

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Francisco Aislan da Silva Freitas

**Sistema IoT para predição da evasão escolar usando técnicas de
Aprendizagem de Máquina em dados socioeconômicos**

Fortaleza – CE

15 de fevereiro de 2022

Francisco Aislan da Silva Freitas

Sistema IoT para predição da evasão escolar usando técnicas de Aprendizagem de Máquina em dados socioeconômicos

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Linha de Pesquisa: Computação Aplicada

Orientador: Prof. Dr. Pedro Pedrosa Rebouças Filho

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PRPI)

Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Fortaleza – CE

15 de fevereiro de 2022

Dados Internacionais de Catalogação na Publicação
Instituto Federal do Ceará - IFCE
Sistema de Bibliotecas - SIBI
Ficha catalográfica elaborada pelo SIBI/IFCE, com os dados fornecidos pelo(a) autor(a)

F862s Freitas, Francisco Aislan da.
 Sistema IoT para predição da evasão escolar usando técnicas de Aprendizagem de Máquina em dados socioeconômicos / Francisco Aislan da Freitas. - 2020.
 71 f. : il. color.

 Dissertação (Mestrado) - Instituto Federal do Ceará, Mestrado em Ciência da Computação, Campus Fortaleza, 2020.

 Orientação: Prof. Dr. Pedro Pedrosa Rebouças Filho.

 1. IoT. 2. Evasão Escolar. 3. Aprendizagem de máquina. I. Título.

CDD 004



Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PRPI)
Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Francisco Aislan da Silva Freitas

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação, sendo aprovada pela Coordenação do Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará e pela banca examinadora:

**Orientador: Prof. Dr. Pedro Pedrosa Rebouças
Filho**

Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

**Prof. Dr. Antonio Wendell de Oliveira
Rodrigues**

Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

Prof. Dr. Antonio Carlos da Silva Barros

Universidade da Integração Internacional da Lusofonia
Afro-Brasileira (Unilab)

Fortaleza – CE
15 de fevereiro de 2022

Agradecimentos

Em primeiro lugar, agradeço a Deus por estar do meu lado guiando os meus pensamentos, soprando palavras, e me mostrando o caminho.

Agradeço aos meus pais Julio Macelio e Maria dos Santos, por sempre terem incentivado os meus estudos e que sem eles jamais teria conseguido.

A minha esposa Tércira Martins pelo seu carinho e contribuição nas noites em que passou em claro me acompanhado nessa jornada. E a minha filha Laura Martins Freitas que me trouxe muitas alegrias. Um agradecimento especial a minha sogra Socorro Roberto e meu sogro Eudo Martins que durante muitos dias cuidou da minha filha para que eu pudesse desenvolver esse trabalho.

Ao meu Irmão Alexandre Freitas e a minha sobrinha Isabele, pelo convívio familiar.

Aos meus amigos do mestrado, Vitor Guimarães, Solon Pinheiro, Leandro Cândido, Carlos Estevão, Elias Paulino, Jackanderson, Monck Charles e Róger Sarmento que me incentivaram a realizar mais esta conquista.

A todos os meus Professores do mestrado, ao Coordenador do PPGCC Pedro Pedrosa e Wendell, pela condução e organização deste curso.

Aos meus Professores orientadores Pedro Pedrosa Rebouças Filho, Cláudio Sá Marques Medeiros e Gilvandenys Leite Sales, pelos ensinamentos que colhi e pela certeza da contribuição árdua desses profissionais para o meu crescimento, pela dedicação, incentivo e apoio nesta grande conquista.

E a todos que de forma direta ou indireta contribuíram para a realização deste trabalho.

*“Não se preocupe com a distância entre seus sonhos e a realidade. Se você pode sonhá-los
você pode realizá-los.”
(Belva Davis)*

