas campanhas publicitárias dos produtos. Nessa comunicação integrada, a mesma ideia divulga todos os níveis educacionais, fortalecendo a marca e apresentando a capilaridade de seu portfólio.

Nas redes sociais, o destaque foi uma divulgação de ações educacionais realizadas pelos alunos durante os cursos, eventos e ações extensivas, com produções audiovisuais e depoimentos dos alunos. Ao todo, foram 1.271 atendimentos no ano. A tabela 4 mostra os números relacionados às visualizações de páginas dos sites.

Tabela 4 – Visualizações de páginas por site

Site	Número Visualizações
Institucional	3.754.307
Faculdade	427.038
Pousada Escola	31.504
Hotel Escola	89.848
Aprender na Instituição	52.781

Fonte: O Autor adaptado relatório anual da instituição (2018).

Também houve um crescimento nas redes sociais. A tabela 5 mostra algumas redes sociais e o crescimento que ocorreu em 2018.

Tabela 5 – Crescimento nas Redes Sociais

Rede Social	Crescimento
Facebook	6%
Youtube	26%
Twitter	3%
Linkedin	185%
Instagram	217%

Fonte: O Autor adaptado relatório anual da instituição (2018).

A figura 16 mostra alguns dos totais de contatos relacionados à imprensa, aos canais de atendimento, aos atendimentos corporativos e rede de carreira. Foram detalhados alguns desses números, para assim compreender o porte dessa instituição e sua preocupação com a evasão escolar.



Figura 16 – Relatório anual - 2018 divulgação



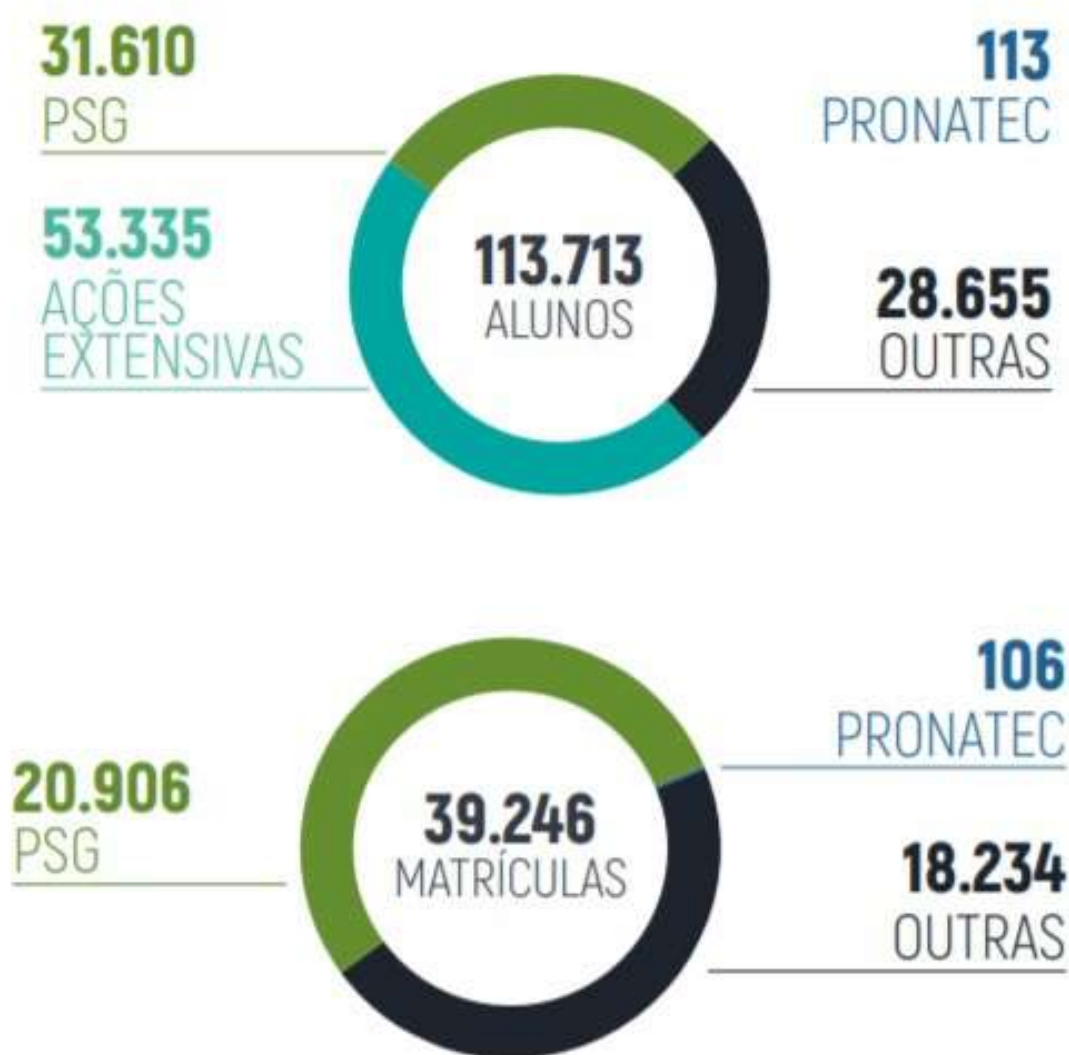
Fonte: O Autor adaptado relatório anual da instituição (2018).

No ano de 2018 a instituição ofertou em todo o Estado de Minas Gerais um rol de cursos que foram da educação básica à pós-graduação, levando conhecimento e capacitação a 169 municípios. Para se conseguir atender ao público dessas cidades a instituição conta ainda com 5 unidades regionais que estão estrategicamente distribuídas de maneira a garantir maior capilaridade das ofertas. Com essa estratégia foi possível atender a 60.378 alunos, sendo que 39.246 alunos foram concluintes e 21.132 alunos ainda em curso, da educação profissional. Além disso, têm-se 53.335 alunos de ações extensivas, totalizando 113.713 alunos atendidos em 2018 e uma carga horaria total de 6.521.063.

Outro ponto a considerar é o número de alunos atendidos pelo PSG que totalizou

31.610. Em relação às matrículas concluídas o PSG representa 53,27% do total de alunos. Já em relação às matrículas em andamento esse valor é de 50,65%. Esses números reafirmam o comprometimento da instituição em expandir, interiorizar e democratizar a oferta de cursos de educação profissional e tecnológica para a população em geral. A figura 17 mostra em mais detalhes essas informações.

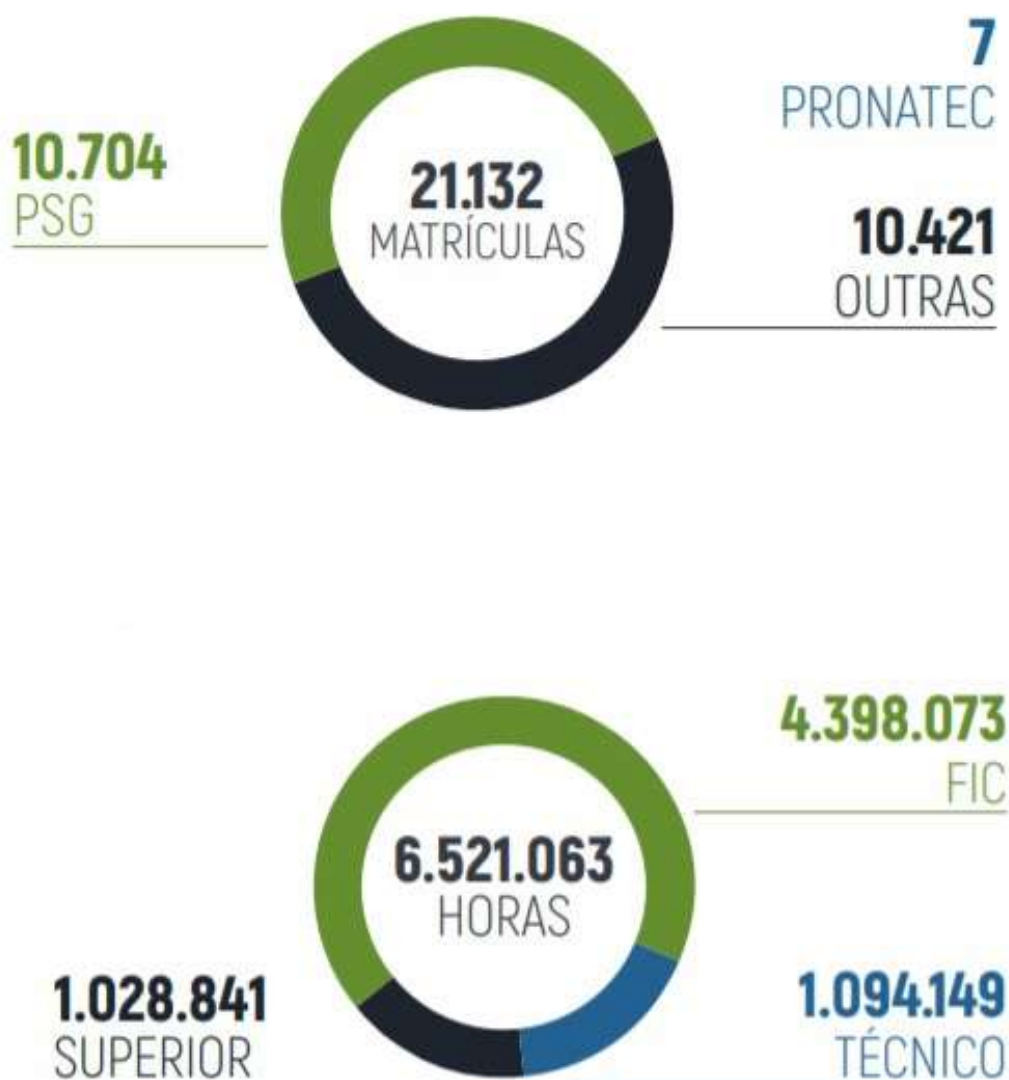
Figura 17 – Total de matrículas por forma de pagamento



Fonte: O Autor adaptado relatório anual da instituição (2018).

A figura 18 a seguir mostra que 21 mil alunos com o status 'em processo' continuaram estudando em 2019 e que 50,65% desses alunos fazem parte do programa PSG e que mais de 6 milhões de horas foram ministradas nas modalidades Formação Inicial Continuada(FIC), técnico e superior em 2018.

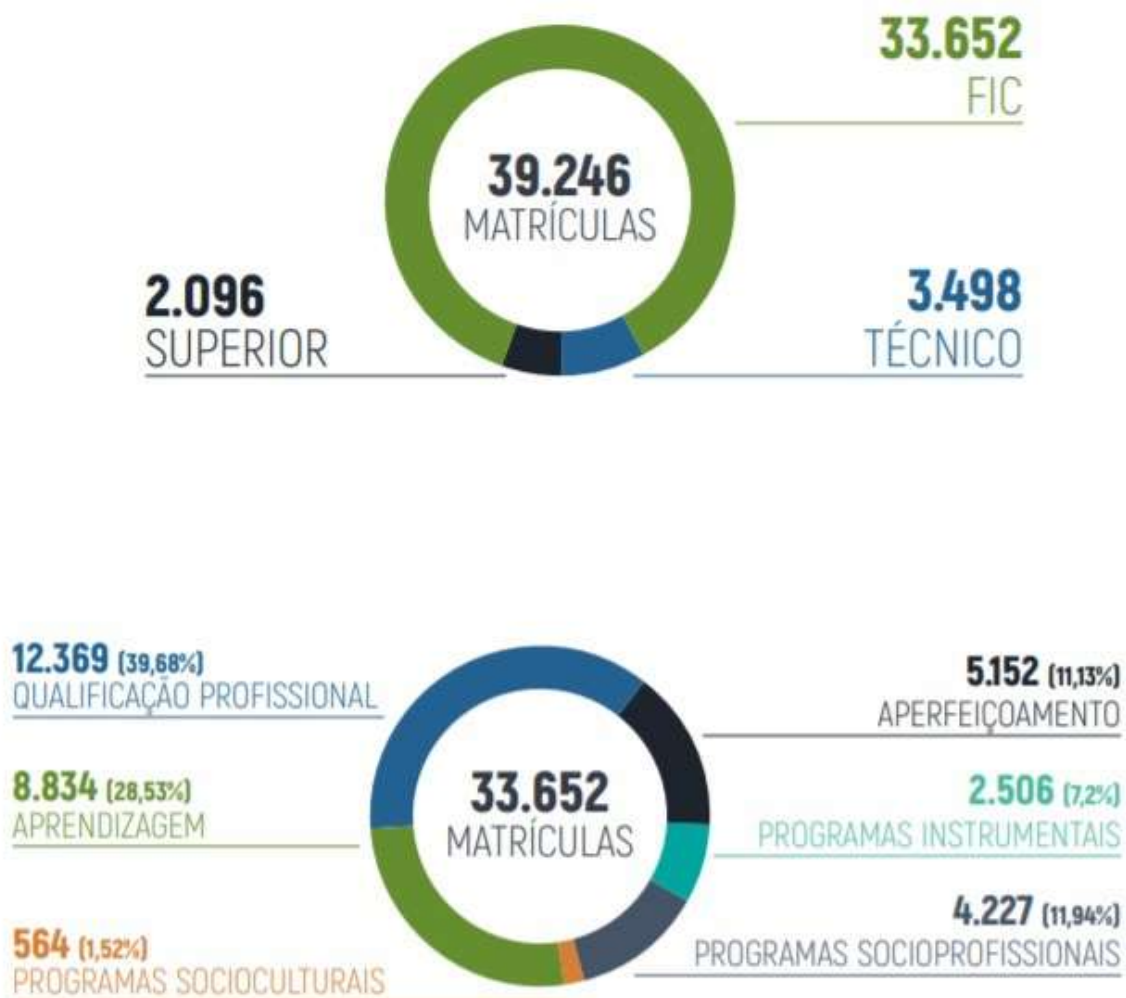
Figura 18 – Matrículas em processo e carga horária concluída



Fonte: O Autor adaptado relatório anual da instituição (2018).

Analisando as matrículas concluídas por modalidade verifica-se que **85,75%** foi **FIC**, **8,91%** **técnico** e apenas **5,34%** **superior**. Quando analisada a modalidade FIC por tipo de curso, entende-se que a procura maior é por qualificação profissional que representa **36,76%** da matrículas da modalidade FIC. Veja na figura 19 a representação desses números.

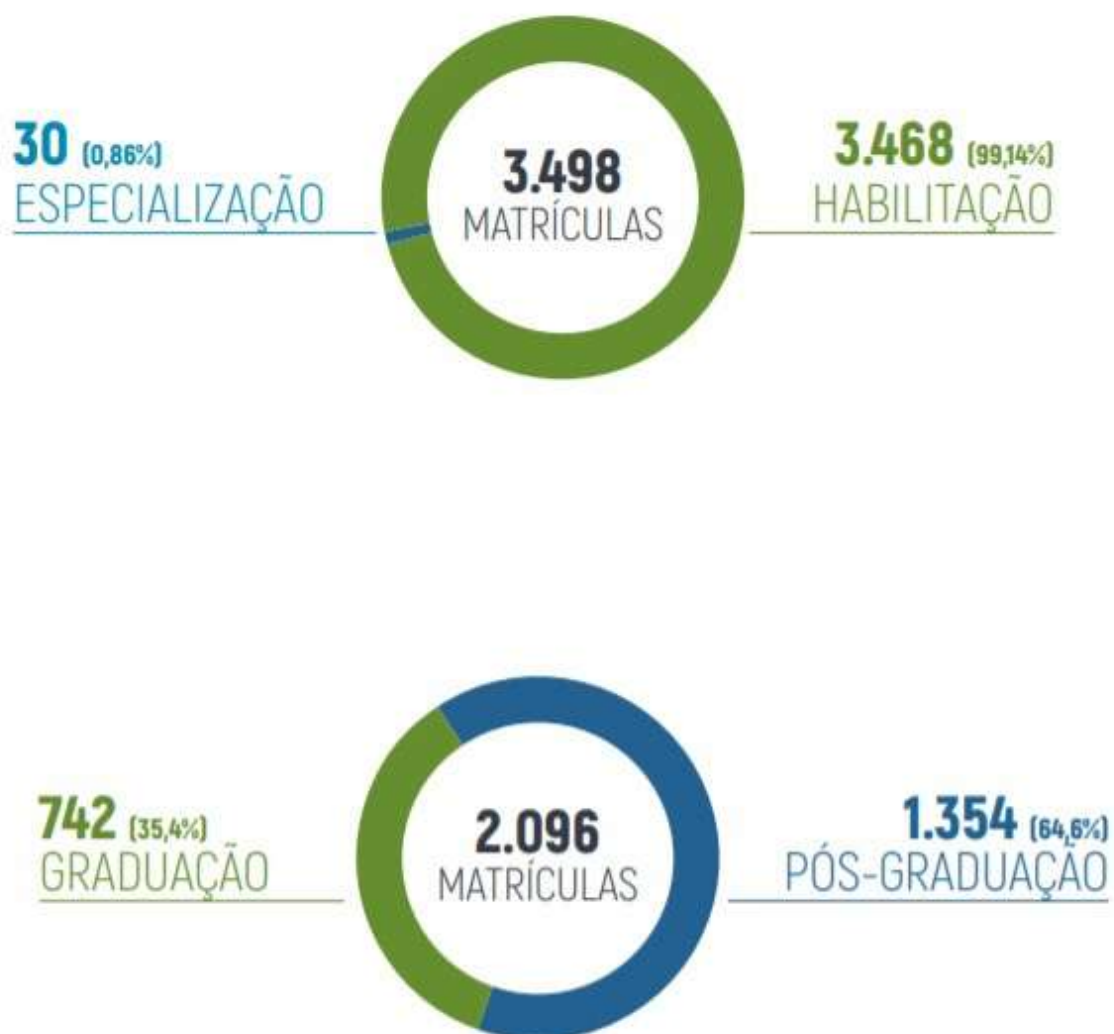
Figura 19 – Matrículas concluídas por modalidade e tipo de curso



Fonte: O Autor adaptado relatório anual da instituição (2018).

Fazendo uma análise da modalidade de cursos técnico, observa-se que a procura foi próxima a **100%** por essa habilitação. Já na modalidade de ensino superior, os cursos de pós-graduação se destacaram com mais de **60%** das matrículas, conforme confirma a figura 20 a seguir.

Figura 20 – Matrículas concluídas por tipo curso

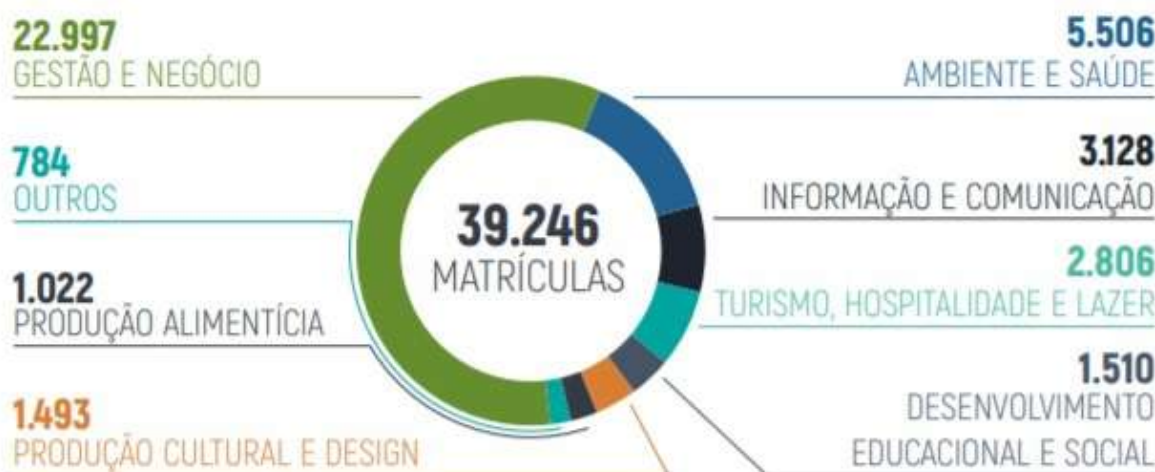


Fonte: O Autor adaptado relatório anual da instituição (2018).

A instituição também tem seus cursos divididos por eixos tecnológicos. A figura 21 a seguir, mostra que o eixo de 'Gestão e Negócio' representou **58,60%** de todas as matrículas concluídas em 2018.



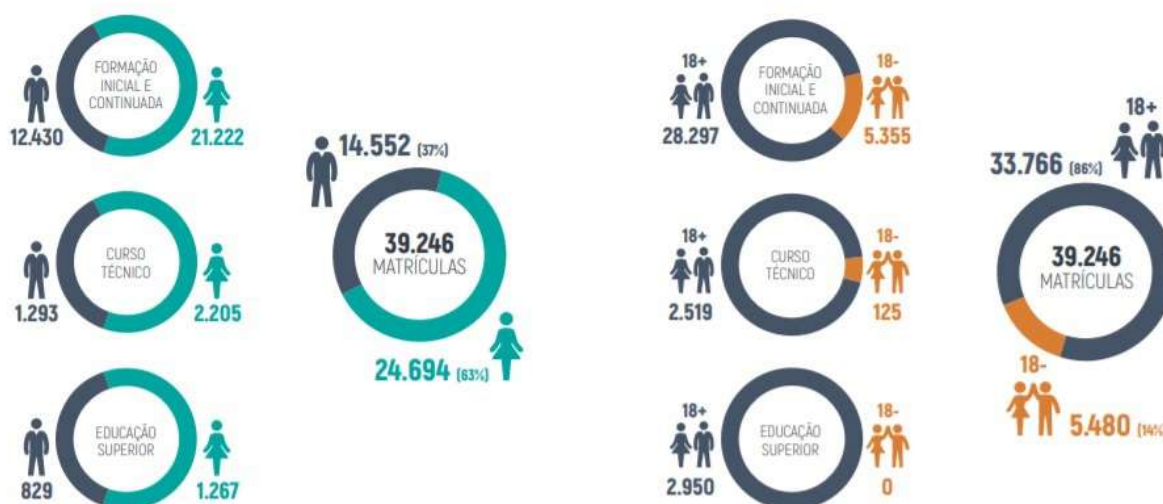
Figura 21 – Matrículas concluídas por êxito tecnológico



Fonte: O Autor adaptado relatório anual da instituição (2018).

A próxima figura 22 mostra as matrículas concluídas por modalidade de educação, sexos e maioridade.

Figura 22 – Matrículas por sexo e maioridade



Fonte: O Autor adaptado relatório anual da instituição (2018).

Com números nessa proporção é preciso uma atenção maior para que a taxa de evasão não cresça. Antes de falar dos números da figura 23 é importante explicar alguns dos status que os alunos podem ter na instituição.

Por prestar contas a um Departamento Nacional (DN), órgão executivo da administração nacional responsável pela coordenação das políticas e diretrizes nacionais, a instituição precisa detalhar a situação de cada aluno em cada curso, já que existem vários programas do governo, muitas matrículas e carga horária envolvida. A tabela 6 a seguir, mostra alguns desses status. Como já mencionado, para esse estudo consideram-se os status: Cancelado com Desistente, Trancado, Abandono, Desistente, Transferido, Cancelado e Evadido como sendo **EVADIDO**. Para esse estudo se o aluno fez a matrícula e não concluiu o curso é considerado evasão.

Tabela 6 – Status do aluno no curso

Status	Descrição
Em Processo	O aluno está matriculado e segue em curso.
Trancado	Em algum momento do curso o aluno interrompeu os estudos.
Transferido curso	Nesse caso o aluno pede para mudar de curso mas permanece na instituição.
Transferido	O aluno muda de instituição.
Cancelado	O aluno faz a matrícula mais antes de começar o curso ele cancela.
Aprovado	O aluno atingiu a nota e frequência exigida para aprovação.
Pendente	O aluno termina o curso mas está pendente de alguma disciplina ou documentação.
Abandono	O aluno começou o curso e em algum momento não voltou mais.
Desistente	O curso inicia mas o aluno não frequenta a aula e não pede para cancelar.
Remanejado	O aluno veio de outra turma.
Reprovado	O aluno não conseguiu nota ou frequência suficiente para ser aprovado.
Evadido	Após o curso ser finalizado, o aluno que saiu do curso o status é alterado para evadido.
Cancelado desistente	O curso inicia mais o aluno não frequenta a aula e pede para cancelar.

Fonte: O Autor adaptado da base de dados da instituição (2019).

A partir do entendimento de cada um dos status, já é possível analisar a taxa de evasão tendo como base o relatório geral de 2017 e 2018 da instituição. A figura 23 mostra que a taxa de evasão em 2017 foi de **17,57%**, a desistência de **6,64%** além disso, 5 foram cancelados e 32 alunos trancaram o curso. Com uma soma desses status e recálculo os valores chegam a uma taxa de evasão de proximamente **23,15%**.