Resumo

A evasão escolar permeia várias modalidades de ensino e tem gerado prejuízos sociais, econômicos, políticos e acadêmicos a todos os envolvidos no processo educacional. É interessante que seja realizada uma análise sobre os dados socioeconômicos desses alunos, a fim de mapear o perfil dos discentes evasores. Essa análise prévia dos possíveis evasores favorece o grupo gestor na tomada de decisão, propondo ações que podem alterar o destino do aluno com perfil de evasão, proporcionando ao discente a oportunidade de conclusão de mais uma etapa acadêmica, como também oportunidades profissionais que muitas vezes são mais disponíveis para as pessoas que possuem ensino superior completo. Neste contexto, este trabalho apresenta uma estrutura de *Internet of Things* (IoT) para prever abandono usando métodos de aprendizado de máquina, como Árvore de Decisão, Regressão Logística, *Support Vector Machine* (SVM), *k – Nearest Neighbor* (KNN), *Multilayer Perceptron* (MLP) e *Deep Learning* em dados socioeconômicos. Este trabalho propõe a automação do processo de previsão por um método capaz de obter informações que seriam difíceis e demoradas para os humanos obterem, contribuindo para uma previsão mais precisa. Além disso, com o advento da IoT, é possível criar uma ferramenta altamente eficiente e flexível para melhorar os problemas relacionados a gerenciamento e serviços, o que pode fornecer uma previsão de evasão de novos alunos que ingressam em cursos de nível superior, permitindo acompanhamento personalizado para alunos a reverter uma possível desistência. Para sua implementação, utiliza os dados socioeconômicos do formulário de pré-matrícula do sistema Q-Acadêmico do Instituto Federal de Ciência e Tecnologia do Ceará (IFCE) campus Fortaleza. Nos casos quando o aluno evadir ou concluir o curso, o gestor informará esse dado no sistema, favorecendo para realizar as devidas análises sobre esse conjunto. O sistema Q-Acadêmico não apresenta nenhuma ferramenta ou módulo que possa utilizar os dados para realizar a predição se o aluno irá evadir ou não. Assim, para a previsão prévia da evasão escolar com o uso dos dados do sistema Q-Acadêmico, busca-se utilizar os dados socioeconômicos, por se tratar de informações que remetem à vida externa a escola dos discentes. O modelo tem como entrada uma matriz com os dados socioeconômicos contendo sexo, idade, raça, índice de desenvolvimento humano por bairro, renda, origem e a distância entre o campus e a casa do aluno. O modelo é aplicado em quatro cursos de engenharia Telecomunicações, Mecatrônica, Civil, Computação do IFCE campus Fortaleza. Como resultado desta pesquisa é concebido um mecanismo de predição da evasão escolar baseado nos dados socioeconômicos, a fim de auxiliar a gestão escolar na tomada de decisões para que se possa aumentar as chances de reversão de uma propensão à evasão. A abordagem foi validada analisando os parâmetros *Accuracy*, *F1 – Score*, *Recall* e *Precision*. Os resultados mostraram que o sistema desenvolvido obteve 99,34 % de *Accuracy*, 99,34 % de *F1 – Score*, 100 % de *Recall* e 98,69 % de *Precision* usando a Árvore de Decisão. Assim, o sistema desenvolvido apresenta-se como

uma opção viável para uso nas universidades para prever estudantes com probabilidade de deixar a universidade.

Palavras-chave: IoT. Evasão Escolar. Aprendizagem de máquina.

Abstract

School dropout permeates various teaching modalities and has generated social, economic, political and academic losses for everyone involved in the educational process. It is interesting that an analysis is carried out on the socioeconomic data of these students, in order to map the profile of the evading students. This prior analysis of possible evaders favors the management group in decision making, proposing actions that can change the destiny of the student with an evasion profile, providing the student with the opportunity to complete an academic stage too, as well as professional opportunities that are often more available to the students. people who have completed higher education. In this context, this work presents an Internet of Things (IoT) structure to predict abandoning machine learning methods, such as Decision Tree, Logistic Regression, Support Vector Machine (SVM), k Nearest Neighbor (KNN), Multilayer Perceptron (MLP) eDeep Learning in socioeconomic data. This work proposes the automation of the forecasting process by a method capable of obtaining information that would be difficult for humans to obtain, contributing to a more accurate forecast. In addition, with the advent of IoT, it is possible to create a highly efficient and flexible tool to improve the problems related to management and services, which can provide a drop-out forecast for new students entering higher-level courses, allowing personalized monitoring for students to reverse a possible dropout. For its implementation, it uses the socioeconomic data from the pre-registration form of the Q-Academic system of the Federal Institute of Science and Technology of Ceará (IFCE) campus Fortaleza. In cases when the student escapes or finishes the course, the manager will inform this data in the system, favoring to carry out the due analyzes on this set. The Q-Academic system does not have any tools or modules that can use the data to make the prediction whether the student will evade or not. Thus, for the prior prediction of school dropout using data from the Q-Academic system, we seek to use socioeconomic data, as it is information that refers to the life outside the students' school. The model has as input a matrix with socioeconomic data containing sex, age, race, human development index by neighborhood, income, origin and the distance between the campus and the student's home. The model is applied in four engineering courses Telecommunications, Mechatronics, Civil, Computing at IFCE campus Fortaleza. As a result of this research, a prediction mechanism for school dropout based on socioeconomic data is conceived, in order to assist school management in decision making so that the chances of reversing a propensity to drop out can be increased. The approach was validated by analyzing the parameters Accuracy, F1 Score, RecallePrecision. The results showed that the developed system obtained 99.34% of Accuracy, 99.34% of F1 Score, 100% of Recalle and 98.69% of Precision using the Decision Tree. Thus, the developed system presents itself as a viable option for use in universities to predict students likely to leave the university.