Na figura 24 é possível concluir que a taxa de evasão em 2018 foi de **18,51%**, a desistência de **5,71%** e de 33 alunos que trancaram. A soma desses status e a partir do recálculo, os valores chegam a uma taxa de evasão de proximamente **23,25%**.

Figura 23 – Aproveitamento por status em 2017

MATRÍCULA CONCLUÍDA	DESISTÊNCIA	TAXA DE DESISTÊNCIA	CANCELAMENTO	TRANSFERÊNCIA	TRANCAMENTO	SAÍDA INTERMEDIÁRIA
<b>36.034</b>	<b>2.392</b>	<b>6,64%</b>	<b>5</b>	<b>0</b>	<b>32</b>	<b>0</b>
MATRÍCULA EFETIVA	EVASÃO	TAXA DE EVASÃO	REPROVAÇÃO	TAXA DE REPROVAÇÃO	APROVAÇÃO	TAXA DE APROVAÇÃO
<b>33.642</b>	<b>5.912</b>	<b>17,57%</b>	<b>1.878</b>	<b>5,58%</b>	<b>25.852</b>	<b>76,84%</b>

Fonte: O Autor adaptado relatório anual da instituição (2017).

No relatório anual, a instituição analisa o total de alunos que concluíram cursos durante o ano, tendo em vista que esse relatório é uma exigência do DN que fiscaliza a instituição. Entretanto, a partir da análise dos dados de um prisma diferente, é possível se surpreender com os índices de evasão dos últimos 10 anos. Pela análise dos dados a partir das matrículas realizadas por ano, de 2009 a 2018, a taxa de evasão aumenta bastante em alguns casos. Em 2009 a taxa de evasão chegou a **26,43%** das matrículas realizadas. Em 2018 essa taxa sobe para **40,02%** de evasão das matrículas realizadas. Entende-se que para esse estudo é de suma importância analisar a evasão a partir das matrículas efetuadas, pois desta forma, é possível analisar melhor o contexto da evasão na instituição.

Figura 24 – Aproveitamento por status em 2018

MATRÍCULA CONCLUÍDA	DESISTÊNCIA	TAXA DE DESISTÊNCIA	CANCELAMENTO	TRANSFERÊNCIA	TRANCAMENTO	SAÍDA INTERMEDIÁRIA
<b>39.246</b>	<b>2.240</b>	<b>5,71%</b>	<b>0</b>	<b>0</b>	<b>33</b>	<b>0</b>
MATRÍCULA EFETIVA	EVASÃO	TAXA DE EVASÃO	REPROVAÇÃO	TAXA DE REPROVAÇÃO	APROVAÇÃO	TAXA DE APROVAÇÃO
<b>37.006</b>	<b>6.850</b>	<b>18,51%</b>	<b>2.221</b>	<b>6%</b>	<b>27.935</b>	<b>75,49%</b>

Fonte: O Autor adaptado relatório anual da instituição (2018).

Para que a instituição continue a cumprir a sua meta, sempre pautada na entrega de uma educação profissional de qualidade e alinhada ao que o mercado necessita, é necessário entender os motivos que levam a evasão escolar.





busca por determinado curso ou turma em determinado período. Ao executar a consulta clicando no botão Pesquisar conforme a seta mostra. A consulta pode retornar uma ou varias turmas dependendo da pesquisa. Tem-se a opção de clicar nessa turma consultada, conforme a seta e obter mais informações sobre a turma pesquisada. A figura 27 ajuda na compreensão.

Figura 27 – Pesquisa curso no sistema acadêmico - SA

Cód.Turma / Nome curso:

Que inicia entre:  à  Que termina entre:  à

Situação:  Liberado Matrícula  Bloqueada Matrícula  Turma Concluída  Turma Cancelada

Ordem:  Código da turma / Nº módulo  Nome do Curso  Data de Início

Detalhes da turma:  Todos os alunos  Somente alunos ativos

[Pesquisar](#) | [Limpar filtros](#)

Turma	Local	Status	Dt. Início Matrícula	Período / Horário / Dias	Carga Horária	Valor Turma	Qtd Vagas	Qtd Alunos	Turma Fechada	Area Profissional
006.2018.0001 - BOAS PRÁTICAS NA MANIPULAÇÃO DE ALIMENTOS	CEP Belo Horizonte	Turma Concluída	2/10/2017	19/2/2018 à 23/2/2018 18:00 às 22:00 ( Seg Ter Qua Qui Sex )	20 hs	R\$ 180,00	21	19		HOSPITALIDA DE

Fonte: O Autor adaptado Sistema Acadêmico (2019).

O cadastro do aluno é bem extenso visto que a instituição precisa de algumas informações para prestar contas ao DN e as auditorias que ocorrem frequentemente. A figura 28 apresenta apenas uma parte desse cadastro.

Figura 28 – Cadastro do aluno no sistema acadêmico - SA

CPF\*:

Nome\*:  Sexo\*:

Possui Nome Social?  Sim  Não

Nome Social:

Data de Nascimento\*:  Naturalidade\*:

Nacionalidade\*:  UF\*:

E-mail:

Faixa de renda familiar\*:  Estado Civil\*:

Nível de Escolaridade\*:

Cor\*:  Portador de deficiência\*:

Trabalha atualmente?\*:  Sim  Não / Situação\*:

Fonte: O Autor adaptado Sistema Acadêmico (2019).

# 5 METODOLOGIA

Neste capítulo, foi descrita a metodologia de pesquisa utilizada nesse estudo a fim de alcançar os objetivos propostos, a caracterização da pesquisa, os procedimentos para a coleta e análise de dados, a ferramenta e metodologia de desenvolvimento.

## 5.1 Caracterização da Pesquisa

Para Gil (2010) a pesquisa pode ser definida como “o procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos”.

Ainda segundo Gil (2010), as razões que determinam a realização de uma pesquisa podem ser classificadas em dois grandes grupos denominados como razões de ordem intelectual, em que se afirma o desejo de conhecer pela própria satisfação de conhecer, e de razões de ordem prática. Também se afirma o desejo de conhecer com vistas a fazer algo de maneira mais eficiente ou eficaz. No caso desse estudo, ambas as razões são inerentes ao propósito do autor.

VERGARA (2013) qualifica uma pesquisa em relação a dois critérios básicos de classificação: quanto aos fins e quanto aos meios. Quanto aos fins esta pesquisa demonstra caráter predominantemente exploratório. Segundo Gil (2008), a pesquisa exploratória tem como principal objetivo procurar dispor maior familiaridade com o problema a fim de torná-lo explícito, tendo como objetivo central o aprimoramento de ideias e a descoberta de intuições.

Quanto à forma de abordagem, a presente pesquisa se enquadra como qualitativa, que segundo Cooper e Schindler (2016), permite que uma situação seja compreendida com profundidade, condizente com “estudos que visam percepções, motivações, sentimentos ou comportamentos, os diferentes significados que as pessoas atribuem às suas experiências”. Além de procurar responder como e por que algo acontece a partir de dados obtidos em diversas fontes, como textos, objetos, pessoas e organizações.

Também é um estudo de caso, que utilizando métodos dedutivos, ou seja, a dedução “parte do entendimento das leis e teorias que abrangem determinado fenômeno e a partir da definição de premissas e análise da relação entre elas se constrói o conhecimento” (DRESCH; LACERDA; JÚNIOR, 2015).

Quanto à natureza essa pesquisa se caracterizou como aplicada, pois conforme Kauark, Manhães e Medeiros (2010), objetivou gerar conhecimentos para aplicação prática, dirigida à solução de problemas específicos, envolvendo verdades e interesses. Assim, foi realizada com o propósito de resolver um problema concreto, neste caso a evasão

escolar.

Para atingir o objetivo geral, foi necessário recuperar as informações por meio de consultas nos bancos de dados da instituição. O mapeamento dos dados armazenados e o levantamento das informações deram origem à construção do modelo ML. A Tabela 7 apresenta uma síntese da coleta de dados, relacionando os objetivos específicos deste estudo.

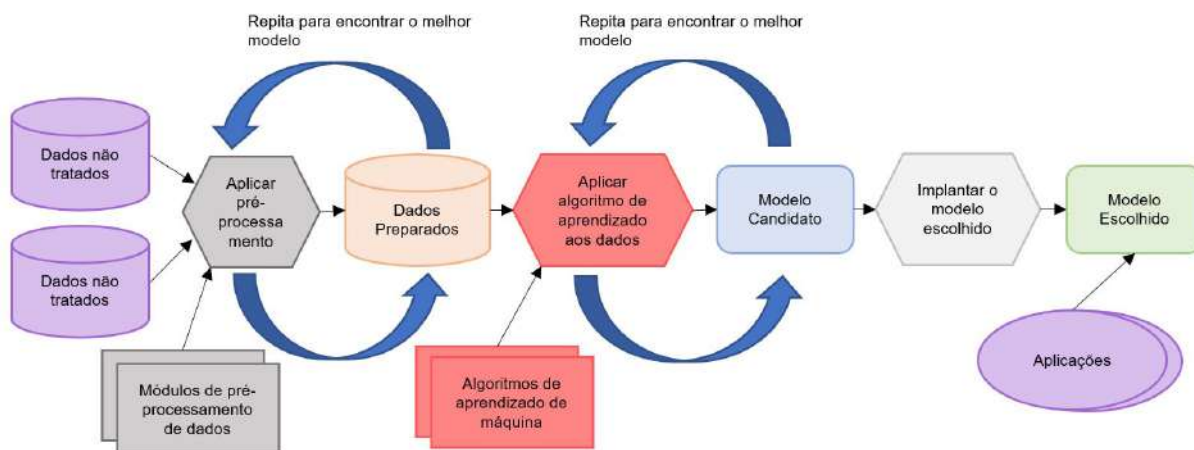
Tabela 7 – Síntese dos objetivos relacionados com coleta de dados

Objetivos específicos	Instrumento de coleta de dados
Verificar na literatura quais as variáveis que faram parte do modelo.	Referencial Teórico.
Identificar quais dados e informações, conforme levantamento bibliográfico, estão presentes nos bancos de dados do sistema acadêmico.	Levantamento das informações presentes nos bancos de dados.
Preparar os conjuntos de dados para treinamento e teste do modelo.	SQL, DataTools, Machine Learning Studio

Fonte: Elaborado pelo Autor (2019).

Como o objetivo desse trabalho é criar um modelo preditivo que seja capaz de detectar estudantes em risco de evasão através de técnicas de ML, o fluxo do algoritmo segue conforme a figura 29 abaixo:

Figura 29 – Fluxo do modelo ML



Fonte: O Autor adaptado de (CHAPPELL, 2015).

Parar entender melhor a figura 29, cada uma das sete etapas será explicada a seguir:

1. **Coleta de dados:** trata-se de um momento crucial para o resultado final, pois a quantidade e a qualidade das informações determinam o quão preditivo o modelo de ML vai ser.

2. **Preparação dos dados:** verificar se as informações coletadas estão bem distribuídas ou se são tendenciosas. Nessa etapa, os dados coletados são separados em duas amostras — uma a ser utilizada no treinamento (etapa 4) e outra para a avaliação de performance do modelo (etapa 5).
3. **Escolha do modelo:** há uma infinidade de modelos de ML disponíveis, cada um destes voltado ao cumprimento de uma determinada função. Por isso, a escolha do modelo mais adequado foi feita de acordo com o objetivo proposto.
4. **Treinamento:** essa etapa é fundamental não apenas para preparar a máquina, mas para aprimorar constantemente suas habilidades de previsão. Dessa forma, a máquina efetivamente aprende com seus erros e torna-se cada vez mais aperfeiçoada.
5. **Avaliação:** testa o modelo com as informações não utilizadas no treinamento. Isso permite verificar se a máquina realmente foi capaz de aprender, e não apenas de memorizar respostas anteriores.
6. **Aprimoramento dos parâmetros:** visando sempre melhorar a qualidade e a eficiência do modelo de ML que está sendo utilizado, essa etapa identifica valores que afetam diretamente a acurácia do modelo e o tempo de treinamento necessário.
7. **Predição:** é quando a máquina dotada de ML pode efetivamente ser usada para responder as perguntas para as quais foi treinada.

## 5.2 CRISP-DM

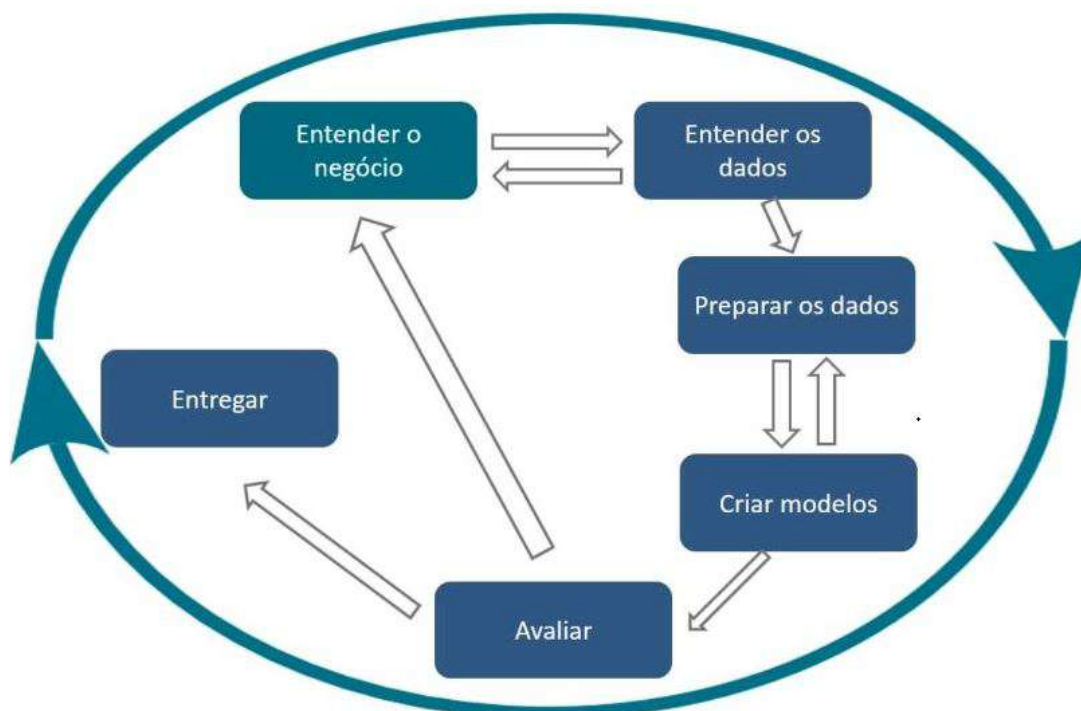
O modelo de processo CRISP-DM é uma metodologia concebida visando auxiliar a estruturação de projetos de mineração de dados ([CHAPMAN et al., 2000](#)). A metodologia tem por base um processo cíclico e interativo. Cada fase do CRISP-DM tem processos específicos que serão descritos a seguir:

1. **Compreensão do Negócio:** esta fase inicial esforçou-se em compreender os objetivos, requisitos e as necessidades do projeto a partir do ponto de vista do negócio, transformando esse conhecimento na definição de um problema a ser solucionado com os conhecimentos adquiridos com a mineração de dados, juntamente com um plano de trabalho projetado para atingir estes objetivos.
2. **Compreensão dos Dados:** nesta etapa do projeto, houve uma coleta e análise dos dados visando conhecer e entender profundamente sua natureza podendo assim indicar aspectos relevantes, o que inclui a identificação de anomalias, tipos de dados, relacionamentos, quantidade de registros, problemas de qualidade e outros pontos que permitam desenvolver hipóteses iniciais acerca dos dados coletados.

3. **Preparação dos dados:** esta etapa teve como objetivo de preparar, transformar atributos de modo a tornar o conjunto de dados adequado para servir de entrada para os algoritmos de mineração dados. Também faz parte dessa etapa a identificação e exclusão de variáveis com anomalias identificadas na etapa anterior.
4. **Modelagem:** nesta etapa pretendeu-se selecionar e aplicar os algoritmos de mineração buscando alcançar o melhor resultado possível, documentando todo o processo desde a geração dos modelos, a sua validação, a seleção das melhores regras e a sua interpretação.
5. **Avaliação:** esta etapa da metodologia buscou-se avaliar se os modelos finais gerados foram aprovados e se os conhecimentos adquiridos com estes modelos foram utilizados na etapa de implantação. Assim, buscou-se analisar os modelos de acordo com os parâmetros de qualidades e objetivos comerciais do projeto de mineração.
6. **Implementação:** nesta etapa do projeto buscou-se descrever como o conhecimento adquirido com o projeto de mineração de dados pode ser aplicado na organização em suas atividades corriqueiras.

O processo não se encerra na implementação. Após a implementação, o conhecimento adquirido servirá ainda como subsídio para o desenvolvimento do projeto ML em futuras versões, iniciando um novo ciclo. A figura 30 apresenta as etapas do modelo CRISP-DM:

Figura 30 – Fluxo do modelo CRISP-DM



Fonte: Etapas do Framework CRISP-DM Adaptado de (WIRTH; HIPP, 2000).

## 5.3 Compreensão do Negócio

Esta etapa descreveu os objetivos principais do estudo, mas em uma perspectiva comercial, alinhando esses entendimentos com o problema resolvido com os conhecimentos adquiridos com o ML, juntamente com o plano projetado para se atingir os objetivos descritos na seção 1.4.

### 5.3.1 Determinar os objetivos do negócio

Por se tratar de uma instituição que recebe repasse do governo, é necessário um cuidado ainda maior na forma de utilizar os recursos dos contribuintes e com a máxima eficiência. Nesse contexto, a instituição possui vários cursos em suas várias modalidades e eixos tecnológicos. Podem-se segmentar os cursos em 3 tipos: FIC, Técnico e Superior. Em um cenário ideal, a instituição deveria formar todos os alunos matriculados, mas em seu relatório anual referente a 2018 observam-se que dos 39.246 alunos que finalizaram seus cursos em 2018, 9.123 alunos evadiram mostrando que a instituição tem atualmente uma taxa de evasão de aproximadamente **25%**.

Assim, pode-se estabelecer que o principal objetivo desse trabalho, para a instituição, foi testar e/ou aplicar técnicas de ML na busca de possíveis razões para a evasão escolar, tornando assim, a instituição mais eficiente com a aplicação dos seus recursos públicos, proporcionando o máximo retorno do imposto pago pelo cidadão brasileiro com uma educação de qualidade e a redução da porcentagem nos índices de evasão por meio de ações decorrentes dos conhecimentos adquiridos com o presente estudo.

### 5.3.2 Avaliar a situação

Esse estudo conta inicialmente com um hardware Notebook Dell G5 com processador Intel Core i7 9th Gen, 16gb de memória RAM, 256gb HD SSD + HDD de 1TB (5.400 RPM) e placa de vídeo GeforceGTX 8GB do próprio autor. Sobre os softwares utilizados: o sistema operacional Windows 10 Pro, o Microsoft Office 365, o software de banco de dados SQL Serve e o Datatools da Microsoft. Para a execução dos modelos utilizou *Microsoft Azure Machine Learning*(MLS).

Os dados cedidos pela instituição estão em um banco de dados SQL Server e foram manipulados para uma base que foi criada. Sobre os requisitos necessários para o projeto foi solicitado o acesso aos dados acadêmicos e financeiros dos alunos. A obtenção dos dados foi realizada por intermédio de solicitação formal via comunicado interno para o setor responsável Gerência Tecnologia Informação (GTI). A cópia dessa autorização se encontra no anexo B. Após esse processo foi disponibilizada uma cópia do banco de dados do SA em um ambiente de desenvolvimento e depois foi efetuada a cópia das tabelas que foram utilizadas para o estudo.



Como esse banco de dados possui informações acadêmicas, financeiras e pessoais de alunos e responsável financeiro, o autor se comprometeu a manter sigilo sobre as informações ali contidas, além de se responsabilizar por não divulgar e nem repassar nenhuma das informações mencionadas a terceiros. Os dados foram utilizados somente com o objetivo de realização da pesquisa e não serão divulgadas informações pessoais ou que permitam a identificação pessoal de alunos. Todas as informações extraídas do banco de dados e utilizadas para publicação do estudo foram informações estatísticas e agregadas, que não possibilitam identificação individual de alunos.

Uma restrição do estudo foi o tamanho da base de dados do SA que é muito grande. Sendo assim, foi necessária uma análise cuidadosa das tabelas que foram utilizadas. Após a carga realizada, o banco que foi disponibilizado no ambiente de desenvolvimento da instituição foi apagado. Outro ponto é que se utilizou uma conta gratuita para acessar o MLS. Isso limita alguns algoritmos e testes, mas esperou-se conseguir bons resultados com esse ambiente.

Como meta, foi estabelecido utilizar pelo menos três tipos de algoritmos diferentes. Sobre critério para um resultado bem-sucedido, foi estabelecido que os modelos gerados devessem possuir a eficiência de no mínimo **75%** na classificação correta.

### 5.3.3 Ferramentas e técnicas

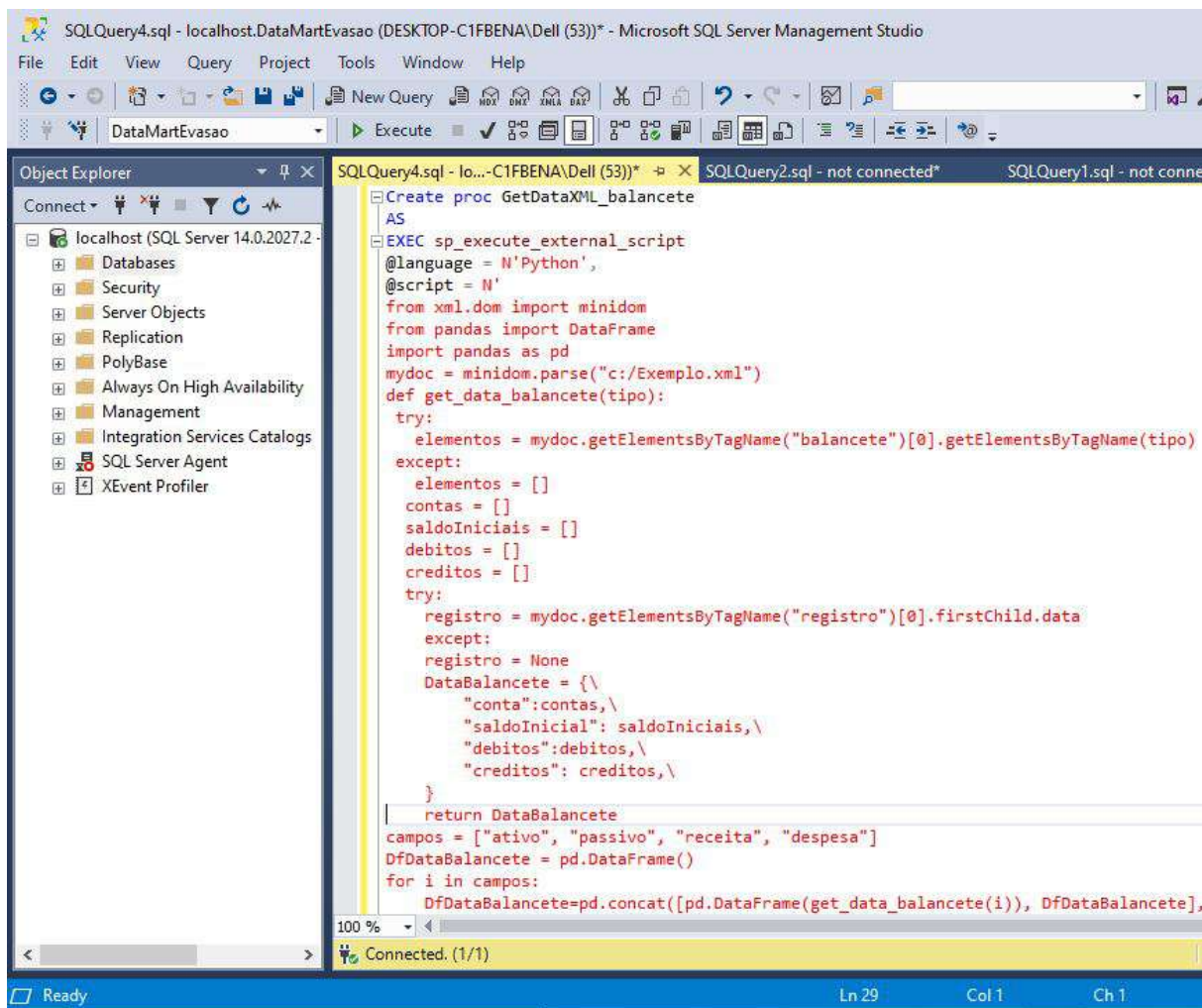
Sobre as ferramentas e técnicas utilizadas, foram analisadas algumas formas comuns de realizar esse trabalho.