Keywords: IoT. School dropout. Machine learning.

Lista de ilustrações

Figura 1 – Etapas do processo da metodologia: Universo da pesquisa, extração dos dados, características socioeconômicas, análise estatística, aprendizagem de máquina, resultados coletados, análises e discussões e estrutura IoT.	29
Figura 2 – Identificação dos valores ausentes nos atributos.	31
Figura 3 – Etapas para o processo de análise exploratória dos dados: 1º Identificação dos atributos, 2º Etapa análise univariada, 3º Etapa análise bivariada multivariada, 4º Etapa detectando valores ausentes, 5º Etapa detectando valores discrepantes e 6º Etapa validando hipóteses	34
Figura 4 – Neurônio artificial - Perceptron com a camada de entrada, os pesos sinápticos, a função soma e a função de ativação	35
Figura 5 – Representação de uma rede MLP expondo a camada de entrada, as ligações dos pesos sinápticos, a camada oculta, os pesos sinápticos que interligam a camada de saída e a camada de saída com a função soma e ativação	37
Figura 6 – Histogramas dos alunos que concluíram (A (Idade), C (IDHB), E (Renda) e G (Distância)) e os que evadiram (B (Idade), D (IDHB), F (Renda) e H (Distância))	45
Figura 7 – <i>Boxplot</i> dos alunos que concluíram (A (Idade), C (IDHB), E (Renda) e G (Distância)) e os que evadiram (B (Idade), D (IDHB), F (Renda) e H (Distância))	47
Figura 8 – Gráfico dispersão: Distância (entre 0 à 70) x IDHB (entre 0 à 1), na cor azul os que não evadiram e em laranja os que evadiram	48
Figura 9 – Gráfico dispersão: Distância (entre 0 à 70) x Origem (Filantrópica, Pública Municipal, Pública Federal, Pública Estadual, Privada e Outros), na cor azul os que não evadiram e em laranja os que evadiram	49
Figura 10 – Gráfico dispersão: Renda (de 1 Salário mínimo à mais de 20 salários) x Origem (Filantrópica, Pública Municipal, Pública Federal, Pública Estadual, Privada e Outros)	49
Figura 11 – Gráfico dispersão: Raça (Brancos, Pardos, Pretos, Amarelos, Indígenas e não declaram) x IDHB (entre 0 à 1)	50
Figura 12 – Gráfico dispersão: Raça (Brancos, Pardos, Pretos, Amarelos, Indígenas e não declaram) x Renda (de 1 Salário mínimo à mais de 20 salários)	50
Figura 13 – Correlação de Pearson entre os atributos, sexo, idade, raça, IDHB, renda, origem, distância e evasão	51
Figura 14 – Fluxograma da abordagem proposta para detecção e classificação on-line de possíveis alunos que abandonam a escola.	61

Figura 15 – Visão geral do processamento de registros socioeconômicos na estrutura da IoT. 62