*Structured Query Language* (SQL) é uma linguagem de programação em que o principal objetivo é a manipulação, controle, transação e consultas de dados. Desse modo tem como função ser a interface entre o utilizador e o sistema gestor da base de dados seja, por exemplo, SQL Server ou Oracle. Essa linguagem é utilizada para modelos relacionais.

O SQL Server é um sistema que gera bases de dados relacionais, desenvolvido pela Microsoft. A sua principal função é de armazenar dados que sejam fornecidos por outros softwares. Em suas versões mais recentes, a Microsoft incorporou ao SQL Server a linguagem R e Python. Isso permitiu que o desenvolvedor pudesse, por exemplo, desenvolver um código em Python e executá-lo como uma procedure no SQL SERVER. A figura 31 mostra a IDE do SQL Server com um exemplo de um procedure em Python.



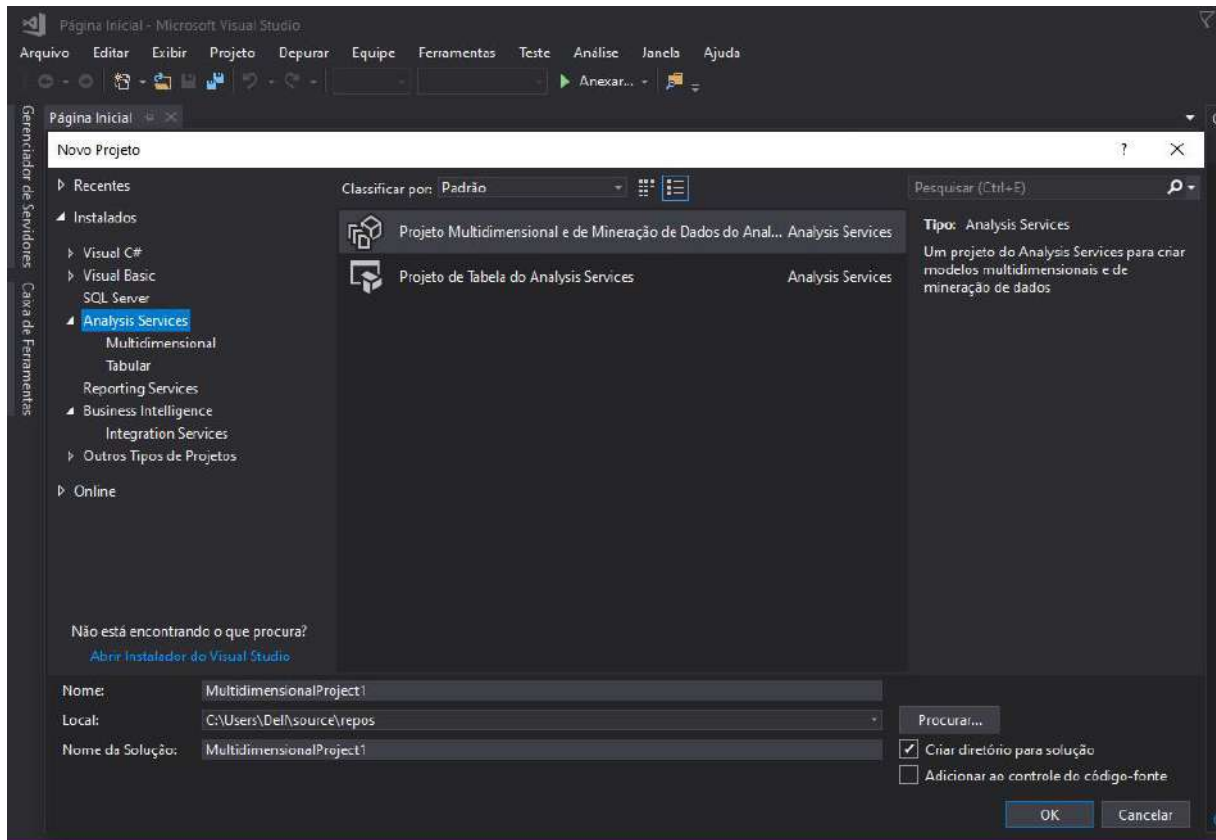
Figura 31 – Ambiente de desenvolvimento do SQL Server



Fonte: Elaborado pelo Autor (2019).

O Visual Studio Data Tools é um conjunto de programas criado pela Microsoft. Esta ferramenta é conhecida como uma "Integrated Development Environment", um software com um editor de texto muito poderoso. O Data Tools encontra-se dividido em três módulos: *SQL Server Integration Services* - SSIS, *SQL Server Analysis Services* - SSAS e *SQL Server Reporting Services* - SSRS. Essa ferramenta será bem útil, tendo em vista que todo o processo de carga, ou seja, todas as tabelas relacionadas ao processo de evasão que fazem parte do banco de dados do SA serão copiadas para o nosso ambiente de estudo utilizando o SSIS. Na figura 32 podem-se observar os módulos SSIS, SSAS e SSRS.

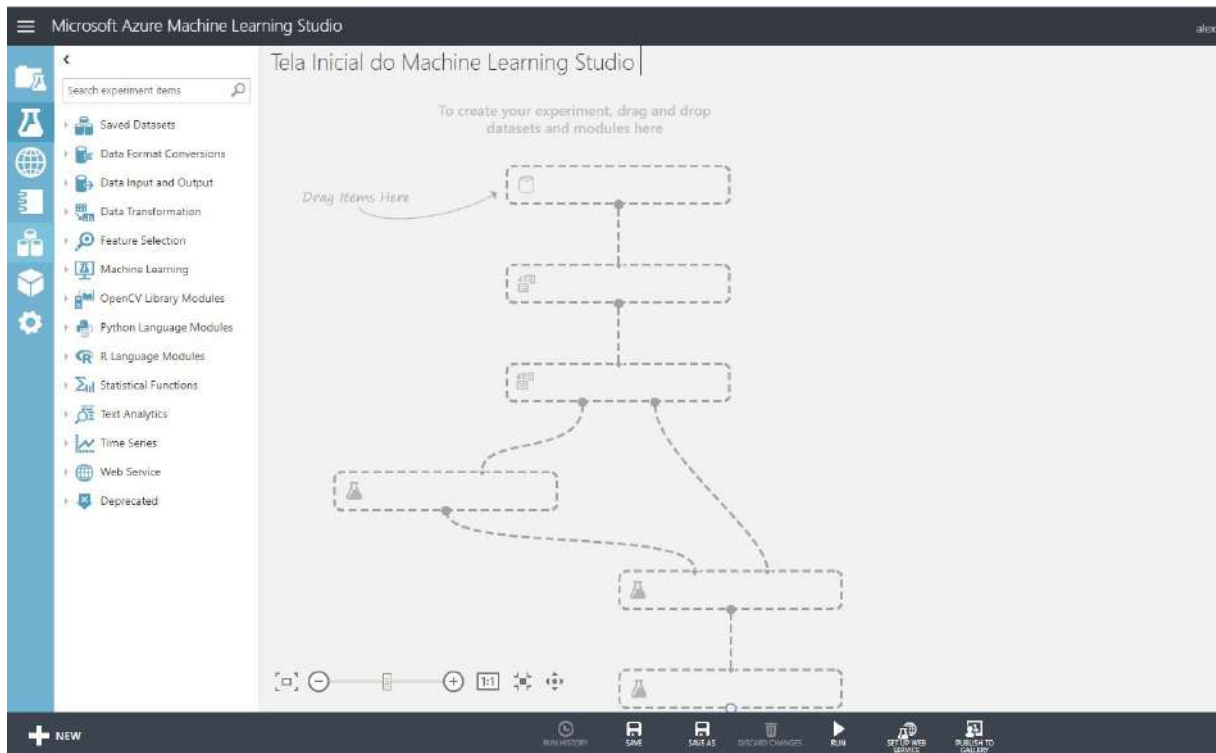
Figura 32 – Ambiente de inicial do visual studio data tools



Fonte: Elaborado pelo Autor (2019).

Uma ferramenta da Microsoft que vem sendo utilizada no Azure é MLS que foi desenvolvido para ser uma ferramenta do tipo assistente na criação de soluções de ML. Com o MLS os blocos de comando podem ser arrastados e assim montar seu modelo. Por ser uma ferramenta que faz parte do Azure, é possível publicar um modelo como um serviço na Web e ser consumidos por aplicativos personalizados ou ferramentas de BI como o Power BI. Na figura 33 é mostrada a tela principal a qual foi trabalhada nesse estudo na maioria das vezes.

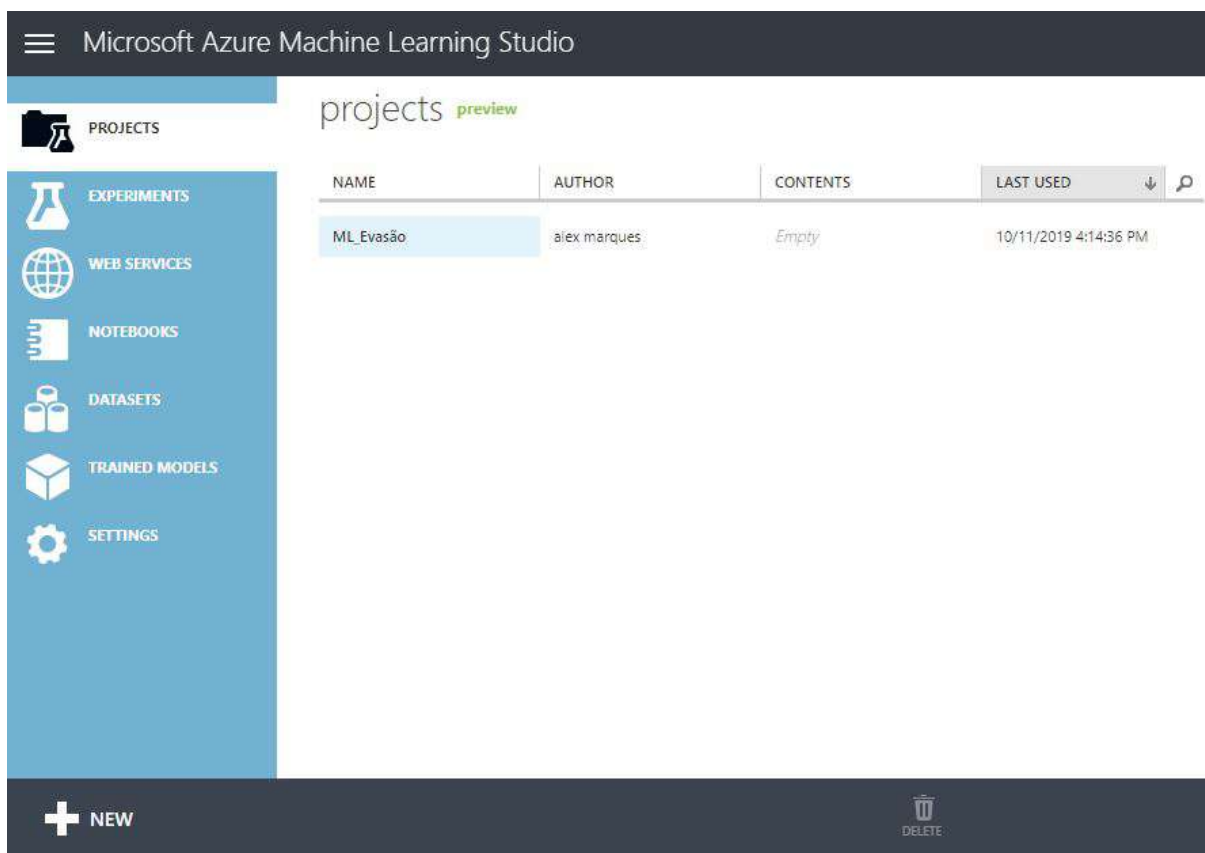
Figura 33 – Machine learning studio - MLS



Fonte: O Autor adaptado da tela do MLS (2019).

Para desenvolver um modelo de análise preditiva, geralmente utilizam-se dados de uma ou mais fontes, os preparam e os analisam por meio de várias funções estatísticas e de manipulação de dados, além de produzir um conjunto de resultados. Desenvolver um modelo como este é um processo iterativo. À medida que se altera as diversas funções e seus parâmetros, seus resultados convergem até que você esteja satisfeito, com um modelo treinado e eficiente. A figura 34 mostra a tela inicial do MLS.

Figura 34 – Tela principal do machine learning studio - MLS

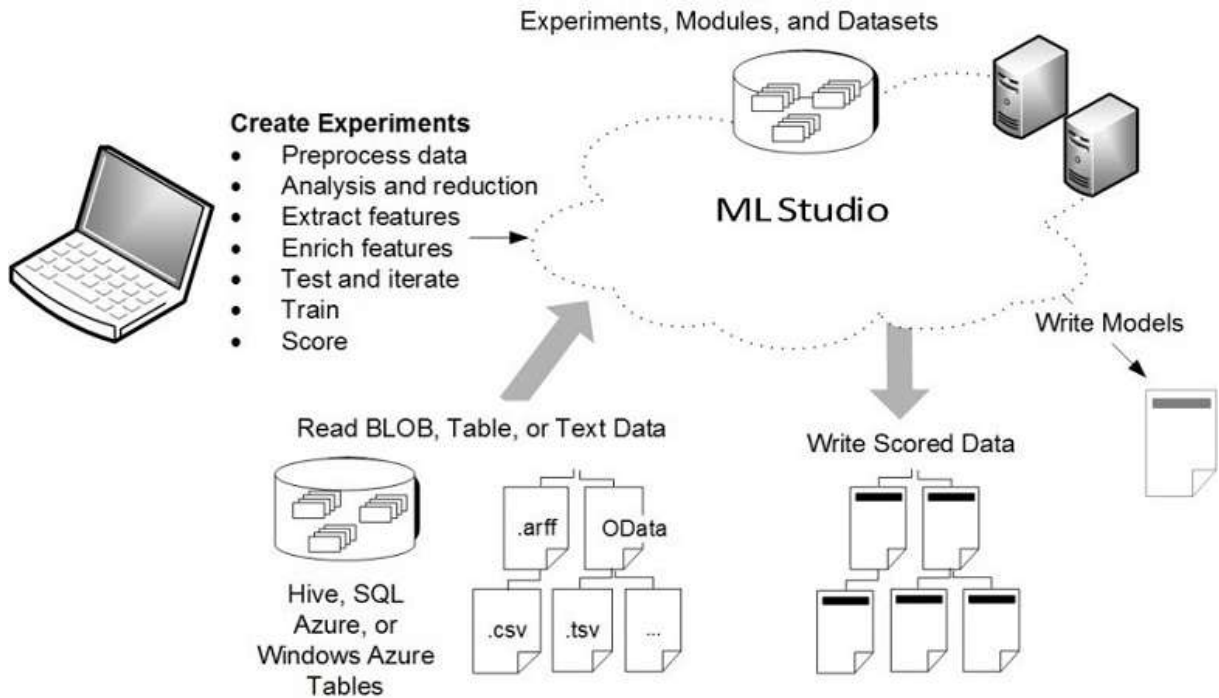


Fonte: O Autor adaptado da tela do MLS (2019).

O Azure MLS oferece um workspace visual e comunicativo para compilar, testar e iterar em um modelo de análise preditivo. Arrastam-se e se soltam conjuntos de dados e módulos de análise em telas interativas ligando-as para criar um teste conforme a figura 28. Para iterar no design de modelo, edita-se o teste, salva uma cópia, se desejado, e executa-o novamente. Quando estiver finalizado, pode-se mudar o teste de treinamento em uma experiência preditiva e, em seguida, publicá-la como um serviço Web para que o modelo possa ser acessado por outras pessoas.

Não há necessidade de programação, basta conectar visualmente os conjuntos de dados e módulos para construir seu modelo de análise preditivo. A figura 35 a seguir mostra o diagrama MLS.

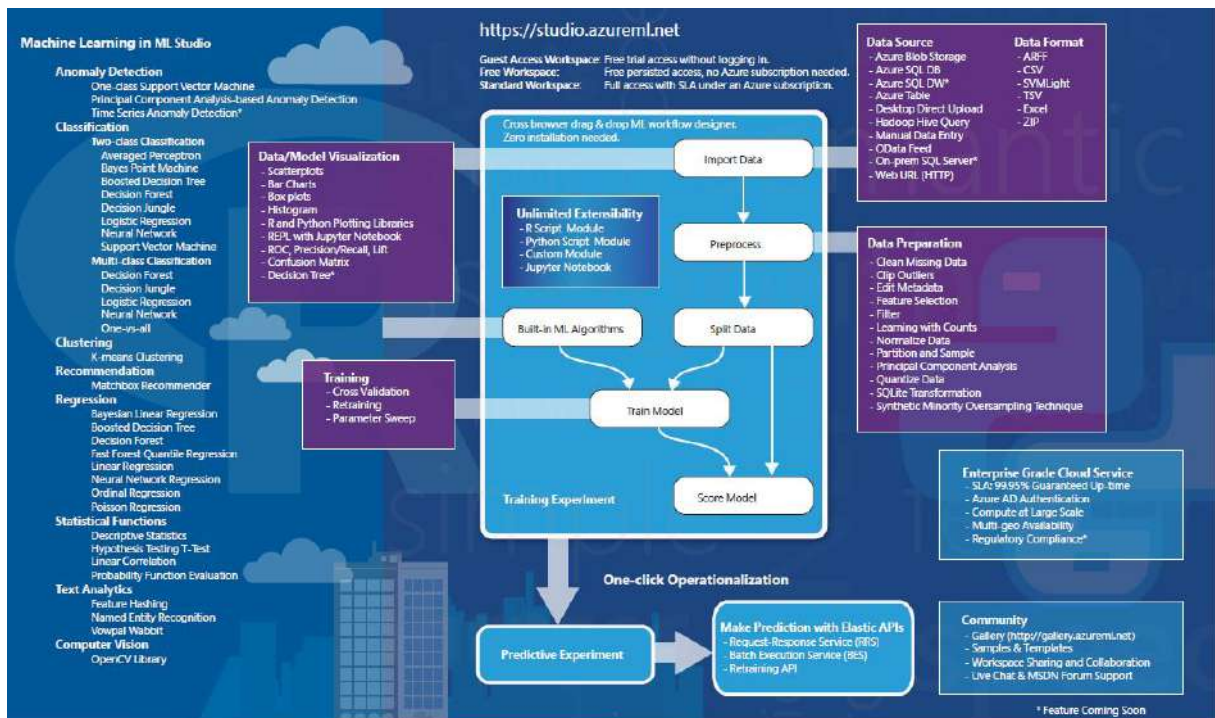
Figura 35 – Experimentos, módulos e base de dados



Fonte: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio>.

O MLS é uma ferramenta muito poderosa com muitos recursos. Com o objetivo de entender melhor o seu funcionamento, a figura 36 mostra um diagrama com a visão geral dos recursos do Microsoft Azure MLS.

Figura 36 – Diagrama azure machine learning studio - MLS



Fonte: <https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio>.

Como Ferramenta foi escolhida a MLS para utilização desse estudo, levando-se em consideração que o autor possui certo conhecimento adquirido durante disciplina de Inteligência Artificial do presente curso de mestrado. Além disso, o autor possui mais afinidade e conhecimento técnico das ferramentas da Microsoft.

## 5.4 Compreensão dos Dados

Nessa etapa do estudo, pretendeu-se conhecer mais profundamente a fonte de dados, seus atributos e aspectos relevantes, o que inclui a identificação de anomalias, tipos dos dados, quantidade de registros, entre outros.

### 5.4.1 Coleta de dados

A instituição pesquisada disponibilizou um banco de dados no ambiente de desenvolvimento para que uma primeira análise fosse feita a fim de entender quais atributos de quais tabelas seriam utilizadas, visto que o banco de dados da instituição é muito grande. Como o autor já trabalha há muito tempo na instituição e tem conhecimento do tamanho e da complexidade do sistema, foi necessário rever algumas regras com as áreas de negócio, mapear todas as tabelas que fariam parte do estudo e por fim fazer uma extração das informações necessárias para o estudo em questão. Primeiro, foram mapeadas as tabelas que contêm as informações sobre os alunos, sobre os cursos e sobre a situação financeira dos alunos nos cursos.

O SA é um sistema muito grande com vários módulos e com um banco de dados SQL SERVER muito grande. Possui mais de 20.000 objetos entre eles: tables, procedures, views e functions, por exemplo. Assim, foi necessário fazer uma extração das informações necessárias para outro banco de dados. Para esse estudo criou-se um banco de dados, **DataMartEvasao**, onde se concentram todas as informações necessárias para a elaboração do modelo. Foram listadas mais de 170 tabelas contendo informações de suma importância para o estudo sobre a evasão. Essas tabelas foram divididas entre os seguintes temas: curso, turma, aluno e financeiro. A tabela 8 ajuda a entender melhor esse total por tema. Com essas tabelas é possível atingir uma quantidade de atributos e informações muito importante a respeito do aluno, do curso e da instituição. Isso sugere inicialmente uma boa capacidade de análise para os modelos de ML. A relação de todas as tabelas relacionadas por tema se encontra no anexo C.



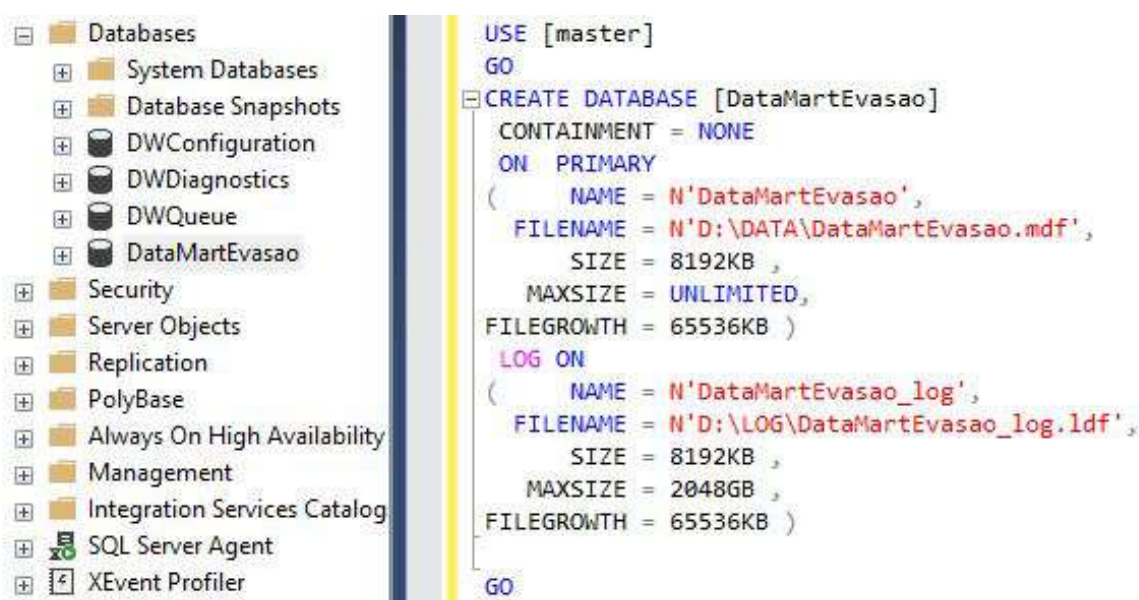
Tabela 8 – Quantidade de tabelas por tema

Tabelas por tema	Quantidade
Curso	29
Turma	73
Aluno	55
Financeiro	16
Total	173

Fonte: Elaborado pelo Autor (2019).

Para a criação do banco de dados foi utilizado o script conforme a figura 37 a seguir:

Figura 37 – Criação do Banco de Dados



```

USE [master]
GO
CREATE DATABASE [DataMartEvasao]
    CONTAINMENT = NONE
    ON PRIMARY
    (
        NAME = N'DataMartEvasao',
        FILENAME = N'D:\DATA\DataMartEvasao.mdf',
        SIZE = 8192KB ,
        MAXSIZE = UNLIMITED,
        FILEGROWTH = 65536KB )
    LOG ON
    (
        NAME = N'DataMartEvasao_log',
        FILENAME = N'D:\LOG\DataMartEvasao_log.ldf',
        SIZE = 8192KB ,
        MAXSIZE = 2048GB ,
        FILEGROWTH = 65536KB )
GO

```

Fonte: Elaborado pelo Autor (2019).

Para realizar carga dos dados para o Banco de dados criado, **DataMartEvasao**, foi utilizado o conceito de ETL, do inglês **Extract Transform Load** (Extração, Transformação e Carregamento).

ETL é o processo que coleta os dados relevantes dos bancos de dados transacionais, transformando-os em um padrão (por meio de processos de limpeza, tratamento e classificação) e os carrega nas bases analíticas (BARBIERI, 2011). Resumidamente esse processo é composto por:

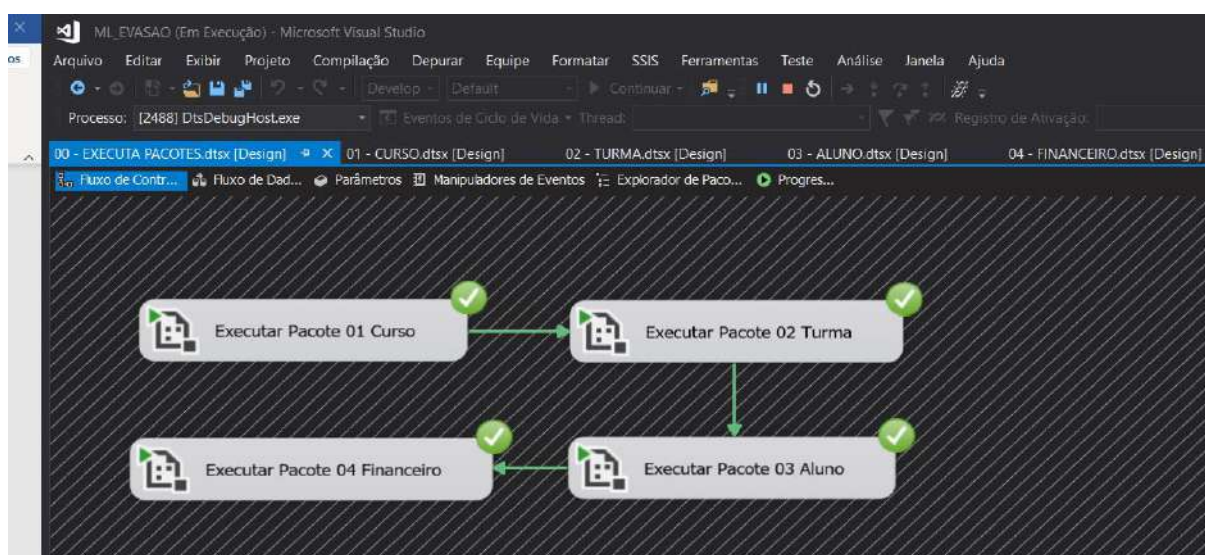
1. **Extração:** leitura dos dados de um ou mais banco e dados;
2. **Transformação:** conversão dos dados extraídos de sua forma anterior para a forma em que precisam estar, a fim de serem colocados em um data warehouse, data mart ou apenas em outro banco de dados;

3. **Carga:** colocação dos dados no data warehouse ou data mart.

Para essa atividade utilizou-se a ferramenta de ETL da Microsoft, o SSIS. Criou-se um projeto no dataTools com nome **ML-EVASAO**. Para essa atividade criou-se um pacote para cada tema com o nome: **01 - Curso**, **02 - Turma**, **03 - Aluno** e **04 - Financeiro** e o **00 - Executa Pacotes** que foi criado para executar todos os pacotes de uma vez.

Quando se executou o pacote **00 - Executa Pacotes**, todos os blocos foram executados com sucesso. Pode-se observar isso pelo símbolo verde em cada um dos blocos. A figura 38 ajuda a entender melhor o sucesso da execução desse pacote.

Figura 38 – ETL - Carga das tabelas



Fonte: Elaborado pelo Autor (2019).

#### 5.4.2 Descrição dos dados

O objetivo dessa etapa foi examinar detalhadamente todos os dados que foram copiados para o banco de dados **DataMartEvasao**. Assim, conforme descrito anteriormente os dados que foram coletados estavam espalhados em várias tabelas com atributos distintos entre elas. Neste sentido, verificou-se a necessidade de unificar essas tabelas em uma tabela com todos os dados antes de se realizar a descrição dos dados, para não se desperdiçar dados que não poderiam ser utilizados e assim não se desperdiçar tempo e esforço de trabalho. A figura 39 mostra o script da consulta criada e posteriormente os atributos que fariam parte da tabela que foi criada.

A partir da consulta gerada, criaram-se alguns parâmetros para analisar os atributos e o seu significado. Assim, segue alguns dos parâmetros analisados sobre cada atributo e o significado dessas informações:



- **Tabela Origem:** visto que foram copiadas muitas tabelas para a base de estudo, essa coluna mostrou de qual tabela o campo se originou.
- **Atributo Origem:** demonstrou o nome do atributo em sua tabela de origem.
- **Nome Atual:** nome que foi dado ao atributo após a junção dos atributos em uma única tabela dando origem ao dataSet de estudo.
- **Tipo:** essa informação foi importante, pois informa se uma variável era um número representado por meio de números inteiros ou reais. Sendo assim classificada como **numérica**, ou se a variável era **String**, sendo representada por valores não numéricos como letras ou palavras e classificada como **nominal**.
- **Ausência:** do inglês Missing, essa variável ajudou a verificar a quantidade de valores que não estavam preenchidos e se devesse assim usar ou não a variável. Por exemplo: se uma determinada variável tem **20%** de missing, isso significa que **20%** dos registros não possuem informações. Esperou-se no dataSet um valor aproximado de **0%** de missing.
- **Distintos:** essa variável ajudou a identificar valores diferentes. Por exemplo, a variável **Sexo** é do tipo nominal e é representa pelos valores **M** ou **F**. Nesse caso essa variável teve apenas 02 valores distintos. Já uma variável numérica tendeu a ter uma quantidade bem alta de números distintos.

Figura 39 – Script da consulta SQL com todas as tabelas

The screenshot displays the Microsoft SQL Server Management Studio interface. The main window shows a SQL query script with the following content:

```
--select * from tbModalidadeEnsino
IF OBJECT_ID('tempdb..#TbBaseFumec') IS NOT NULL
DROP TABLE #TbBaseFumec

SELECT distinct b.idaluno,b.sexo, Idade,b.chvcidade,b.idestadoCivil,b.idTitulacaoAluno,b.chvCorRaca,
b.chvOcupacao,b.idfaixaRendaMensal,b.chvOrigemEscolaEnsinofundamental,b.chvOrigemEscolaEnsinomedio,a.idturmaalu
Case when a.idsituacaoturma in(1,6,10,15,24) then 'Não Evadido' else 'Evadido' end as SituacaoAluno,e.chvturmag
g.chvmodalidadeensino,d.turma,d.sigla,d.codturma,d.vlrturma,d.datiniocioturma,d.datterminoturma,
d.qtdcargahorariaturma,d.chvhorario,i.idturno,e.codturmagrade,e.chvidundnegmodopeedif,e.vlrturmagrade,e.noacur
year(e.datiniocioturmagrade)AnoInicioTurma, case when month(e.datiniocioturmagrade)<=6 then 1 else 2 end as Semes
year(e.datterminoturmagrade)AnoTerminoTurma, case when month(e.datterminoturmagrade)<=6 then 1 else 2 end as Se
e.qtdCargaHorariaTurmagrade,j.unidadenegocio,j.nomcidade,l.desRegiao,j.chvUnidadeNegocio,j.chvRegiao
--,isnull(k.inadimplencia,0)inadimplencia
```

The Results pane shows the following data:

	SituacaoAluno	sexo	idade	grupoidade	CidadeAluno	EstadoCivil	Titulacao	CorRaca	Ocupacao	IgEmpregado	FaixaRendaMensal	OrigemEscolaEnsinofundamental	OrigemEscolaEnsinomedio
1	0	0	18	0	2619	1	7	1	2	1	1	1	1
2	0	0	18	0	2703	1	7	1	2	1	3	1	1
3	0	0	18	0	2703	1	7	2	8	0	3	1	1
4	0	0	18	0	2703	1	7	3	0	0	3	1	1
5	0	0	18	0	2722	1	7	0	8	0	0	1	1
6	0	0	18	0	3136	1	7	1	0	0	0	1	1
7	0	0	18	0	3136	1	7	1	2	1	0	1	1
8	0	0	18	0	3136	1	7	1	8	0	0	1	1
9	0	0	18	0	3201	1	7	2	8	0	2	1	1
10	0	0	18	0	3201	1	7	3	8	0	1	1	1
11	0	0	18	0	3211	1	7	2	1	0	0	1	1
12	0	0	18	0	3211	1	7	3	7	0	0	1	1
13	0	0	18	0	3281	1	7	2	2	1	2	1	1

The status bar at the bottom indicates: Query executed successfully. 00:00:01 | 113.683 rows

Fonte: Elaborado pelo Autor (2019).

Criou-se uma tabela com os parâmetros citados e foram classificados os atributos que poderão ser utilizados nos modelos. A tabela 9 apresenta alguns desses atributos. A relação completa pode ser vista no anexo D.

Tabela 9 – Relação de atributos

<b>Atributo</b>	<b>001</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	idSexo
<b>Nome Atual</b>	Sexo
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	002
<b>Valores e quantidade por instâncias</b>	M(934530), F(475508)
<b>Atributo</b>	<b>002</b>
<b>Tabela Origem</b>	tbaluno
<b>Nome de Origem</b>	idCidade
<b>Nome Atual</b>	CidadeAluno
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	1.234
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	<b>003</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	idEstadoCivil
<b>Nome Atual</b>	EstadoCivil
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	006
<b>Valores e quantidade por instâncias</b>	1(1044057), 3(313565), 5(35817), 7(4109), 10(9909), 11(2581)

Fonte: Elaborado pelo Autor (2019).

### 5.4.3 Exploração dos dados

Essa atividade foi realizada em conjunto com a anterior, visto que para descrever a união de várias tabelas em uma única e consistir os dados, foi necessário explorar todos os atributos disponíveis nas mais de 170 tabelas. Também foi utilizando o software

SQL Server para realizar uma exploração mais detalhada dos dados, quando foi possível identificar os registros e seus aspectos, possibilitando a descrição detalhada de todos os dados (realizado na tarefa anterior). Analisando alguns atributos, como por exemplo, a **DataNascimento**, era possível perceber a necessidade de transformar seus valores, a fim de balancear os registros e agrupar informações.

#### 5.4.4 Qualidade dos dados

Analisando a tabela 14, foi possível perceber que os dados possuem um bom nível de qualidade no que diz respeito aos dados ausentes, pois todos os atributos possuem **0%** dados ausentes. Também a respeito dos tipos dos dados, na sua grande parte mais de **95%** dos dados são do tipo nominal. Também pode-se ressaltar a qualidade dos dados na gama de informações, apresentando **1.410.038** informações diferentes a respeito de um aluno, possibilitando mais opções de testes com diferentes atributos na criação dos modelos de regras. Por fim, ressalta-se que a quantidade de registros é grande o suficiente para o uso de todos os algoritmos de mineração disponíveis no MLS. Sendo assim, conclui-se que a base de dados apresentou requisitos de qualidade consideráveis para o cumprimento dos objetivos do projeto, o que tornou viável sua continuação com os dados selecionados até então.

### 5.5 Preparação dos Dados

Essa etapa teve como objetivo organizar os atributos do **DataSet**, ou seja, do conjunto de dados para aplicação dos algoritmos de ML. Caso existisse alguma variável, identificada na etapa anterior, com algum tipo de irregularidade, aqui foi feita uma remoção ou adequação.

#### 5.5.1 Seleção de dados

Essa tarefa teve como objeto a construção da base de dados que foi utilizada na geração dos modelos para aplicação dos algoritmos de ML. Foram realizados testes na base de dados, e foi possível verificar quais dos atributos se comportavam melhor com os algoritmos disponíveis no MLS. E por fim, foram verificados mais a fundo quais atributos faziam sentido para a geração da base de dados dos alunos evadidos e se as informações acrescentavam de fato algum conhecimento para a instituição.

Após algumas análises, conclui-se que alguns dos atributos deveriam ser excluídos do **dataSet**. A tabela 10 apresenta a lista com os atributos que não foram utilizados. Os motivos que levaram a tal decisão estão disponíveis no anexo E. Alguns atributos não foram utilizados em sua forma original. Como foi visto mais a frente na presente pesquisa, alguns atributos deram origem a outros.

Tabela 10 – Relação dos atributos que foram excluídos

Nº	Nome do Atributo
01	DatMatricula
02	idIngresso
03	datInicioTurma
04	DatNasc
05	idNaturid
06	idestado
07	flgCanhoto
08	codturma
09	sigla
10	turma
11	flgResponsavel Financeiro
12	qtdVagas
13	chvsala
14	codturmagrade
15	nomCurso
16	vlrurma
17	datFimTurma
18	desProfissao
19	idEmprego
20	desCargo

Fonte: Elaborado pelo Autor (2019).

### 5.5.2 Limpeza de dados

Nessa tarefa foram feitas as correções dos dados de acordo com as verificações de qualidade. Foram alterações nos nomes dos atributos, removendo acentos e adaptando os nomes de forma a se compreender melhor o seu significado. Outro ponto foi a quantidade de registro. Havia uma preocupação se a quantidade de registros seria suficiente, visto que, normalmente, os algoritmos de ML precisam de uma quantidade razoável. Felizmente conseguiu-se gerar um dataset com uma quantidade de atributos e de registros suficientes para se analisar o perfil do aluno.

### 5.5.3 Construção de dados

O principal objetivo desta tarefa foi descrever os atributos que foram criados a partir de outros. Segue a relação dos atributos derivados e presentes no dataSet final conforme a tabela 11:

Tabela 11 – Relação de atributos derivados

<b>Atributo</b>	<b>Idade</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	datNasc

<b>Valores dos registros criados</b>	Diversas idades entre 16 e 80 anos.
<b>Finalidade</b>	Verificar qual a idade que mais esta evadindo.
<b>Atributo</b>	
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	datNasc
<b>Valores dos registros criados</b>	Criamos um total de 6 grupos. A tabela nos mostra os grupos que foram criados e as regras utilizadas na criação dos grupos
<b>Finalidade</b>	Entender em qual grupo de idade está ocorrendo o maior número de evasão.
<b>Atributo</b>	
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	datInicioTurma
<b>Valores dos registros criados</b>	Turma que iniciaram no ano 2006 até 2019
<b>Finalidade</b>	Saber qual o ano que a turma iniciou e relacionar com os alunos que evadiram.
<b>Atributo</b>	
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	datInicioTurma
<b>Valores dos registros criados</b>	1º ou 2º Semestre
<b>Finalidade</b>	Saber qual o semestre que os alunos estão mais evadindo.
<b>Atributo</b>	
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	datFimTurma
<b>Valores dos registros criados</b>	Turma que terminaram no ano 2007 até 2023
<b>Finalidade</b>	Saber qual o ano que a turma iniciou e relacionar com os alunos que evadiram.
<b>Atributo</b>	
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	datFimTurma

<b>Valores dos registros criados</b>	1º ou 2º Semestre
<b>Finalidade</b>	Saber qual o semestre que os alunos estão mais evadindo.

Fonte: Elaborado pelo Autor (2019).

Tabela 12 – Grupo de idade criado a partir do atributo  
DatNasc

<b>Grupo Idade</b>	<b>Regra</b>	<b>Valor</b>
16 a 30	idade 15 and idade =30	0
31 a 40	idade 30 and idade =40	1
41 a 50	idade 40 and idade =50	2
51 a 60	idade 50 and idade =60	3
61 a 70	idade 60 and idade =70	4
70	idade 70	5

Fonte: Elaborado pelo Autor (2019).

### 5.5.4 Integração de dados

Nesta tarefa procurou-se pelos atributos que poderiam ser mesclados, com o objetivo de melhorar o desempenho dos algoritmos ML e também facilitar a sua compreensão. Seguem as informações sobre os atributos transformados neste processo conforme a tabela 13.

Tabela 13 – Relação de atributos mesclados

Atributo	NecessidadeEspecial
Tabela Origem	tbAluno
Nome de Origem	idNecessidadeEspecial
Objetivo da modificação	Os registros estavam divididos pelo tipo de necessidade especial que o aluno marcou no formulário no momento da matrícula. Porém, apenas é importante compreender se o aluno possui ou não uma necessidade. Assim, se o aluno possui alguma necessidade especial convertemos para (1) caso contrário (0).
Atributo	SituacaoAluno
Tabela Origem	tbTurmaGrade
Nome de Origem	idSituacaoTurmaAluno
Objetivo da modificação	Os registros estavam divididos em vários status conforme apresentado na tabela . Assim fizemos uma conversão desses status para Evadido e Não Evadido conforme explicado na seção 4.2.

Fonte: Elaborado pelo Autor (2019).

### 5.5.5 Formatação de dados

Buscou-se nessa tarefa transformar os dados a fim de melhorar o desempenho dos algoritmos de ML, para isso foram realizados testes com todos os 32 atributos selecionados para a base de dados final. A tabela 14 mostra o dataset final que foi usado nos modelos de ML.

Tabela 14 – Base final utilizada na geração do modelo ML

ID	Variável	Classificação
1	Turno	Acadêmica
2	Vlrturmagrade	Acadêmica
3	Curso	Acadêmica
4	AnoInicioTurma	Acadêmica
5	SemestreIncioTurma	Acadêmica
6	AnoTerminoTurma	Acadêmica
7	SemestreTerminoTurma	Acadêmica
8	qtdCargaHorariaTurmagrade	Acadêmica
9	Unidadenegocio	Acadêmica
10	CidadeEscola	Acadêmica
11	RegiaoEscola	Acadêmica
12	Frequência	Acadêmica
13	IdDisciplina	Acadêmica
14	Idturma	Acadêmica
15	Idaluno	Acadêmica
16	StatusFinanceiro	Acadêmica
17	TipoBaixa	Acadêmica
18	Status	Acadêmica
19	NecessidadeEspecial	SocioEconômica
20	Sexo	SocioEconômica
21	Idade	SocioEconômica
22	GrupoIdade	SocioEconômica
23	OrigemEscolaEnsinomedio	SocioEconômica
24	CidadeAluno	SocioEconômica
25	EstadoCivil	SocioEconômica
26	Titulacao	SocioEconômica
27	CorRaca	SocioEconômica
28	Ocupacao	SocioEconômica
29	FlgEmpregado	SocioEconômica
30	FaixaRendaMensal	SocioEconômica
31	OrigemEscolaEnsinofundamental	SocioEconômica
32	SituacaoAluno(Evadido?)	Acadêmica/SocioEconômica

Fonte: Elaborado pelo Autor (2019).



## 5.6 Modelagem

Existem diversos algoritmos que podem ser utilizados para realizar a ML, sendo esses divididos basicamente em duas categorias: (i) os supervisionados, que utilizam uma classe padrão para agrupar os registros e gerar regras de acordo com essa classe, e (ii) os não supervisionados, que não utilizam uma classe padrão e geram regras com as associações mais comuns entre os registros. Nesse estudo, foram utilizados algoritmos supervisionados, tendo em vista que a classe padrão é o atributo **SituaçãoAluno** que contém a informação se o aluno é evadido ou não.

Os algoritmos supervisionados também são conhecidos como algoritmos de classificação. O MLS da microsoft separa os algoritmos em 4 categorias:

- **Anomaly Detection**
- **Classification**
- **Clustering**
- **Regression**

Nesse estudo a concentração foi em utilizar apenas os algoritmos que estavam na categoria **Classification** do MLS. Conforme estabelecido na etapa de Compreensão do negócio do presente projeto, foi estabelecido que os modelos gerados deveriam possuir a eficiência de no mínimo **75%** na Classificação Correta dos dados e que seriam utilizados no mínimo 03 algoritmos de classificação diferentes. Chegou-se a esse índice mínimo após os primeiros testes na criação de modelos, quando se percebeu que os modelos alcançaram índices de eficiência entre **65%** a **75%**. Considerando os tipos e quantidades de registros da base de dados, foram escolhidos os seguintes algoritmos de classificação:

- Algoritmo Two-Class Support Vector Machine;
- Algoritmo Two-Class logistic Regression;
- Algoritmo Two-Class Locally-Deep VSM;
- Algoritmo Two-Class Decision Jungle;
- Algoritmo Two-Class Neural Network;
- Algoritmo Two-Class Boosted Decision Tree.

## 5.7 Avaliação

Essa etapa da metodologia teve como objetivo avaliar se os modelos apresentados no final dos testes realizados seriam aprovados e se as percepções adquiridas com estes modelos poderiam ser utilizadas na etapa seguinte da implementação. Dessa forma, procurou-se avaliar os modelos seguindo os parâmetros de qualidade e os objetivos propostos nesse projeto.

Na etapa de preparação dos dados e na etapa de modelagem, foram apresentados em detalhe todos os parâmetros, modelos e as regras geradas, aceitas e as desconsideradas. Assim, dentro do próprio projeto, só foram descritos os modelos que foram aprovados pelos testes e que cumpriram os objetivos do projeto. Durante a elaboração desse projeto, foram gerados vários modelos com diversas configurações de parâmetros. Não se julgou necessário documentar todos os modelos para essa dissertação, sendo assim, todos os modelos descritos no projeto foram aceitos e utilizados na fase de implementação.

Ao se revisar o processo de ML, não foi percebida nenhuma tarefa que possa ter sido desconsiderada. Todos os procedimentos foram realizados dentro dos parâmetros e processos definidos na metodologia CRISP-DM.

Considerando-se a aceitação dos modelos e regras geradas na Etapa de modelagem da metodologia, os próximos passos do projeto compreenderam na realização da implantação dos conhecimentos no contexto da instituição objeto de estudo de acordo com os objetivos presentes nesse projeto. A implantação dos conhecimentos dentro da instituição se dará, por sua vez, após a apresentação dos resultados do projeto para o comitê de TI em um momento futuro.

## 5.8 Implementação

Essa etapa tem como principal objetivo apresentar futuramente os conhecimentos adquiridos com o projeto de ML e como ele poderá ser aplicado da instituição.

Após a conclusão do projeto de ML, foi possível entender algumas questões relacionadas a evasão. Esse conhecimento será levado em um relatório para o setor responsável dentro da instituição pesquisada para que sejam sugeridas algumas ações de combate à evasão com base nestes conhecimentos. Essas sugestões poderão ser usadas em todas as unidades com o intuito de diminuir os índices de evasão.

O projeto de ML contendo a descrição de todas as tarefas e resultados gerados compõem por si só a presente dissertação de mestrado. Assim, o relatório final citado acima corresponde a essa dissertação.

Considerando que o presente projeto foi construído com finalidades acadêmicas e

que o autor do projeto possuía relativa experiência no processo de ML, é possível concluir que o projeto gerado descreve em detalhes todos os procedimentos realizados, desde a sua concepção até a sua finalização.

Neste sentido, como revisão do projeto, pode-se considerar como assertiva o uso dos algoritmos escolhidos e o software MLS. Também foi muito positivo a receptividade dos conhecimentos gerados por meio do ML pela instituição de ensino, pois algo com essa tecnologia não havia sido ainda realizado na instituição.

## 6 TESTES E RESULTADOS OBTIDOS

Após os processos de seleção e transformação dos dados descritos nas etapas anteriores da metodologia, foram iniciados os testes ML com os atributos aplicando os algoritmos de classificação escolhidos. A fim de comparar o desempenho dos algoritmos, a tabela 15 demonstra os resultados dos testes realizados.

Tabela 15 – Desempenho dos algoritmos

Métricas	Two-Class Support Vector Machine	Two-Class logistic Regression	Two-Class Locally-Deep	Two-Class Decision Jungle	Two-Class Neural Network	Two-Class Decision Tree
True Positive	5202	6078	6612	8044	8015	8433
False Positive	2570	3110	1860	1433	626	505
False Negative	3666	2790	2256	824	853	435
True Negative	15004	14464	15714	16141	16948	17069
<b>Accuracy</b>	<b>0.764</b>	<b>0.777</b>	<b>0.844</b>	<b>0.915</b>	<b>0.944</b>	<b>0.964</b>
Recall	0.587	0.685	0.746	0.907	0.904	0.951
Precision	0.669	0.662	0.78	0.849	0.928	0.943
F1 Score	0.625	0.673	0.763	0.877	0.916	0.947
<b>AUC</b>	<b>0.856</b>	<b>0.868</b>	<b>0.917</b>	<b>0.974</b>	<b>0.988</b>	<b>0.994</b>

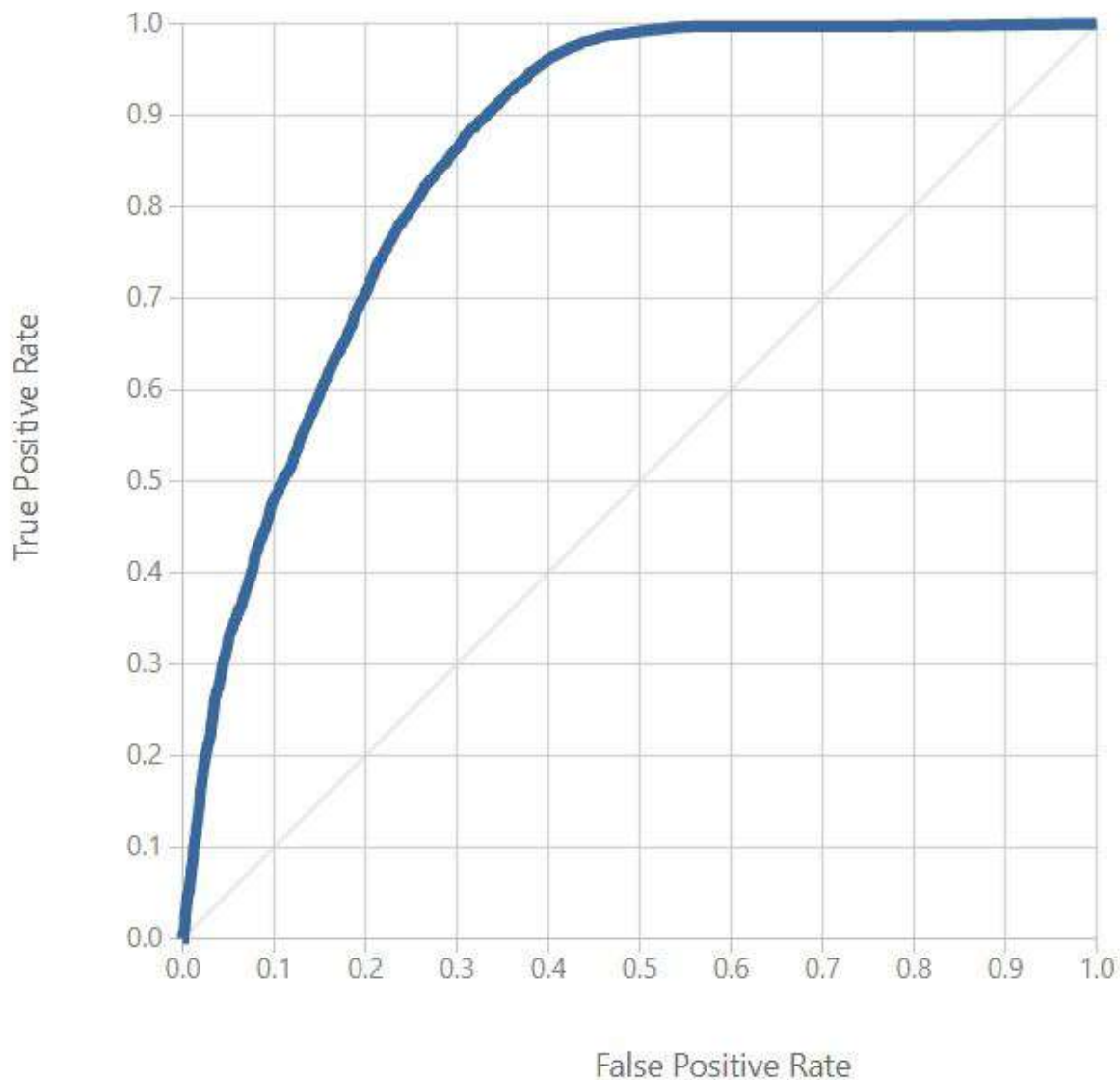
Fonte: Elaborado pelo Autor (2019).

Conforme verificado na tabela 15 é possível concluir que o algoritmo que obteve o melhor desempenho na métrica Accuracy é o **Two-Class Decision Tree**. Segue uma análise dos resultados obtidos.

### 6.1 Algoritmo Two-Class Support Vector Machine

Esse modelo requereu dados rotulados e, no processo de treinamento, o algoritmo analisou os dados de entrada e reconheceu padrões em um espaço de característica multi-dimensional chamado **hiperplane**. As figuras 40, 41 e 42 apresentam os resultados desse algoritmo.

Figura 40 – Algoritmo Two-Class Support Vector Machine



Fonte: Elaborado pelo Autor (2019).

Figura 41 – Algoritmo Two-Class Support Vector Machine

True Positive	False Negative	Accuracy	Precision	Threshold	<input type="text" value="0.5"/>	AUC
<b>5202</b>	<b>3666</b>	<b>0.764</b>	<b>0.669</b>	<b>0.5</b>		<b>0.856</b>
False Positive	True Negative	Recall	F1 Score			
<b>2570</b>	<b>15004</b>	<b>0.587</b>	<b>0.625</b>			
Positive Label	Negative Label					
<b>1</b>	<b>0</b>					

Fonte: Elaborado pelo Autor (2019).

Figura 42 – Algoritmo Two-Class Support Vector Machine

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall
(0.900,1.000]	178	47	0.009	0.670	0.039	0.791	0.020
(0.800,0.900]	1257	294	0.067	0.706	0.270	0.808	0.162
(0.700,0.800]	1442	533	0.142	0.740	0.456	0.767	0.324
(0.600,0.700]	1064	680	0.208	0.755	0.549	0.717	0.444
(0.500,0.600]	1261	1016	0.294	0.764	0.625	0.669	0.587
(0.400,0.500]	1651	1520	0.414	0.769	0.692	0.626	0.773
(0.300,0.400]	1188	1862	0.529	0.744	0.703	0.575	0.907
(0.200,0.300]	655	1766	0.621	0.702	0.688	0.530	0.981
(0.100,0.200]	122	1508	0.682	0.649	0.655	0.489	0.994
(0.000,0.100]	50	8348	1.000	0.335	0.502	0.335	1.000

Fonte: Elaborado pelo Autor (2019).

A primeira coisa que se pôde analisar do resultado desse algoritmo foi a matriz de confusão, composta pelos quatro primeiros valores: **True positive**, **False negative**, **False positive** e **True negative**.

A matriz foi muito útil, principalmente por dois motivos: primeiro porque os dados dela descreveram o resultado da classificação de cada registro, e segundo porque é por meio dela que obteve-se as demais métricas. O que representa cada um desses valores da matriz:

- **True positive (TP)**: indicou a quantidade de registros que foram classificados como "evadido corretamente", ou seja, a resposta do classificador foi que o aluno era evadido e o aluno realmente era evadido.
- **True negative (TN)**: indicou a quantidade de registros que foram classificados como "não evadido de maneira correta", ou seja, a resposta do classificador foi que o aluno não evadiu e o aluno realmente não evadiu.
- **False positive (FP)**: indicou a quantidade de registros que foram classificados como "evadido de maneira incorreta", ou seja, a resposta do classificador foi que o aluno evadiu, mas o aluno não evadiu.
- **False negative (FN)**: indicou a quantidade de registros que foram classificados como "não evadido de maneira incorreta", ou seja, a resposta do classificador foi que o aluno não é evadido, mas o aluno é evadido.

Por meio desses quatro valores, foram calculados os indicadores: **Accuracy**, **Precision**, **Recall** e **F1 Score**.

- **Accuracy:** foi o indicador mais simples de se calcular. Ele é simplesmente a divisão entre todos os acertos pelo total conforme cálculo a seguir:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

$$\text{Accuracy} = (5202 + 15004) / (5202 + 2570 + 3666 + 15004)$$

$$\text{Accuracy} = 20206 / 26442$$

$$\text{Accuracy} = \mathbf{0.764163}$$

- **Precision:** foi utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões positivas (incluindo as falsas). Também forneceu informação sobre falsos positivos, identificando um determinado resultado de maneira precisa. Na base de dados ela respondeu à seguinte pergunta: de todos os alunos classificados como positivos, qual percentual realmente é positivo?

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Precision} = 5202 / (5202 + 2570)$$

$$\text{Precision} = 5202 / 7772$$

$$\text{Precision} = \mathbf{0.669325}$$

- **Recall:** foi utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas (True Positives e False Negatives). Essa métrica foi bastante útil quando precisaram minimizar os falsos negativos. Essa métrica foi capaz de responder a questão: de todos os alunos que realmente são positivos, qual percentual é identificado corretamente pelo modelo?

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Recall} = 5202 / (5202 + 3666)$$

$$\text{Recall} = 5202 / 8868$$

$$\text{Recall} = \mathbf{0.586603}$$

- **F1 Score:** foi uma maneira de visualizar as métricas "Precision" e "Recall" juntas. Outra maneira de se unir as duas métricas seria simplesmente calcular a média aritmética. O problema é que existem casos que a "Precision", ou a "Recall", podem ser muito baixas, enquanto a outra permanece alta. Isso indicaria problemas na geração de falsos positivos ou negativos, conforme já visto nos tópicos anteriores. Para ajustar isso, o cálculo foi um pouco diferente, mas ainda acaba sendo uma média entre as duas métricas anteriores.

$$\text{F1 Score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{F1 Score} = 2 * 0.928 * 0.904 / (0.928 + 0.904)$$

$$\text{F1 Score} = 1.677824 / 1.832$$

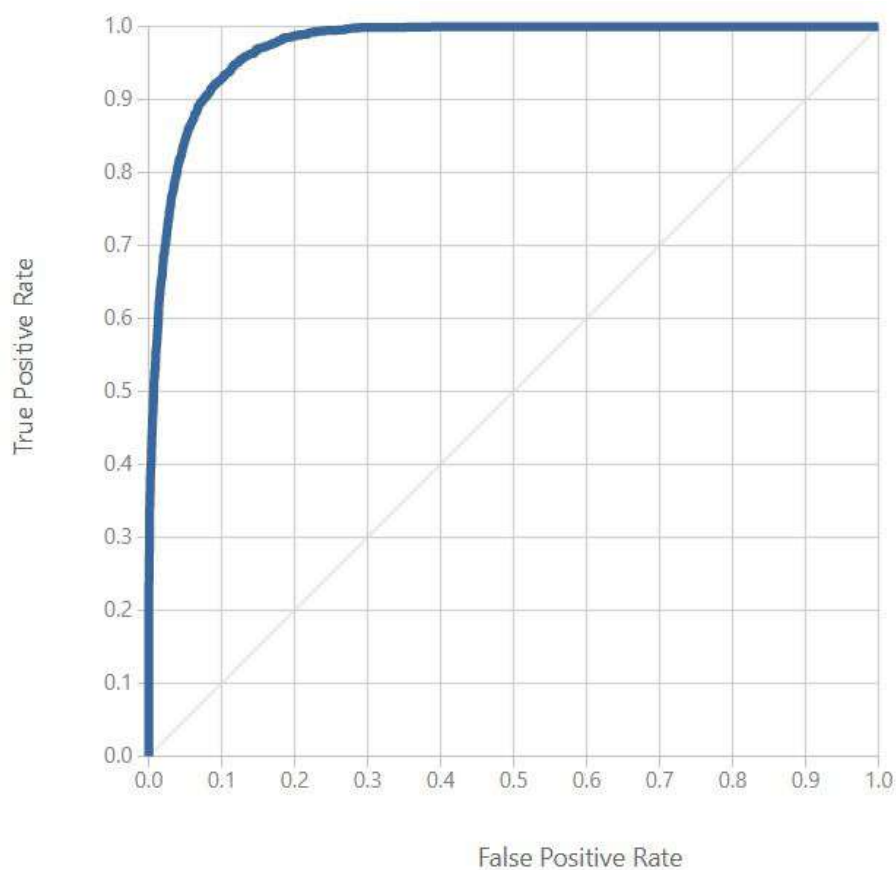
F1 Score = **0.915842**

Conforme demonstrado na Matriz de Confusão do modelo, o algoritmo classificou corretamente 5.202 registros de alunos evadidos e 15.004 registros de não evadidos, ou seja, considerando que a base de dados possui 10.000 alunos evadidos o algoritmo gerou regras para mais de 50% dos registros, o que pode ser considerado bom. Um ponto negativo foi o valor **Precision 0,669** tendo em vista que a precisão foi a proporção das instâncias classificadas corretamente, ela foi a primeira métrica a ser avaliada e o **Recall 0,587** que foi o mais baixo entre os modelos avaliados.

## 6.2 Algoritmo Two-Class logistic Regression

A regressão logística é um método bem conhecido nas estatísticas em que é usado para prever a probabilidade de um resultado e é especialmente popular para tarefas de classificação. Para a presente pesquisa, o algoritmo previu a probabilidade de ocorrência de um evento ajustando dados a uma função logística. As figuras 43, 44 e 45 apresentam os resultados desse algoritmo.


Figura 43 – Algoritmo Two-Class logistic Regression



Fonte: Elaborado pelo Autor (2019).



Figura 44 – Algoritmo Two-Class logistic Regression

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
<b>8044</b>	<b>824</b>	<b>0.915</b>	<b>0.849</b>	<b>0.5</b>		<b>0.974</b>
False Positive	True Negative	Recall	F1 Score			
<b>1433</b>	<b>16141</b>	<b>0.907</b>	<b>0.877</b>			
Positive Label	Negative Label					
<b>1</b>	<b>0</b>					

Fonte: Elaborado pelo Autor (2019).

Figura 45 – Algoritmo Two-Class logistic Regression

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall
(0.900,1.000]	1249	3	0.047	0.712	0.247	0.998	0.141
(0.800,0.900]	1606	23	0.109	0.772	0.486	0.991	0.322
(0.700,0.800]	1965	142	0.189	0.841	0.696	0.966	0.544
(0.600,0.700]	2007	403	0.280	0.901	0.839	0.923	0.770
(0.500,0.600]	1217	862	0.358	0.915	0.877	0.849	0.907
(0.400,0.500]	561	1251	0.427	0.889	0.854	0.762	0.970
(0.300,0.400]	208	1480	0.491	0.840	0.807	0.679	0.994
(0.200,0.300]	45	1242	0.539	0.795	0.766	0.621	0.999
(0.100,0.200]	3	1171	0.584	0.751	0.729	0.574	0.999
(0.000,0.100]	7	10997	1.000	0.335	0.502	0.335	1.000

Fonte: Elaborado pelo Autor (2019).

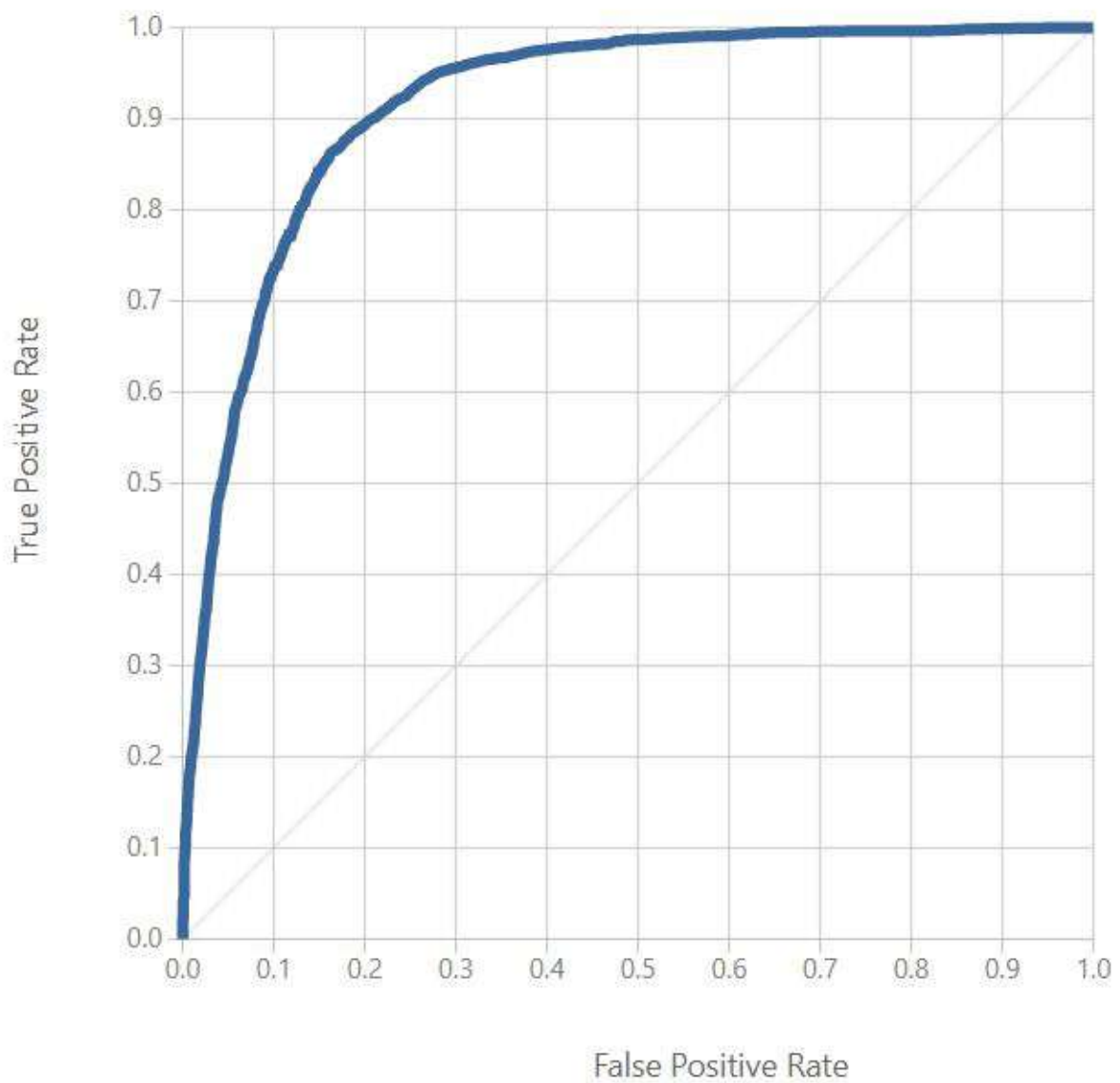
Esse algoritmo apresentou um valor baixo de **0,662** para a métrica *Precision* sendo também considerado baixo para esse estudo. Entretanto como a *Accuracy* apresentou um valor **0,777** o modelo esteve dentro dos valores considerados como bons. Além disso, foi possível inspecionar a taxa de **TP** versus a taxa de **FP** na curva **ROC** (*Receiver Operating Characteristic*) e o valor correspondente de *AUC* (*Area Under the Curve*) que nesse algoritmo *0,868*. Quanto mais próxima essa curva estiver do canto superior esquerdo, melhor estará o desempenho do classificador (que é maximizar a taxa de *TP* enquanto minimiza os **FP**).

### 6.3 Algoritmo Two-Class Locally-Deep VSM

As máquinas de vetores de suporte (SVMs) são uma classe extremamente popular e bem pesquisada de modelos de aprendizado supervisionado, que podem ser usados em

tarefas de classificação linear e não linear. Pesquisas recentes se concentraram em maneiras de otimizar esses modelos para serem escalados com eficiência para conjuntos de treinamento maiores. Nessa implementação da **Microsoft Research**, a função do **kernel** usada para mapear pontos de dados para destacar o espaço foi projetada especificamente para reduzir o tempo necessário para o treinamento, mantendo a maior parte da precisão da classificação. As figuras 46, 47 e 48 apresentam os resultados desse algoritmo.

Figura 46 – Algoritmo Two-Class Locally-Deep VSM



Fonte: Elaborado pelo Autor (2019).

Figura 47 – Algoritmo Two-Class Locally-Deep VSM

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
<b>6612</b>	<b>2256</b>	<b>0.844</b>	<b>0.780</b>	<b>0.5</b>		<b>0.917</b>
False Positive	True Negative	Recall	F1 Score			
<b>1860</b>	<b>15714</b>	<b>0.746</b>	<b>0.763</b>			
Positive Label	Negative Label					
<b>1</b>	<b>0</b>					

Fonte: Elaborado pelo Autor (2019).

Figura 48 – Algoritmo Two-Class Locally-Deep VSM

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall
(0.900,1.000]	1872	204	0.079	0.728	0.342	0.902	0.211
(0.800,0.900]	2398	481	0.187	0.800	0.618	0.862	0.482
(0.700,0.800]	1303	588	0.259	0.827	0.709	0.814	0.628
(0.600,0.700]	689	320	0.297	0.841	0.749	0.797	0.706
(0.500,0.600]	350	267	0.320	0.844	0.763	0.780	0.746
(0.400,0.500]	289	269	0.342	0.845	0.771	0.764	0.778
(0.300,0.400]	432	368	0.372	0.848	0.784	0.746	0.827
(0.200,0.300]	558	943	0.429	0.833	0.781	0.696	0.890
(0.100,0.200]	787	3967	0.608	0.713	0.696	0.540	0.979
(0.000,0.100]	190	10167	1.000	0.335	0.502	0.335	1.000

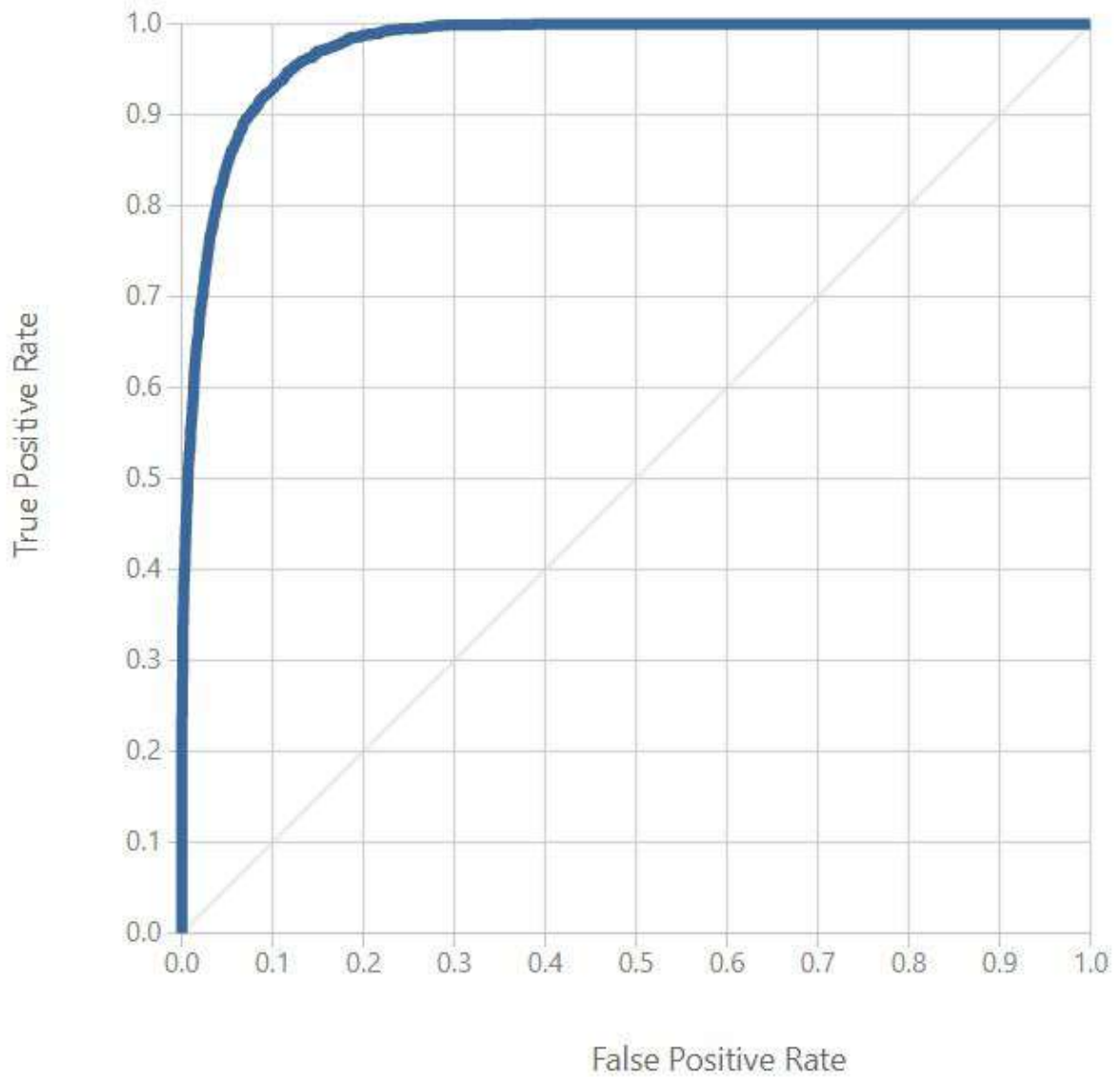
Fonte: Elaborado pelo Autor (2019).

A partir desse algoritmo os valores apresentam uma melhora significativa. Todas as métricas (*Accuracy*, *Recall*, *Precision* e *F1 Score*) atingem valores superiores a **0,70** e a **AUC** o valor de **0,917**. Os valores atingidos por esse algoritmo são considerados muito bom.

## 6.4 Algoritmo Two-Class Decision Jungle

Esse algoritmo retorna um classificador não treinado. Foi preciso treinar esse modelo em um conjunto de dados de treinamento rotulado usando o **Train Model** ou o **Tune Model Hyperparameters**. Optou-se por usar **Train Model**. As figuras 49, 50 e 51 apresentam o resultados desse algoritmo.

Figura 49 – Algoritmo Two-Class Decision Jungle



Fonte: Elaborado pelo Autor (2019).

Figura 50 – Algoritmo Two-Class Decision Jungle

True Positive	False Negative	Accuracy	Precision	Threshold	<input type="range" value="0.5"/>	AUC
<b>8044</b>	<b>824</b>	<b>0.915</b>	<b>0.849</b>	<b>0.5</b>		<b>0.974</b>
False Positive	True Negative	Recall	F1 Score			
<b>1433</b>	<b>16141</b>	<b>0.907</b>	<b>0.877</b>			
Positive Label	Negative Label					
<b>1</b>	<b>0</b>					

Fonte: Elaborado pelo Autor (2019).

Figura 51 – Algoritmo Two-Class Decision Jungle

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall
(0.900,1.000]	1249	3	0.047	0.712	0.247	0.998	0.141
(0.800,0.900]	1606	23	0.109	0.772	0.486	0.991	0.322
(0.700,0.800]	1965	142	0.189	0.841	0.696	0.966	0.544
(0.600,0.700]	2007	403	0.280	0.901	0.839	0.923	0.770
(0.500,0.600]	1217	862	0.358	0.915	0.877	0.849	0.907
(0.400,0.500]	561	1251	0.427	0.889	0.854	0.762	0.970
(0.300,0.400]	208	1480	0.491	0.840	0.807	0.679	0.994
(0.200,0.300]	45	1242	0.539	0.795	0.766	0.621	0.999
(0.100,0.200]	3	1171	0.584	0.751	0.729	0.574	0.999
(0.000,0.100]	7	10997	1.000	0.335	0.502	0.335	1.000

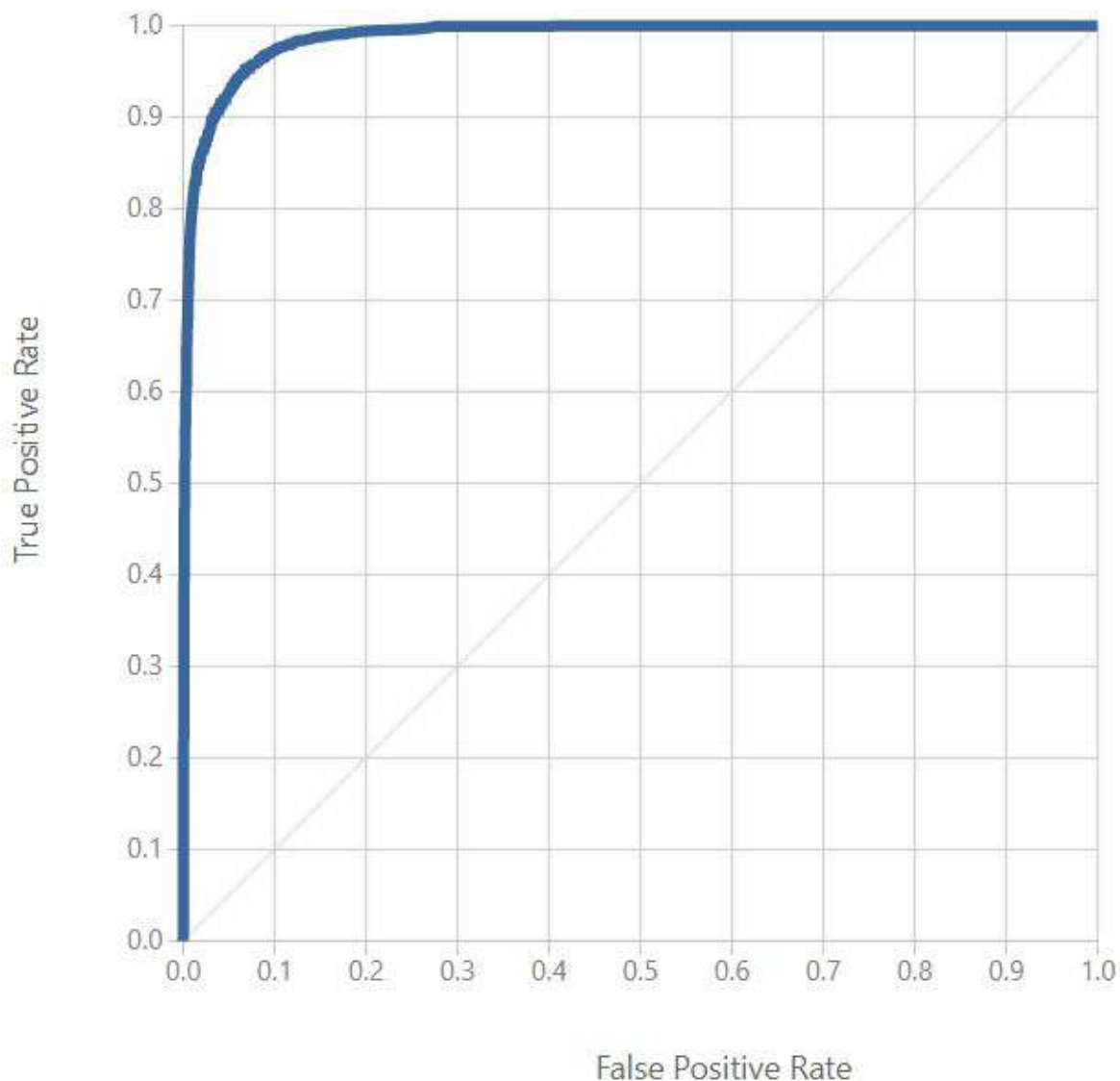
Fonte: Elaborado pelo Autor (2019).

O algoritmo alcançou **Accuracy 0,915** e uma **Precision 0,849** demonstrando um excelente desempenho. A matriz de Confusão classificou corretamente **8.044** registros de alunos evadidos e **16.141** registros de não evadidos.

## 6.5 Algoritmo Two-Class Neural Network

Esse algoritmo de rede neural de duas classes pôde criar um modelo para prever um destino de até dois valores. As figuras 52, 53 e 54 apresentam os resultados desse algoritmo.

Figura 52 – Algoritmo Two-Class Neural Network



Fonte: Elaborado pelo Autor (2019).

Figura 53 – Algoritmo Two-Class Neural Network

True Positive	False Negative	Accuracy	Precision	Threshold	<input type="text" value="0.5"/>	AUC
<b>8015</b>	<b>853</b>	<b>0.944</b>	<b>0.928</b>	<b>0.5</b>		<b>0.988</b>
False Positive	True Negative	Recall	F1 Score			
<b>626</b>	<b>16948</b>	<b>0.904</b>	<b>0.916</b>			
Positive Label	Negative Label					
<b>1</b>	<b>0</b>					

Fonte: Elaborado pelo Autor (2019).

Figura 54 – Algoritmo Two-Class Neural Network

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall
(0.900,1.000]	6341	101	0.244	0.901	0.828	0.984	0.715
(0.800,0.900]	834	87	0.278	0.929	0.884	0.974	0.809
(0.700,0.800]	343	99	0.295	0.938	0.902	0.963	0.848
(0.600,0.700]	255	171	0.311	0.941	0.909	0.944	0.877
(0.500,0.600]	242	168	0.327	0.944	0.916	0.928	0.904
(0.400,0.500]	176	214	0.342	0.943	0.915	0.907	0.924
(0.300,0.400]	194	282	0.360	0.939	0.913	0.882	0.946
(0.200,0.300]	193	451	0.384	0.930	0.902	0.845	0.967
(0.100,0.200]	117	537	0.409	0.914	0.884	0.805	0.980
(0.000,0.100]	173	15464	1.000	0.335	0.502	0.335	1.000

Fonte: Elaborado pelo Autor (2019).

Esse algoritmo foi um dos melhores. Observou-se isso no gráfico ao ser analisada a curva **AUC** que chegou a **0,988**. Todas as métricas passaram de **0,90**. Ao se analisar os valores de **FP** e **FN** na Matriz de Confusão, notou-se que atingiram valores bem baixos respectivamente **626** e **853**. Isso demonstrou que o algoritmo errou menos e por isso o resultado foi excelente.

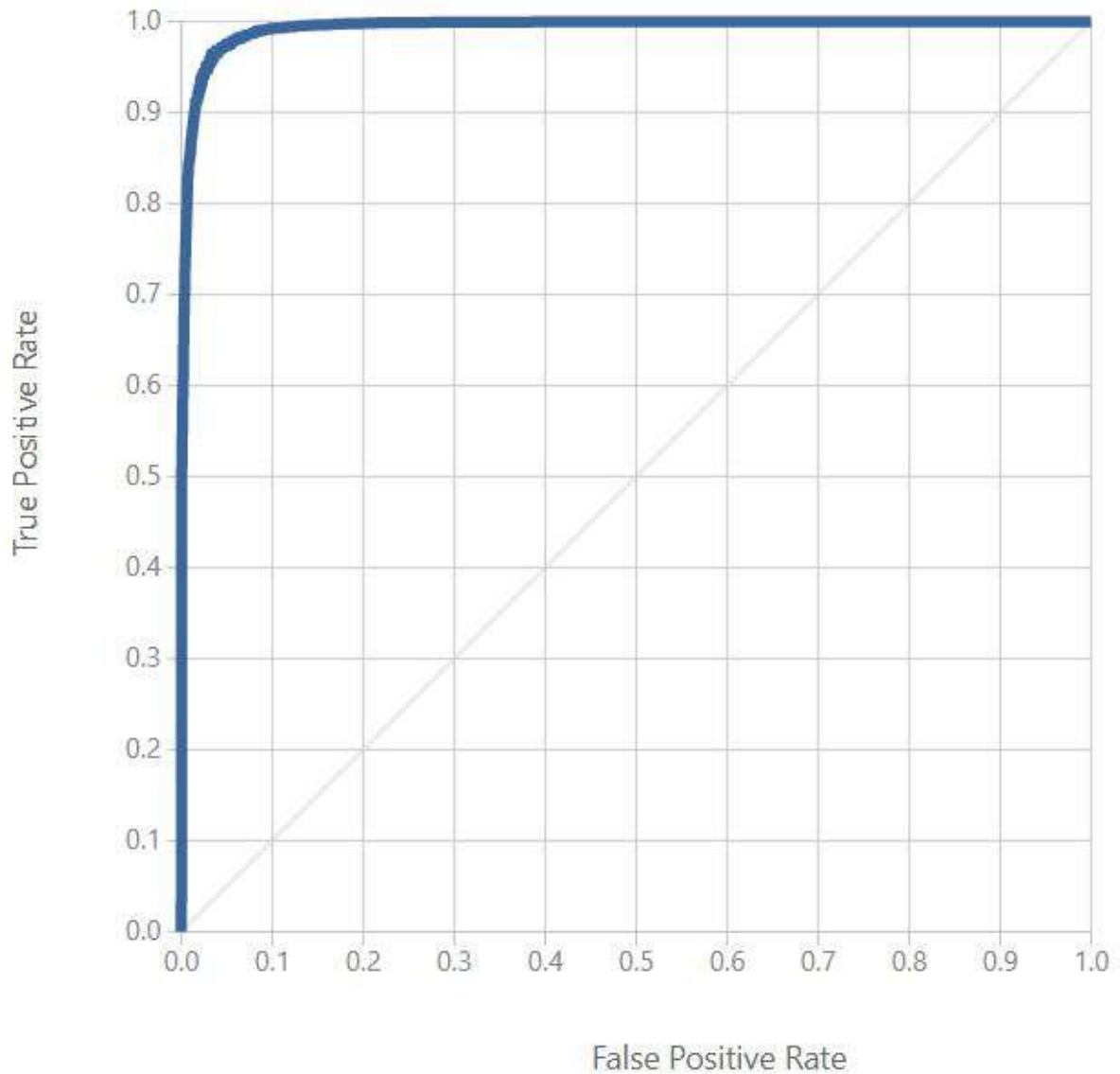
## 6.6 Algoritmo Two-Class Boosted Decision Tree

Uma Árvore de Decisão aprimorada é um método de aprendizado de conjunto no qual a segunda Árvore corrige os erros da primeira Árvore, a terceira Árvore corrige os erros da primeira e da segunda Árvores e assim por diante. As previsões são baseadas em todo o conjunto de Árvores que fazem as correções.

Geralmente, quando configuradas corretamente, as Árvores de Decisão aprimoradas são os métodos mais fáceis para se obter o melhor desempenho em uma ampla variedade de tarefas de aprendizado de máquina. No entanto, elas também são um dos métodos que mais consomem memória. As figuras 55, 56 e 57 apresentam o resultado desse algoritmo.



Figura 55 – Algoritmo Two-Class Boosted Decision Tree



Fonte: Elaborado pelo Autor (2019).

Figura 56 – Algoritmo Two-Class Boosted Decision Tree

True Positive	False Negative	Accuracy	Precision	Threshold	<input type="text" value="0.5"/>	AUC
<b>8433</b>	<b>435</b>	<b>0.964</b>	<b>0.943</b>	<b>0.5</b>		<b>0.994</b>
False Positive	True Negative	Recall	F1 Score			
<b>505</b>	<b>17069</b>	<b>0.951</b>	<b>0.947</b>			
Positive Label	Negative Label					
<b>1</b>	<b>0</b>					

Fonte: Elaborado pelo Autor (2019).



Figura 57 – Algoritmo Two-Class Boosted Decision Tree

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall
(0.900,1.000]	7322	119	0.281	0.937	0.898	0.984	0.826
(0.800,0.900]	514	100	0.305	0.953	0.926	0.973	0.884
(0.700,0.800]	275	92	0.319	0.960	0.938	0.963	0.915
(0.600,0.700]	195	88	0.329	0.964	0.945	0.954	0.937
(0.500,0.600]	127	106	0.338	0.964	0.947	0.943	0.951
(0.400,0.500]	102	101	0.346	0.964	0.948	0.934	0.962
(0.300,0.400]	68	171	0.355	0.961	0.943	0.917	0.970
(0.200,0.300]	91	278	0.369	0.954	0.934	0.892	0.980
(0.100,0.200]	89	500	0.391	0.938	0.915	0.850	0.990
(0.000,0.100]	85	16019	1.000	0.335	0.502	0.335	1.000

Fonte: Elaborado pelo Autor (2019).

Esse algoritmo foi o que apresentou melhor resultado. A **Accuracy 0,964**, a **Precision 0,943**, o **Recall 0,951**, **F1 Score 0,947** e a **AUC** atingiu **0,994**.

## 7 Considerações Finais

A educação é um dos assuntos mais importantes nas sociedades mais avançadas. Todo dinheiro viabilizado em educação nunca deveria ser compreendido como um gasto ou como um custo, mas sim como um investimento a longo prazo. Assim, também dentro dos orçamentos para a Educação pública e privada, com cada Real disponibilizado entendido como um investimento. Assim é necessário que os investimentos na educação continuem e que não haja desperdícios, como por exemplo, com a evasão escolar.

Por intermédio da literatura existente e dos levantamentos feitos, reafirma-se que a evasão pode afetar tanto os indivíduos, como a sociedade, prejudicando de forma particular o orçamento, a estrutura e a qualidade do ensino oferecido. Essa ideia é reforçada pelo Modelo Teórico de [Tinto \(1975\)](#) intitulado de "Modelo de Integração Acadêmico e Social", em que se ressaltam os danos e prejuízos causados pela evasão para a gestão universitária em particular, no entanto, mostra que os prejuízos afetam a sociedade como um todo.

Para adotar medidas capazes de minimizar os processos de evasão, as instituições educativas precisam adotar algumas medidas importantes além de conhecerem as causas que motivaram essa ocorrência e suas diferentes nuances. Existem, na literatura, diversas causas que se inter cruzam ([PRESTES; FIALHO, 2018](#)).

Combater a evasão não é uma tarefa fácil e além de atingir toda a comunidade escolar. Toda a sociedade precisa entender que quando uma instituição fica com vagas desocupadas isso representa um desperdício de recursos financeiros e um prejuízo à sociedade.

Nos trabalhos relacionados houve alguns bons resultados com o trabalho de [Márquez-Vera et al. \(2013\)](#) que explorou a evasão no ensino médio em uma cidade mexicana e atingir valores de 93.4%, 94.0% e 88.3%, respectivamente. Já o trabalho de [Hoffmann et al. \(2016\)](#) utilizou a metodologia CRISP-DM no curso de Zootecnia e alcançou uma acurácia de 98% na previsão, e mais de 70% de sucesso na previsão de alunos que abandonaram o curso.

O presente trabalho procurou dar a instituição uma nova ferramenta para combater à evasão. As tecnologias de ML estão crescendo consideravelmente nos últimos anos, principalmente no meio educacional. Nesse sentido o principal objetivo tem sido gerar conhecimento oculto nos milhares de registros dos bancos de dados de sistemas acadêmicos e que não são perceptíveis por meio de relatórios gerenciais gerados por consultas estruturadas na base de dados.

Foram traçados objetivos na presente pesquisa para utilizar a ML de forma consistente e que os conhecimentos gerados possuíssem a confiabilidade esperada pela ins-

tituição. Dentre os objetivos estabelecidos, todos foram alcançados durante o desenvolvimento da pesquisa, sendo o primeiro deles a compreensão da problemática da evasão e seus impactos. Para cumprir esse objetivo, foi realizado um levantamento bibliográfico para conhecer as possíveis causas da evasão e os modelos e teorias relacionadas à questão.

O segundo objetivo traçado foi verificar com o público alvo, os registros dos alunos e curso mais relevantes. Levantamos todas as variáveis sugeridas pelos usuários chave.

O terceiro objetivo foi identificar e mapear quais dados e informações estão presentes no banco de dados e que precisam ser considerados para elaboração do modelo. Analisaram-se todas as tabelas que poderia conter alguma informação e foi feita uma carga para um banco de dados.

O quarto objetivo foi relacionar e correlacionar as características mais relevantes para compor o banco de dados das amostras. Foi feita uma consulta SQL com todas as variáveis que foram mapeadas.

O quinto objetivo foi preparar os conjuntos de dados para treinamento e teste dos modelos. Após analisar a consulta e verificar a integridade dos dados, gerou-se um arquivo para carga no MLS.

O sexto objetivo foi treinar, testar e validar os modelos. Para cumprir esse objetivo utilizou-se a metodologia CRISP-DM para guiar todos os procedimentos realizados durante o processo de mineração e o software MLS forneceu os algoritmos utilizados e uma interface gráfica amigável e de fácil utilização.

O sétimo objetivo foi analisar os resultados obtidos com o modelo proposto. Foram testados 6 modelos. Todos apresentaram resultados aceitáveis conforme estabelecidos em nossa metodologia.

Entre os conhecimentos gerados pode-se destacar:

- É possível definir quais os alunos estão em tendência à evasão com base nos dados já existentes de registro e controle acadêmico;
- O modelo foi capaz de identificar uma *Accuracy* de 0,964 e AUC de 0.994;

Assim, a partir deste conhecimento gerado, será possível traçar estratégias de combate à evasão proporcionando uma mudança organizacional significativa na forma de compreender e combater a evasão escolar.

Sobre as limitações e desafios encontrados durante a pesquisa, destaca-se:

- Integrar os registros de todas as tabelas tendo em vista que o banco do sistema acadêmico é muito grande.

- A quantidade de registros disponíveis na base de dados final teve que ser reduzida, pois o MLS estava demorando dias para processar tendo em vista que para esse trabalho utilizou-se uma conta gratuita e seus recursos são limitados.
- Introduzir o conhecimento gerado pelo ML nos setores responsáveis pelo acompanhamento do aluno.

Com estudos futuros da presente pesquisa, espera-se utilizar outros algoritmos e outros parâmetros e comparar os resultados atuais com os índices que serão coletados no futuro com os novos algoritmos e os novos parâmetros. Espera-se também aplicar ML utilizando uma conta que não seja gratuita e assim utilizar todos os registros disponíveis em sua base de dados.

## Referências

- ALGHAMDI, M. et al. Predicting diabetes mellitus using smote and ensemble machine learning approach: The henry ford exercise testing (fit) project. *PloS one*, Public Library of Science, v. 12, n. 7, 2017. Citado na página 39.
- ALPAYDIN, E. *Introduction to Machine Learning Cambridge*. [S.l.]: MIT Press, 2004. Citado na página 33.
- BALANIUK, R. et al. Predicting evasion candidates in higher education institutions. In: SPRINGER. *International Conference on Model and Data Engineering*. [S.l.], 2011. p. 143–151. Citado na página 43.
- BARBIERI, C. *BI2–Business intelligence: Modelagem & Qualidade*. [S.l.]: Elsevier Editora, 2011. Citado na página 70.
- BERRY, M. J.; LINOFF, G. S. *Data mining techniques: for marketing, sales, and customer relationship management*. [S.l.]: John Wiley & Sons, 2004. Citado na página 39.
- BIAMONTE, J. et al. Quantum machine learning. *Nature*, Nature Publishing Group, v. 549, n. 7671, p. 195, 2017. Citado na página 17.
- BOTCHKAREV, A. Evaluating hospital case cost prediction models using azure machine learning studio. *arXiv preprint arXiv:1804.01825*, 2018. Citado na página 44.
- BOTCHKAREV, A. Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. *Available at SSRN 3177507*, 2018. Citado na página 43.
- CALDEIRA, C. *Data Warehousing: Conceitos e Modelos*. [S.l.]: Edições Sílabo, 2012. Citado na página 16.
- CHANG, C.-C.; LIN, C.-J. Libsvm: a library for support vector machines, "2001. software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. Citeseer, 2001. Citado na página 37.
- CHAPMAN, P. et al. Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, v. 16, 2000. Citado na página 60.
- CHAPPELL, D. Introducing azure machine learning. *A guide for technical professionals, sponsored by microsoft corporation*, 2015. Citado na página 59.
- CIELEN, D.; MEYSMAN, A.; ALI, M. *Introducing data science: big data, machine learning, and more, using Python tools*. [S.l.]: Manning Publications Co., 2016. Citado na página 32.
- COELHO, V. C. G. et al. Mineração de dados educacionais no ensino à distância governamental. *Conferências Ibero-Americanas WWW/Internet e Computação Aplicada. Brasília, Brasil*, p. 77–84, 2016. Citado 2 vezes nas páginas 43 e 44.

- COOPER, D. R.; SCHINDLER, P. S. *Métodos de Pesquisa em Administração-12ª Edição*. [S.l.]: McGraw Hill Brasil, 2016. Citado na página 58.
- CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 37.
- COUTINHO, C. P.; LISBÔA, E. S. Sociedade da informação, do conhecimento e da aprendizagem: desafios para educação no século xxi. *Revista de Educação*, Universidade de Lisboa. Instituto de Educação, v. 18, n. 1, p. 5–22, 2011. Citado na página 45.
- DEKKER, G. W. et al. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, ERIC, 2009. Citado na página 42.
- DRESCH, A.; LACERDA, D. P.; JÚNIOR, J. A. V. A. *Design science research: método de pesquisa para avanço da ciência e tecnologia*. [S.l.]: Bookman Editora, 2015. Citado na página 58.
- FIALHO, M. G. D. et al. A evasão escolar e a gestão universitária: o caso da universidade federal da paraíba. Universidade Federal da Paraíba, 2014. Citado na página 22.
- FRITSCH, R.; VITELLI, R.; ROCHA, C. S. Defasagem idade-série em escolas estaduais de ensino médio do rio grande do sul. *Revista Brasileira de Estudos Pedagógicos*, v. 95, n. 239, 2016. Citado na página 22.
- FU, J.-H. et al. A support vector regression-based prediction of students' school performance. In: IEEE. *2012 International Symposium on Computer, Consumer and Control*. [S.l.], 2012. p. 84–87. Citado na página 42.
- GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. [S.l.]: O'Reilly Media, 2019. Citado 2 vezes nas páginas 35 e 36.
- GIL, A. C. *Métodos e técnicas de pesquisa social*. [S.l.]: 6. ed. Editora Atlas SA, 2008. Citado na página 58.
- GIL, A. C. Como elaborar projetos de pesquisa. são paulo: Atlas, 2006. gil, antônio carlos. *Como elaborar projetos de pesquisa*, v. 5, 2010. Citado na página 58.
- HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página 34.
- HOFFMANN, G. Z. K. I. L. et al. Uma abordagem para previsão de evasão em cursos de graduação presenciais. 2016. Citado 2 vezes nas páginas 43 e 97.
- INEP. *Relatório Censo da Educação Superior 2018*. [S.l.]: Instituto Nacional de Estudos e Pesquisas, 2019. <[http://download.inep.gov.br/educacao\\_superior/censo\\_superior/documentos/2019/apresentacao\\_censo\\_superior2018.pdf](http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2019/apresentacao_censo_superior2018.pdf)>. Acessado em 15/03/2020. Citado 2 vezes nas páginas 25 e 26.
- JADRIĆ, M. et al. Student dropout analysis with application of data mining methods. *Management: journal of contemporary management issues*, Sveučilište u Splitu, Ekonomski fakultet, v. 15, n. 1, p. 31–46, 2010. Citado na página 41.

- JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado 2 vezes nas páginas 37 e 40.
- JOHANN, C. C. et al. Evasão escolar no instituto federal sul-rio-grandense: um estudo de caso no campus passo fundo. Educação, 2012. Citado na página 31.
- JÚNIOR, W. M. d. S. *Mineração em dados do ENEM para a predição do desempenho acadêmico no âmbito da Rede Federal de Educação Tecnológica*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2018. Citado na página 43.
- KARAMOUZIS, S. T. et al. An artificial neural network for predicting student graduation outcomes. In: *Proceedings of the world congress on engineering and computer science*. [S.l.: s.n.], 2008. p. 991–994. Citado na página 41.
- KAUARK, F. d. S.; MANHÃES, F. C.; MEDEIROS, C. H. Metodologia da pesquisa: um guia prático. Via Litterarum, 2010. Citado na página 58.
- KINGSFORD, C.; SALZBERG, S. L. What are decision trees? *Nature biotechnology*, Nature Publishing Group, v. 26, n. 9, p. 1011–1013, 2008. Citado na página 35.
- KÜCKELHAUS, S. d. S. G. P.; SANTOS, A. P. C. dos; LUZ, C. N. M. Evasão universitária do curso de administração da faculdade itop. *Multidebates*, v. 1, n. 1, p. 8–27, 2018. Citado na página 32.
- KUHN, M.; JOHNSON, K. *Applied predictive modeling*. [S.l.]: Springer, 2013. v. 26. Citado na página 39.
- LENCASTRE, J. A. *Educação On-line: Um estudo sobre o blended learning na formação pós-graduada a partir da experiência de desenho, desenvolvimento e implementação de um protótipo Web sobre a imagem*. Tese (Doutorado), 2009. Citado na página 45.
- LOBO, M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, v. 25, 2012. Citado na página 31.
- LOPES, B.; MUÝLDER, C. F. D.; JUDICE, V. M. M. Inteligência competitiva e o caso de um arranjo produtivo local de eletrônica brasileiro. *Gestão & Planejamento-G&P*, v. 12, n. 2, 2012. Citado na página 16.
- LYKOURENTZOU, I. et al. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, Elsevier, v. 53, n. 3, p. 950–965, 2009. Citado na página 42.
- MANHÃES, L. M. B.; CRUZ, S. M. S. da; ZIMBRÃO, G. Evaluating performance and dropouts of undergraduates using educational data mining. In: *Proceedings of the Twenty-Ninth Symposium on Applied Computing*. [S.l.: s.n.], 2014. Citado na página 41.
- MÁRQUEZ-VERA, C. et al. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, Springer, v. 38, n. 3, p. 315–330, 2013. Citado 3 vezes nas páginas 42, 44 e 97.

- MENDONÇA, A. C. de. Tratado internacional para evitar a dupla tributação e prevenir a evasão fiscal. um estudo de caso: Brasil e israel. *Revista de Direito Internacional Econômico e Tributário*, v. 2, n. 2, Jul/Dez, 2012. Citado na página 31.
- MICHALSKI, R. S. A theory and methodology of inductive learning. In: *Machine learning*. [S.l.]: Springer, 1983. p. 83–134. Citado na página 32.
- MITCHELL, T. M.; LEARNING, M. Mcgraw-hill science. *Engineering/Math*, v. 1, p. 27, 1997. Citado 2 vezes nas páginas 36 e 38.
- MITCHELL, T. M. et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, v. 45, n. 37, 2006. Citado na página 33.
- MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003. Citado na página 33.
- MOURA, D.; ZIVIANI, F.; OLIVEIRA, L. C. V. Utilização do design instrucional em curso ead: análise do ambiente virtual de aprendizagem de curso técnico à distância de uma instituição pública de ensino. *Educação & Tecnologia*, v. 21, n. 1, 2018. Citado na página 16.
- MUSTAFA, M. N. et al. Students dropout prediction for intelligent system from tertiary level in developing country. In: *IEEE. 2012 International Conference on Informatics, Electronics & Vision (ICIEV)*. [S.l.], 2012. p. 113–118. Citado na página 41.
- OLADOKUN, V. et al. Predicting students academic performance using artificial neural network: A case study of an engineering course. Akamai University, Hilo, HI, USA, 2008. Citado na página 42.
- PAL, S. Mining educational data to reduce dropout rates of engineering students. *International Journal of Information Engineering and Electronic Business*, Modern Education and Computer Science Press, v. 4, n. 2, p. 1, 2012. Citado na página 41.
- PEREIRA, P. O fantasma da evasão. *Ensino Superior*, v. 18, 2014. Citado na página 31.
- PINHEIRO, R. Evasões na universidade de Brasília causam prejuízo de R 95 mi. *Correio Braziliense*, v. 10, 2015. Citado na página 31.
- PRESTES, E. M. d. T.; FIALHO, M. G. D. Evasão na educação superior e gestão institucional: o caso da universidade federal da Paraíba. *Ensaio: Avaliação e Políticas Públicas em Educação*, SciELO Brasil, v. 26, n. 100, p. 869–889, 2018. Citado 3 vezes nas páginas 26, 27 e 97.
- RAMEZANKHANI, A. et al. Applying decision tree for identification of a low risk population for type 2 diabetes. tehran lipid and glucose study. *Diabetes research and clinical practice*, Elsevier, v. 105, n. 3, p. 391–398, 2014. Citado na página 35.
- SANTOS, O. F. et al. Involvement of hepatocyte growth factor in kidney development. *Developmental biology*, Elsevier, v. 163, n. 2, p. 525–529, 1994. Citado na página 27.
- SILVA, T. C.; ZHAO, L. *Machine learning in complex networks*. [S.l.]: Springer, 2016. v. 2016. Citado na página 33.



- SOUZA, R. G. D. et al. Previsões dentro e fora da amostra da regra de Taylor utilizando fatores comuns para o período de 2002: 02 à 2015: 04. In: ANPEC-ASSOCIAÇÃO. *Anais do XLIII Encontro Nacional de Economia [Proceedings of the 43rd Brazilian Economics Meeting]*. [S.l.], 2016. Citado na página 33.
- SPADY, W. G. Lament for the letterman: Effects of peer status and extracurricular activities on goals and achievement. *American Journal of Sociology*, University of Chicago Press, v. 75, n. 4, Part 2, p. 680–702, 1970. Citado 2 vezes nas páginas 27 e 28.
- SPADY, W. G. Dropouts from higher education: Toward an empirical model. *Interchange*, Springer, v. 2, n. 3, p. 38–62, 1971. Citado na página 27.
- TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 45, n. 1, p. 89–125, 1975. Citado 5 vezes nas páginas 22, 27, 28, 29 e 97.
- TINTO, V.; CULLEN, J. Dropout in higher education: A review and theoretical synthesis of recent research. ERIC, 1973. Citado na página 26.
- TURBAN, E.; VOLONINO, L. *Tecnologia da Informação para Gestão-: Em Busca de um Melhor Desempenho Estratégico e Operacional*. [S.l.]: Bookman Editora, 2013. Citado na página 16.
- VASCONCELOS, B. F. B. d. Poder preditivo de métodos de machine learning com processos de seleção de variáveis: uma aplicação às projeções de produto de países. 2017. Citado na página 32.
- VEEN, W.; VRAKKING, B. *Homo Zappiens: educando na era digital*. [S.l.]: Artmed Editora, 2009. Citado na página 45.
- VERGARA, S. C. Projetos e relatório de pesquisa em administração. 14. São, 2013. Citado na página 58.
- WIRTH, R.; HIPPEL, J. Crisp-dm: Towards a standard process model for data mining. In: SPRINGER-VERLAG LONDON, UK. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. [S.l.], 2000. p. 29–39. Citado na página 61.
- WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado na página 34.
- WOLFF, K. H.; DURKHEIM, E. *Emile Durkheim, 1858-1917: a collection of essays, with translations and a bibliography*. [S.l.]: The Ohio State University Press, 1960. Citado na página 27.

## ANEXO A – Variáveis relacionadas à evasão

Variáveis	Autores
Prestígio social e/ou valoração do curso / profissão escolhida	(ADACHI, 2009);(CASTRO, 2013); (GERBA, 2014);(HOFFMANN, 2016); (POLYDORO, 2000)
Disponibilidade de estágio	(CASTRO, 2013);(GERBA, 2014)
Vagas disponíveis no mercado de trabalho	(ALMEIDA, 2013);(CASTRO, 2013); (GERBA, 2014);(HOFFMANN, 2016)
Desemprego	(ALMEIDA, 2013);(LOURENÇO, 2014)
Satisfação com o curso	(ALMEIDA, 2013);(BEAN, 1980) ; (CISLAGHI, 2008);(SPADY, 1971)
Satisfação com a estrutura: Biblioteca,Laboratórios, Salas de aulas	(ALMEIDA, 2013);(BEAN, 1980) ; (CASTRO, 2013);(CISLAGHI, 2008); (FÁVERO, 2014);(GERBA, 2014); (GUEDES, 2015);(HOFFMANN, 2016); (LOURENÇO, 2014);(MARTINS, 2007); (POLYDORO, 2000);(SANTIAGO, 2015); (SPADY, 1971)
Satisfação com o corpo docente	(ALMEIDA, 2013);(BEAN, 1980) ; (CASTRO, 2013);(CISLAGHI, 2008); (FÁVERO, 2014);(GERBA, 2014); (HOFFMANN, 2016);(LOBO, 2012); (LOURENÇO, 2014);(MARTINS, 2007); (POLYDORO, 2000);(SANTIAGO, 2015); (SPADY, 1971)
Satisfação com o curso	(ALMEIDA, 2013);(BEAN, 1980) ; (CISLAGHI, 2008);(SPADY, 1971)
Satisfação com o corpo docente	(ALMEIDA, 2013);(BEAN, 1980) ; (CASTRO, 2013);(CISLAGHI, 2008); (FÁVERO, 2014);(GERBA, 2014); (HOFFMANN, 2016);(LOBO, 2012); (LOURENÇO, 2014);(MARTINS, 2007); (POLYDORO, 2000);(SANTIAGO, 2015); (SPADY, 1971)

Satisfação com metodologia de ensino e integração entre teoria e prática.	(ALMEIDA, 2013);(CASTRO, 2013); (CISLAGHI, 2008);(GERBA, 2014); (MARTINS, 2007);(SPADY, 1971)
Satisfação com os serviços administrativos	(CASTRO, 2013);(CISLAGHI, 2008); (FÁVERO, 2014);(HOFFMANN, 2016); (LOBO, 2012);(MARTINS, 2007); (SPADY, 1971)
Satisfação com os serviços de apoio acadêmico	(CASTRO, 2013);(CISLAGHI, 2008); (HOFFMANN, 2016);(LOBO, 2012); (SPADY, 1971)
Imagem institucional	(CABRERA, NORA, CASTAÑEDA, 1992); (HOFFMANN, 2016)
Estudante tem pouca possibilidade de participar das decisões da Universidade	(BEAN, 1980)
Carga elevada de aulas, conteúdo e trabalhos	(ALMEIDA, 2013); (CISLAGHI, 2008)
Clima de pressão	(ALMEIDA, 2013); (CISLAGHI, 2008)
Falta de apoio a inserção profissional	(BEAN, 1980);(GERBA, 2014)
Identidade e afinidade com a instituição	(CABRERA, NORA, CASTAÑEDA, 1992); (SPADY, 1971)
Sentimento de pertencimento na IES	(CABRERA, NORA, CASTAÑEDA, 1992)
Percepção que o currículo do curso: é flexível e integrado com mercado de trabalho	(BEAN, 1980) ;(CISLAGHI, 2008); (FÁVERO, 2014);(GERBA, 2014); (GUEDES, 2015);(MARTINS, 2007); (POLYDORO, 2000);(SANTIAGO, 2015)
Grau de integração das atividades de ensino, pesquisa e extensão	(CISLAGHI, 2008);(GERBA, 2014)
Critérios de avaliação adequados	(MARTINS, 2007)
Nota do vestibular	(ADACHI, 2009);(BRISSAC, 2009); (HOFFMANN, 2016);(VITELLI, 2013)
Nota de matemática	(ADACHI, 2009);(BRISSAC, 2009); (KAWASE, 2015);(SPADY, 1971)

Desempenho Acadêmica (nota)	(ADACHI, 2009);(ALMEIDA, 2013); (BEAN, 1980) ;(BRISSAC, 2009); (CABRERA, NORA, CASTAÑEDA, 1992); (GERBA, 2014);(HOFFMANN, 2016); (KAWASE, 2015);(LOBO, 2012); (SADLER, COHEN, KOCKSEN, 1997); (SPADY, 1971) ;(TINTO, 1997) ; (VITELLI, 2013);(POLYDORO, 2000)
% reprovação de disciplinas	(ALMEIDA, 2013);(CASTRO, 2013); (GERBA, 2014);(HOFFMANN, 2016); (MARTINS, 2007);(SANTIAGO, 2015); (SPADY, 1971) ;(TINTO, 1997); (VITELLI, 2013)
% cancelamento / trancamentos	(HOFFMANN, 2016);(POLYDORO, 2000); (VITELLI, 2013)
% conclusão do curso	(CABRERA, NORA, CASTAÑEDA, 1992); (CISLAGHI, 2008);(LOURENÇO, 2014); (VITELLI, 2013)
Frequência	(ALMEIDA, 2013);(BEAN, 1980); (HOFFMANN, 2016);(LOBO, 2012); (POLYDORO, 2000);(VITELLI, 2013)
Tempo de dedicação aos estudos	(CABRERA, NORA, CASTAÑEDA, 1992); (MARTINS, 2007);(SPADY, 1971)
Forma de ingresso no curso	(VITELLI, 2013)
Nº de disciplinas matriculados	(HOFFMANN, 2016);(VITELLI, 2013)
Interações com outros estudantes	(ADACHI, 2009);(BEAN, 1980) ; (CABRERA, NORA, CASTAÑEDA, 1992); (CASTRO, 2013);(CISLAGHI, 2008); (FÁVERO, 2014);(HOFFMANN, 2016); (MARTINS, 2007);(POLYDORO, 2000); (SANTIAGO, 2015);(SPADY, 1971); (TINTO, 1997)

Interações com o corpo docente	(ADACHI, 2009);(ALMEIDA, 2013); (BEAN, 1980) ;(CASTRO, 2013); (CISLAGHI, 2008);(FÁVERO, 2014); (GUEDES, 2015);(LOURENÇO, 2014); (POLYDORO, 2000);(SANTIAGO, 2015); (SPADY, 1971) ;(TINTO, 1997)
Interações com o coordenador	(ALMEIDA, 2013);(LOURENÇO, 2014); (MARTINS, 2007);(SANTIAGO, 2015); (SPADY, 1971) ;(TINTO, 1997)
Interações com o pessoal técnico administrativo	(BEAN, 1980) ;(FÁVERO, 2014); (POLYDORO, 2000);(SPADY, 1971); (TINTO, 1997)
Interações com projetos de pesquisa, cultural, social ou político	(ADACHI, 2009);(CISLAGHI, 2008); (POLYDORO, 2000)
Financiamento estudantil privado/público/da IES	(CABRERA, NORA, CASTAÑEDA, 1992); (ESTITE, 2005);(MARTINS, 2007); (SADLER, COHEN, KOCKSEN, 1997); (VITELLI, 2013)
Bolsas institucionais / governamentais	(CABRERA, NORA, CASTAÑEDA, 1992); (ESTITE, 2005);(HOFFMANN, 2016); (MARTINS, 2007);(VITELLI, 2013); (SADLER, COHEN, KOCKSEN, 1997)
Oportunidades Acadêmicas (estágio, bolsa científica, culturais, esportivas, monitoria remunerada, emprego e outras)	(BEAN, 1980) ;(CISLAGHI, 2008); (HOFFMANN, 2016)
Mensalidade elevada	(MARTINS, 2007)
Satisfação com o apoio financeiro	(CABRERA, NORA, CASTAÑEDA, 1992)
Inadimplência	(LOBO, 2012);(VITELLI, 2013)
Idade	(ADACHI, 2009);(ESTITE, 2005); (GUEDES, 2015);(HOFFMANN, 2016); (LOURENÇO, 2014);(MARTINHO, 2014); (SADLER, COHEN, KOCKSEN, 1997); (SANTIAGO, 2015);(SANTOS, 2014); (SPADY, 1971) ;(TINTO, 1997); (VITELLI, 2013);(MARTINS, 2007)

Sexo	(ADACHI, 2009);(BRISSAC, 2009); (ESTITE, 2005);(KAWASE, 2015); (LOURENÇO, 2014);(MARTINHO, 2014); (SADLER, COHEN, KOCKSEN, 1997); (SANTIAGO, 2015);(SANTOS, 2014); (SPADY, 1971) ;(TINTO, 1997); (VITELLI, 2013)
Raça	(BRISSAC, 2009);(HOFFMANN, 2016); (MARTINHO, 2014);(SPADY, 1971); (SADLER, COHEN, KOCKSEN, 1997)
Estado civil	(ADACHI, 2009);(BRISSAC, 2009); (ESTITE, 2005);(HOFFMANN, 2016); (KAWASE, 2015);(MARTINHO, 2014); (MARTINS, 2007);(SANTIAGO, 2015); (SANTOS, 2014);(SPADY, 1971); (VITELLI, 2013)
Religião	(SPADY, 1971)
Aberto a mudanças	(SPADY, 1971)
Escolaridade dos pais	(ADACHI, 2009);(BEAN, 1980); (BRISSAC, 2009);(ESTITE, 2005); (HOFFMANN, 2016);(MARTINHO, 2014); (MARTINS, 2007);(SANTOS, 2014); (SPADY, 1971)
Quantidade de irmãos	(SANTIAGO, 2015)
Quantidade de filhos	(ADACHI, 2009);(CISLAGHI, 2008)
Casamento ou nascimento de filhos	(ALMEIDA, 2013);(HOFFMANN, 2016); (LOURENÇO, 2014);(SANTIAGO, 2015)
Quantidade de pessoas que vivem na mesma residência	(ADACHI, 2009);(MARTINHO, 2014); (SANTIAGO, 2015)
Problemas Familiares	(GERBA, 2014);(POLYDORO, 2000); (SPADY, 1971)
Influência da família na escolha do curso	(CISLAGHI, 2008);(GERBA, 2014); (MARTINS, 2007);(SANTIAGO, 2015)
Renda familiar	(ADACHI, 2009);(BEAN, 1980); (ESTITE, 2005);(FÁVERO, 2014); (MARTINHO, 2014);(MARTINS, 2007); (SADLER, COHEN, KOCKSEN, 1997); (SANTIAGO, 2015);(SPADY, 1971)

Compromisso familiares	(ALMEIDA, 2013);(HOFFMANN, 2016)
Apoio da família para cursar ensino superior	(CABRERA, NORA, CASTAÑEDA, 1992)
Estado ou condição de maturidade/grau de certeza para optar por um curso/profissão	(ALMEIDA, 2013);(FÁVERO, 2014); (GERBA, 2014);(LOBO, 2012); (LOURENÇO, 2014);(POLYDORO, 2000); (SANTIAGO, 2015)
Afinidade/identificação com o curso	(CASTRO, 2013);(FÁVERO, 2014); (GUEDES, 2015);(MARTINS, 2007); (POLYDORO, 2000);(SANTIAGO, 2015)
Renda pessoal	(ADACHI, 2009);(BRISSAC, 2009); (ESTITE, 2005);(FÁVERO, 2014); (GUEDES, 2015);(HOFFMANN, 2016)
Conciliar estudo e trabalho	(ADACHI, 2009);(GERBA, 2014); (GUEDES, 2015);(MARTINS, 2007); (POLYDORO, 2000);(SANTIAGO, 2015)
Conciliar estudos e vida familiar	(CASTRO, 2013);(LOURENÇO, 2014)
Adaptação à vida acadêmica	(ADACHI, 2009);(ALMEIDA, 2013); (GERBA, 2014);(LOBO, 2012)
Condições de saúde / falecimento	(ALMEIDA, 2013);(CASTRO, 2013); (LOURENÇO, 2014);(MARTINS, 2007); (POLYDORO, 2000);(SANTIAGO, 2015)
Situação financeira	(ADACHI, 2009);(ALMEIDA, 2013); (CASTRO, 2013);(CISLAGHI, 2008); (GUEDES, 2015);(LOURENÇO, 2014); (MARTINS, 2007);(POLYDORO, 2000); (SANTIAGO, 2015)
Tempo de dedicação aos estudos	(ADACHI, 2009);(CISLAGHI, 2008); (GUEDES, 2015);(MARTINS, 2007); (POLYDORO, 2000)
Dificuldade de acompanhamento do curso	(ALMEIDA, 2013);(CASTRO, 2013); (LOURENÇO, 2014);(MARTINS, 2007); (POLYDORO, 2000)
Exerce atividade remunerada	(ADACHI, 2009);(CASTRO, 2013); (ESTITE, 2005);(GUEDES, 2015); (HOFFMANN, 2016);(MARTINS, 2007); (SANTIAGO, 2015);(SPADY, 1971)

Apoio no emprego atual	(ALMEIDA, 2013);(CISLAGHI, 2008); (POLYDORO, 2000)
Localização da residência (distância da instituição de ensino)	(ADACHI, 2009);(BEAN, 1980); (BRISSAC, 2009);(CASTRO, 2013); (GUEDES, 2015);(HOFFMANN, 2016); (KAWASE, 2015);(LOBO, 2012); (LOURENÇO, 2014);(MARTINHO, 2014); (MARTINS, 2007);(POLYDORO, 2000); (SADLER, COHEN, KOCKSEN, 1997); (SANTIAGO, 2015);(SANTOS, 2014)
Localização do trabalho (distância da instituição de ensino)	(GUEDES, 2015);(LOBO, 2012); (LOURENÇO, 2014);(POLYDORO, 2000)
Mudanças de endereço	(ALMEIDA, 2013);(GERBA, 2014); (HOFFMANN, 2016);(LOBO, 2012); (LOURENÇO, 2014);(MARTINS, 2007); (POLYDORO, 2000)
Tamanho da comunidade natal	(BEAN, 1980)
Localização da IES	(MARTINS, 2007)
Tipo de estabelecimento de ensino (Privado ou Público)	(ADACHI, 2009);(BRISSAC, 2009); (CASTRO, 2013);(ESTITE, 2005); (HOFFMANN, 2016);(KAWASE, 2015); (MARTINHO, 2014);(SANTIAGO, 2015); (SANTOS, 2014);(VITELLI, 2013)
Desempenho no ensino médio	(ADACHI, 2009);(BEAN, 1980); (BRISSAC, 2009);(SPADY, 1971) (CABRERA, NORA, CASTAÑEDA, 1992); (HOFFMANN, 2016); (SADLER, COHEN, KOCKSEN, 1997);
Realizou cursinho pré-vestibular	(ADACHI, 2009);(BRISSAC, 2009)
Participou de vestibulares anteriormente	(ADACHI, 2009);(BRISSAC, 2009); (ESTITE, 2005);(HOFFMANN, 2016); (MARTINS, 2007)
Histórico de evasão em níveis educacionais prévios	(BRISSAC, 2009);(SANTIAGO, 2015)
Preparo na educação básica	(ALMEIDA, 2013);(CISLAGHI, 2008); (LOBO, 2012);(MARTINS, 2007); (POLYDORO, 2000)



Tempo em anos ou períodos acadêmicos entre a conclusão do Ensino Médio e o ingresso na educação superior	(KAWASE, 2015)
Experiências prévias na educação superior	(ADACHI, 2009); (BRISSAC, 2009)
Histórico de evasão na educação superior	(SANTIAGO, 2015)
Motivo para escolha do curso / carreira	(ADACHI, 2009);(CASTRO, 2013); (GUEDES, 2015);(HOFFMANN, 2016); (VITELLI, 2013)
Perspectivas profissionais	(ADACHI, 2009)
Mudança de interesse, opção ou priorização de outro curso	(ALMEIDA, 2013);(FÁVERO, 2014); (GERBA, 2014);(MARTINS, 2007); (SANTIAGO, 2015)
Expectativas em relação ao ensino superior	(ALMEIDA, 2013);(CASTRO, 2013); (HOFFMANN, 2016);(MARTINS, 2007); (SPADY, 1971)
Entendimento que o curso agrega valor ao estudante	(CABRERA, NORA, CASTAÑEDA, 1992); (POLYDORO, 2000);(BEAN, 1980)
Intenção de persistir no curso	(CABRERA, NORA, CASTAÑEDA, 1992); (CISLAGHI, 2008)
Compromisso em gradua-se	(CABRERA, NORA, CASTAÑEDA, 1992); (CASTRO, 2013);(CISLAGHI, 2008); (POLYDORO, 2000)
Grau de segurança quanto a conseguir ser um profissional conceituado	(ALMEIDA, 2013)

# ANEXO B – Pedido de autorização de acesso a dados

## Termo de Autorização – Dissertação

Belo Horizonte, 30 de agosto 2019

A Gerência de Tecnologia da Informação (GTI) da empresa SENAC em Minas , CNPJ 03.447.242/0001-16, Inscrição Estadual 062.896871.00-65 localizada no endereço Rua dos Tupinambás, nº 1086, Centro, CEP 30120-070, Belo Horizonte, Minas Gerais, Telefone: (31) 30485123, autoriza a produção intelectual baseada nas informações do Sistema Acadêmico (SA) para a dissertação do aluno Alex Marques de Souza, a título de pesquisa científica. Serão utilizados somente os dados disponibilizados pela GTI, não sendo utilizada / divulgada nenhuma informação explícita sobre aluno, curso ou qualquer outro dado da instituição, apenas códigos e números serão utilizados como conjunto de dados para treinamento dos modelos de *Machine Learning*.

O Aluno, de código de matrícula 5A231004643, cursa o Programa de Pós-Graduação: Mestrado em Sistema de Informação e Gestão do Conhecimento na Universidade FUMEC, com a orientação do Prof. Dr. Luiz Cláudio Gomes Maia, localizada Rua Cobre, 200 - Cruzeiro, CEP 30310-190, Belo Horizonte, Minas Gerais, Telefone (31) 0800 0300 200. As informações utilizadas serão utilizadas para estudo de técnicas de *Machine Learning*.



Cristiano Mascarenhas da Silva

Gerente de Tecnologia da Informação (Interino)

Cristiano Mascarenhas da Silva  
Coordenador de TI  
SENAC MINAS

## ANEXO C – Relação das tabelas utilizadas no estudo

Segue a relação das tabelas que foram copiadas para o banco de dados "DataMartEvasao".

Relação de tabelas relacionadas ao tema Curso	
TBCURSOCUSTOPLAN	TBCLASSIFICACAO AUTORIZACAOCURSO
TBSEQUENCIALCURSO	TBCENSO_CURSO
TBCONTRATODESCONTO TIPOCURSO	TBGRUPOCURSODETALHE
TNQUADROAVISOTIPOCURSO	TFCURSOS
TBAGRUPAMENTOCURSO COMERCIALCURSO	TBCLASSIFICACAOTIPOCURSO
TBCATEGORIACURSO	TBEIXOCURSO
TBTURNOCURSO	TBTIPOCURSO
TBFORMAVALIDACAOCURSO	TBGRUPOCURSO
TBMETENSCURSO	TBPROCESSOSELETIVO CURSOTURNO
TFMUDANCACURSO	TBTIPOOFERTACURSO
TNGRUPOCONTRATOCURSO	TBTIPOCORRENCIACURSO
TBSTATUSCURSO	TBCURSO
TBTIPOESTRUTURACURSO	TBINFORMACAOCURSO
TFSITUACAOCURSO	TBAGRUPAMENTOCURSO COMERCIAL
TBTIPOMATERIALCURSO	
TBMDICURSO	

Relação de tabelas relacionadas ao tema Turma	
TBTURMAGRADEESTAGIO	TBTIPOESTRUTURATURMA
TBTIPOHISTORICOTURMA	TBREQUERIMENTOTURMA ALUNODISCIPLINA
TBTURMADISCIPLINASITE	TBTURMAPROVA
TBRESERVATURMA MENORAPRENDIZ	TBSEQUENCIALTURMAGRADE

TBDESCONTOTURMAGRADE	TBSEQUENCIALTURMA
TBTURMAGRADEALUNO	TBSTATUSTURMA
TBDIASEMANATURMAGRADE	TBPROCESSOBOLSAACORDO TURMAGRADE
TBREQUERIMENTOTURMAALUNO	TLAGRUPAMALADIRETA TURMAEVENTO
TBSITUACAOTURMAANTERIOR	TBTURMA
TBTRANCAMENTOTURMA	TBACAODIASEMANA TURMAGRADE
TBPLANOTRABALHODOCENTE TURMAHISTORICO	TURMAALUNO
TBPROGRAMACAOTURMA	TBREQUERIMENTOTURMA
TBTURMAITEMCUSTO	TLVINCULOTURMACONTRATO
TBCANCELAMENTOTURMA	TBDESCONTOTURMACOMERCIAL
PR_TURMA	TBTURMACOMERCIAL
TBOCUPACAOTURMADIA INSERIDOS	TBTURMAALUNODISCIPLINA FREQUENCIA
TNTURMAORCAMENTO	TBTIPOTURMACOPIA
TBCOBRANCATURMA	CRITICATURMA
TBPACOTETURMAGRADE	TBDETALHAMENTOTIPO HISTORICOTURMA
TBTURMAALUNOCONTRATO	TBORIGEMTURMA
TBTURMAGRADESEQUENCIA	TBHISTORICOTURMA
TBGRUPOSELECAOTURMAGRADE	TBGRATUIDADERESERVATURMA
TBPLANOACAOTURMA	TBTURMADOCENTE
TBFECHAMENTOTURMA	TBAUXCGUTURMASMANHUACU
TBTURMADISCIPLINA EQUIPAMENTO	TBSITUACAOTURMA
TBCOMPROVACAOALUNOTURMA	TBOCUPACAOTURMADIA
TBSITUACAOTURMAALUNO DISCIPLINA	TBTIPODOCENTETURMA
TBTIPOTURMA	TBDOCENTETURMA
TBTURMAALUNODISCIPLINA HISTORICO	TBCUSTOVIAGEMTURMA
TBTURMADISCIPLINA COMPETENCIA	TBDESCONTOTURMA
TBTURMAALUNODISCIPLINA	TBTURMAGRADE

TBTURMACURSOPENDENCIA PRODUCAODN	TBLOCALREALIZACAO TURMAPSG
TBCONTRATOTURMAALUNO	TBTURMADISCIPLINA
TBTURMAGRADEESTAGIOALUNO	TBTURMAALUNO
TBTURMAALUNODISCIPLINA COMPETENCIA	TBTIPOTURMADISCIPLINA
TBTURMASITE	TBPLANOTRABALHO DOCENTETURMA
TBTURMAGRADEESTAGIOALUNO EMPRESA	

Relação de tabelas relacionadas ao tema Aluno	
TB_CONTROLE_REQUISITO _DOCUMENTO_ALUNO	TBALUNOTRANSFSUGERIDO
TBMOTIVOSITUACAOSTATUSALUNO	TBDESCONTOALUNO INTERESSADO
TB_CONTROLE_REQUISITO _DOCUMENTO_ALUNO_HISTORICO	TBTRANCAMENTOCURSOALUNO
TBALUNOAR	TBBENEFICIOCONTRATOALUNO
TBNOTIFICACAOALUNOGRADE	TBSITUACAOCURSO ALUNOENADE
TBALUNOCOLACAOGRAU	TBDESCONTOALUNO
TBDISCIPLINAALUNO	TBAVALIACAOATIVIDADEALUNO
TBREQUERIMENTOTURMAALUNO	TBALUNOMATRICULAONLINE
TBDISCONTRMATRTURGRADALUNO	TMPCURSOALUNOALTERADO
TNSITUACAOMETALUNO	TBATENDIMENTONOTIFICACAO ALUNO
TBCRACHAACESSOALUNO	TBALUNOCONTRATO
TBSITUACAOTITULACAOALUNO	TBTIPOHISTORICOALUNO
TBCURSOALUNOREF	TBNOTIFICACAOALUNO
TBREALIZACAOATIVIDADE CURSOALUNO	TBCURRICULOALUNO
TBGRADECURRICULARALUNO	TBAVALIACAOINDICADORALUNO
TBCURSOALUNO	TB_CAMPO_ALUNO_REQUISITO
TNTITULACAOALUNO	TBAVALIACAOINDICADOR ALUNODETALHE
TBALUNOREMATRICULA	TBCONTROLEREQUISITO INSCRICAOALUNO

TLCONTROLEDOCUMENTOALUNO	TFDESCONTOALUNO
TL_CONTROLE_REQUISITO_DOCUMENTO_ALUNO	TBCONTROLEDOCUMENTOALUNO
TNALUNOFINANCEIRO	TBALUNOPLANEJADOMENSAL
TLREQUISITOINSCRICAOALUNO	TBALUNO
TBSTATUSCURSOALUNO	TBALUNONECESSIDADEESPECIAL
TBHISTORICOCOMPROVACAOALUNOTURMA	TBAVALIACAOINDICADORALUNOHISTORICO
TBATENDIMENTOALUNO	TBCENSOREGISTROALUNO
TBAVALIACAOATIVIDADEALUNOHISTORICO	TBHISTORICOALTERACAOCONTATOSALUNO
TBALUNOCONTRATADO	TBCOBRANCAALUNO
TBSOLICITACAOCANDIDATOALUNO	TBNECESSIDADEESPECIALALUNO
	TBATENDIMENTOALUNO

Relação de tabelas relacionadas ao tema Financeiro	
TNPLANOFINANCEIRO	TNSITUACAOFINANCEIRO
TBREFINANCIAMENTO	TBTIPORESPONSAVELFINANCEIRO
TBREMESSAFINANCEIROSIGA	TNRETORNOPLANOFINANCEIRO
TBREFINANCIAMENTOMENSALIDADE	TBCONTRATOFINANCEIRO
TBPARCELASREFINANCIAMENTO	TLFINANCIAMENTOPOLITICASDESCONTO
TBREMESSAFINANCEIROSIGAPARCELA	TBREFINANCIAMENTORECIBO
TNRESULTADOATENDIMENTOFINANCEIRO	TBRESPFINANCCONTRATOMATRICULA
TNALUNOFINANCEIRO	TBRESPONSAVELFINANCEIRO

## ANEXO D – Relação dos atributos gerados para aplicação dos modelos

<b>Atributo</b>	<b>001</b>
Tabela Origem	tbAluno
Nome de Origem	idSexo
Nome Atual	Sexo
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	002
Valores e quantidade por instâncias	M(934530), F(475508)
<b>Atributo</b>	<b>002</b>
Tabela Origem	tbaluno
Nome de Origem	idCidade
Nome Atual	CidadeAluno
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	1.234
Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	<b>003</b>
Tabela Origem	tbAluno
Nome de Origem	idEstadoCivil
Nome Atual	EstadoCivil
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	006
Valores e quantidade por instâncias	1(1044057), 3(313565), 5(35817), 7(4109), 10(9909), 11(2581)
<b>Atributo</b>	<b>004</b>
Tabela Origem	tbAluno
Nome de Origem	idTitulacao
Nome Atual	Titulacao

<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	008
<b>Valores e quantidade por instâncias</b>	1(383), 2(1799), 3(22892), 4(398246), 5(587), 6(75212), 7(910466), 8(453)
<b>Atributo</b>	
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	desCorRaca
<b>Nome Atual</b>	CorRaca
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	006
<b>Valores e quantidade por instâncias</b>	0(257825), 1(516964), 2(122064), 3(491180), 4(18327), 5(3678)
<b>Atributo</b>	
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	idocupacao
<b>Nome Atual</b>	Ocupacao
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	019
<b>Valores e quantidade por instâncias</b>	0(176612), 1(251074), 2(517413), 3(132939), 4(4546), 5(11619), 6(1046), 7(31147), 8(184949), 9(14734),10(5924),11(47569), 12(1163), 13(8476), 14(182), 15(784), 16(1293), 17(252), 18(18316)
<b>Atributo</b>	
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	idEmpregado
<b>Nome Atual</b>	Empregado
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	002



<b>Valores e quantidade por instâncias</b>	0(674248), 1(735790)
<b>Atributo</b>	<b>008</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	idFaixaRendaMensal
<b>Nome Atual</b>	FaixaRendaMensal
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	006
<b>Valores e quantidade por instâncias</b>	0(763676), 1(121541), 2(67408), 3(303260), 4(110711), 5(43442)
<b>Atributo</b>	<b>009</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	desEscolaFundamental
<b>Nome Atual</b>	EscolaFundamental
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	002
<b>Valores e quantidade por instâncias</b>	1(1315663), 2(94375)
<b>Atributo</b>	<b>010</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	desEscolaMedio
<b>Nome Atual</b>	EscolaMedio
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	002
<b>Valores e quantidade por instâncias</b>	1(1288693), 2(121345)
<b>Atributo</b>	<b>011</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	desTurno
<b>Nome Atual</b>	Turno
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	006

<b>Valores e quantidade por instâncias</b>	1(285434), 2(276351), 3(748562), 4(87561), 5(7397), 6(4733)
<b>Atributo</b>	
<b>Atributo</b>	<b>012</b>
<b>Tabela Origem</b>	tbTurmaGrade
<b>Nome de Origem</b>	vlrTurmaGrade
<b>Nome Atual</b>	ValorCurso
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	1226
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>013</b>
<b>Tabela Origem</b>	tbTurmaGrade
<b>Nome de Origem</b>	CodCurso
<b>Nome Atual</b>	Curso
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	883
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>014</b>
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	datInicioTurma
<b>Nome Atual</b>	InicioTurma
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	2.381
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>015</b>
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	datFimTurma
<b>Nome Atual</b>	FimTurma
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	2.386

Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	<b>016</b>
Tabela Origem	tbTurma
Nome de Origem	qtdCargaHoraria
Nome Atual	CargaHoraria
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	109
Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	<b>017</b>
Tabela Origem	tbTurmaGrade
Nome de Origem	chvUnidadeNegocio
Nome Atual	UnidadeNegocio
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	044
Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	<b>018</b>
Tabela Origem	tbTurmaGrade
Nome de Origem	idCidade
Nome Atual	CidadeEscola
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	1234
Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	<b>019</b>
Tabela Origem	tbTurmaGrade
Nome de Origem	desRegiao
Nome Atual	RegiaoEscola
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	005

<b>Valores e quantidade por instâncias</b>	3(217847), 4(201019), 8(143772), 10(269699), 12(577701)
<b>Atributo</b>	<b>020</b>
<b>Tabela Origem</b>	tbTurmaGrade
<b>Nome de Origem</b>	idSituacaoTurmaAluno
<b>Nome Atual</b>	SituacaoAluno
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	017
<b>Valores e quantidade por instâncias</b>	1(133454), 2(2852), 3(14), 5(63359), 6(822867), 7(246), 9(44161),10(872), 11(15866), 12(78367), 13(216244), 14(29071), 16(62), 17(1442), 23(90), 24(1051), 25(20)
<b>Atributo</b>	<b>021</b>
<b>Tabela Origem</b>	tbMatricula
<b>Nome de Origem</b>	datMatricula
<b>Nome Atual</b>	Matricula
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	1.410.038
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	<b>022</b>
<b>Tabela Origem</b>	tbTurmaGrade
<b>Nome de Origem</b>	idIngresso
<b>Nome Atual</b>	Ingresso
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	003
<b>Valores e quantidade por instâncias</b>	5.000
<b>Atributo</b>	<b>023</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	idNaturid

<b>Nome Atual</b>	Naturalidade
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	009
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo 024</b>	
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	idEstado
<b>Nome Atual</b>	Estadado
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	07
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo 025</b>	
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	flgCanhoto
<b>Nome Atual</b>	Canhoto
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	001
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo 026</b>	
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	datNasc
<b>Nome Atual</b>	Nascimento
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	20.612
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo 027</b>	
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	codTurma
<b>Nome Atual</b>	Turma

<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	27.018
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>028</b>
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	codsigla
<b>Nome Atual</b>	Sigla
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	27.018
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>029</b>
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	nomTurma
<b>Nome Atual</b>	Turma
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	883
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>030</b>
<b>Tabela Origem</b>	tbTurmaAluno
<b>Nome de Origem</b>	flgResponsavelFinanceiro
<b>Nome Atual</b>	ResponsavelFinanceiro
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	001
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>031</b>
<b>Tabela Origem</b>	tbTurma
<b>Nome de Origem</b>	qtdVagas
<b>Nome Atual</b>	Vagas
<b>Tipo</b>	Nominal

Ausência(Missing)	0
Distintos(Distinct)	003
Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>032</b>
Tabela Origem	tbTurma
Nome de Origem	chvSala
Nome Atual	Sala
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	245
Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>033</b>
Tabela Origem	tbTurmaGrade
Nome de Origem	codTurmaGrade
Nome Atual	TurmaGrade
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	883
Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>034</b>
Tabela Origem	tbTurma
Nome de Origem	vlrTurma
Nome Atual	ValorTurma
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	1226
Valores e quantidade por instâncias	1.410.038
<b>Atributo</b>	
<b>Atributo</b>	<b>035</b>
Tabela Origem	tbAluno
Nome de Origem	desprofissao
Nome Atual	Profissao
Tipo	Nominal
Ausência(Missing)	0

Distintos(Distinct)	002
Valores e quantidade por instâncias	1.410.038
<b>Atributo 036</b>	
Tabela Origem	tbturma
Nome de Origem	idturma
Nome Atual	Turma
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	27.018
Valores e quantidade por instâncias	1.410.038
<b>Atributo 037</b>	
Tabela Origem	tbAluno
Nome de Origem	desCargo
Nome Atual	Cargo
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	002
Valores e quantidade por instâncias	1.410.038
<b>Atributo 038</b>	
Tabela Origem	tbTurmaAlunoDisciplina
Nome de Origem	idFrequencia
Nome Atual	Frequencia
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	9.119
Valores e quantidade por instâncias	1.410.038
<b>Atributo 039</b>	
Tabela Origem	tbdisciplina
Nome de Origem	idDisciplina
Nome Atual	Disciplina
Tipo	Nominal
Ausência(Missing)	0
Distintos(Distinct)	2.030



<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	<b>040</b>
<b>Tabela Origem</b>	tbAluno
<b>Nome de Origem</b>	idaluno
<b>Nome Atual</b>	aluno
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	331.637
<b>Valores e quantidade por instâncias</b>	1.410.038
<b>Atributo</b>	<b>041</b>
<b>Tabela Origem</b>	tbFinanceiro
<b>Nome de Origem</b>	idStatusFinanceiro
<b>Nome Atual</b>	StatusFinanceiro
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	009
<b>Valores e quantidade por instâncias</b>	0(20089),1(1328278),2(9), 4(615),6(40318),8(19387), 9(798),10(488),13(56)
<b>Atributo</b>	<b>042</b>
<b>Tabela Origem</b>	tbMensalidade
<b>Nome de Origem</b>	idTipoBaixa
<b>Nome Atual</b>	TipoBaixa
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0
<b>Distintos(Distinct)</b>	002
<b>Valores e quantidade por instâncias</b>	M(1500), F(1850)
<b>Atributo</b>	<b>043</b>
<b>Tabela Origem</b>	tbTurmaGrade
<b>Nome de Origem</b>	chvModalidadeEnsino
<b>Nome Atual</b>	ModalidadeEnsino
<b>Tipo</b>	Nominal
<b>Ausência(Missing)</b>	0

---

<b>Distintos(Distinct)</b>	003
<b>Valores e quantidade por instâncias</b>	1(551988), 2(486154), 3(371896)

## ANEXO E – Relação dos atributos excluídos

Nome do Atributo	Motivo
DatMatricula	<p>Observa-se que esse atributo não tinha muito a acrescentar na construção do dataset visto que na instituição existe um período de matrícula e que a maioria dos alunos fazem sua matrícula nesse período. Além disso, no caso dos cursos técnicos e superior essa matrícula ocorre no início e no meio do ano. Ficou decidido assim por não utilizar esse atributo.</p>
idIngresso	<p>Esse atributo identifica a forma como o aluno ingressou na instituição. Atualmente esse ingresso pode ser por: processo seletivo, Enem ou transferência. Ao analisar esse atributo percebeu-se que apenas 5.000 registros têm informação preenchida. Optou-se por remover esse atributo do nosso dataset visto que praticamente todos os registros estão vazios(nulos).</p>
datInicioTurma	<p>Esse atributo informa a data de início da turma. Não foi possível, pela quantidade de registros diferentes, utilizar esse atributo pois ele não trouxe informações relevantes nos primeiros testes que fizemos. Entretanto será feita uma alteração nesse atributo gerando assim 2 novos atributos.</p>
DatNasc	<p>Esse atributo contém a data de nascimento do aluno. Não foi possível, pela quantidade de registros diferentes, utilizar esse atributo pois ele não trouxe informações relevantes para os primeiros testes que foram feitos. Entretanto será feita uma alteração nesse atributo gerando assim 2 novos atributos.</p>

idNaturid	Esse atributo informa a naturalidade do aluno. Infelizmente a grande maioria dos registros não estavam preenchidos. Como isso poderia confundir os modelos, foi decidido por não utilizar.
idestado	Analisando esse atributo optou-se por não utiliza-lo visto que 95% dos dados preenchidos são de alunos do estado de Minas Gerais. Entende-se que essa informação não seria relevante nesse momento.
flgCanhoto	Esse atributo diz se o aluno é canhoto. Todo os registros estavam vazios (nulos).
codturma	Esse atributo é um código de 11 números composto por 3 informações: os 3 primeiros número representam a unidade. os 4 próximos números, o ano da turma e os 3 últimos são um número auto incremento. Como o código do curso é um atributo mais forte optou-se por remover esse atributo.
sigla	Ao ser criado uma turma o SA gera uma sigla. Não será utilizado esse atributo pelo mesmo motivo do atributo codturma.
turma	Esse atributo contém o nome do curso. Como já utilizou-se o código do curso, seria uma informação redundante.
flgResponsavel Financeiro	Todos os registros estavam com a opção 1. Assim não faz sentido incluir na geração do modelo uma informação que é comum.
qtdVagas	Esse atributo contem a informação referente a quantidade de vagas do curso. A instituição mantém um padrão de vagas por modalidade. Assim como possuímos e modalidades, ou seja, FIC, Técnicos e Superior, podemos ter apenas 3 registros diferentes. Não achamos essa informação relevante.

chvsala	Esse atributo informa a sala dentro da unidade. Como a instituição possui várias unidades no estado de minas gerais esses códigos se repetem em cada unidade. Já temos uma informação mais relevante que é unidade de negócio e a região. Assim optamos por não utilizar esse atributo.
codturmagrade	Esse atributo é o mesmo do codturma que é replicado em outra tabela. Assim pelo mesmo motivo não utilizaremos.
nomCurso	Nesse Atributo temos o nome do curso. Como já estamos utilizando o código do curso, decidimos não utilizar esse atributo.
vlrTurma	Esse atributo nos informa o Valor do curso. Como já estamos utilizando o atributo vlrTurmagrade, não utilizaremos esse atributo pois os valores são iguais.
datFimTurma	Esse atributo nos informa a data de termino da turma. Não conseguimos, pela quantidade de registros diferentes, utilizar esse atributo pois ele não trouxe informações relevantes nos primeiros testes que fizemos. Entretanto ao faremos uma alteração nesse atributo gerando assim 2 novos atributos
desProfissao	Esse atributo nos informa a profissão do aluno. Resolvemos não utilizar esse atributo por que mais de 65% dos registros estavam vazios.
idEmprego	Esse atributo nos informa se o aluno está empregado. Resolvemos não utilizar por que já estamos utilizando o atributo idocupacao. Além disso 85% dos registros não estão preenchidos.
desCargo	Esse atributo nos informa o cargo do aluno. Resolvemos não utilizar esse atributo por que mais de 90% dos registros estavam vazios.

## ANEXO F – Breve descrição do Currículo do autor

Possui graduação em Ciência da Computação em 2003 pela Pontificia Universidade Católica de Minas Gerais (PUC - Betim) e especialização em Banco de Dados e BI pela Universidade Newton Paiva. Possui experiência de mais de 15 anos em administração de banco de dados e BI, já tendo trabalhado como programador, analista de sistemas, DBA e consultor em BI. Já trabalhou com as seguintes ferramentas: Delphi, .Net. Tem experiência com os seguintes sistemas de banco de dados: MS SQL Server, Oracle, Sybase, DB2 e PostgreSQL e com as seguintes ferramentas de BI: SSIS, SSAS, SSRS e Power BI, Tableau, Click View e Cognos. Possui várias certificações na área de Banco de Dados e BI. Possui experiência em especial com sistemas do ramo Educação.

Maiores detalhes do currículo do autor podem ser conferidos através dos endereços:

<http://lattes.cnpq.br/4717450794507499>

<https://mcp.microsoft.com/Anonymous//Transcript/Validate>

Transcript Id: 989798 Access Code: 06525538