Figura 16 – Camadas da arquitetura IoT 63

Lista de tabelas

Tabela 1 – Variáveis contínuas com os correspondentes mínimos e máximos observados.	31
Tabela 2 – Variáveis categóricas com as respectivas categorias e os valores que podem ser assumidos pelas respectivas categorias. O valor atribuído é o percentual de abandono observado na categoria (o valor atribuído para cada categoria da variável renda corresponde a média salarial da categoria)	32
Tabela 3 – Conjunto de parâmetros para treinamento da posição a_1 até a_7 (sexo, idade, raça, IDHB, renda, escola de origem e dist. campus / casa) e a classe y (não-evasão e evasão)	33
Tabela 4 – Dados estatísticos - Alunos não-evasores (Qnt, Média, DP, Mín, 25 %, 50 %, 75 % e Máx)	39
Tabela 5 – Dados estatísticos - Alunos evasores (Qnt, Média, DP, Mín, 25 %, 50 %, 75 % e Máx)	39
Tabela 6 – TF: Atributo Sexo - Alunos não-evasores (Fa, Fr, M, F e FT)	41
Tabela 7 – TF: Atributo Sexo - Alunos evasores (Fa, Fr, M, F e FT)	41
Tabela 8 – TF: Atributo Raça - Alunos não-evasores (Pardos, Brancos Pretos, Amarelos, Indígenas, Não declararam, FT, FA, FR)	42
Tabela 9 – TF: Atributo Raça - Alunos evasores (Pardos, Brancos Pretos, Amarelos, Indígenas, Não declararam, FT, FA, FR)	42
Tabela 10 – TF: Atributo Origem - Alunos não-evasores (Privada, Estadual, Federal, Municipal, Filantrópica, Outros, FT, FA e FR)	42
Tabela 11 – TF: Atributo Origem - Alunos evasores (Privada, Estadual, Federal, Municipal, Filantrópica, Outros, FT, FA e FR)	42
Tabela 12 – Tabela de Contingência: Atributo Sexo (Masculino e Feminino), a Evasão (Sim e Não) e o Total	43
Tabela 13 – Tabela de Contingência: Atributo Raça (Brancos, Pardos, Pretos, Amarelos, Indígenas e não declaram), a Evasão (Sim e Não) e o total)	43
Tabela 14 – Tabela de Contingência: Atributo Origem (Privada, Pública Federal, Pública Estadual, Pública Municipal, Filantrópica e Outros), a Evasão (Sim e Não) e o Total)	44
Tabela 15 – Resultado da predição de alunos com risco de abandono escolar usando dados socioeconômicos aplicando técnicas de aprendizagem de máquina, considerando <i>Recall</i> (Rec), <i>Precision</i> (Pre), <i>Fscore</i> e <i>Accuracy</i> (Acc) como métricas de avaliação	53

Tabela 16 – Comparação dos classificadores com métodos recentes usando conjuntos de dados diferentes e acurácia (Acc) como métrica de avaliação	55
Tabela 17 – Relação dos atributos previsores de cada autor referenciado neste trabalho como objeto comparativo entre os atributos selecionados	56
Tabela 18 – Comparação do treinamento dos classificadores e tempo médio de teste da amostra (τ) em segundos.	64

Lista de abreviaturas e siglas

Acc	<i>Accuracy</i>
CRISP-DM	<i>Cross Industry Standard Process for Data Mining</i>
DCR	Distância entre o Campus e a Residência do aluno
DP	Desvio Padrão
FA	Frequência Absoluta
FN	Falso Negativo
FP	Falso Positivo
FR	Frequência Relativa
FT	Frequência Total
IA	Inteligência Artificial
ICA	Inteligência Computacional Aplicada
IDH	Índice de Desenvolvimento Humano
IDHB	Índice Desenvolvimento Humano por Bairro
IFCE	Instituto Federal de Educação, Ciência e Tecnologia do Ceará
KNN	K-Nearest Neighbors
MLP	Multi Layer Perceptron
OECD	Organization for Economic Cooperation and Development
PDE	Plano de Desenvolvimento da Escola
PNE	Plano Nacional de Educação
Prec	<i>Precision</i>
Rec	<i>Recall</i>
RNA	Redes Neurais Artificiais
SPSS	<i>Statistical Package for the Social Sciences</i>

SVM	Support Vector Machine
TF	Tabelas de Frequências
TN	Verdadeiro Negativo
TP	Verdadeiro Positivo

Lista de símbolos

Σ	Letra grega maiúscula Sigma
α	Letra grega minúscula Alfa
Θ	Letra grega maiúscula Teta
λ	Letra grega minúscula Lambda
∇	Letra grega Nabla
Γ	Letra grega maiúscula Gama
Λ	Letra grega maiúscula Lambda

Sumário

1	INTRODUÇÃO	23
1.1	Motivação para a Dissertação	25
1.2	Descrição do problema	26
1.3	Objetivos Geral e Específicos	26
1.4	Produção científica	26
1.5	Estrutura da Dissertação	27
2	METODOLOGIA	29
2.1	Universo da Pesquisa	29
2.2	Extração dos dados	30
2.3	Características socioeconômicas	30
2.4	Análise Estatística da base de dados	33
2.5	Aprendizagem de máquina	34
2.6	Métricas	38
3	RESULTADOS	39
3.1	Resultados da análise estatística dos dados	39
3.2	Resultados dos classificadores de Aprendizagem de Máquina	51
4	ESTRUTURA IOT	61
5	CONCLUSÃO	65
	REFERÊNCIAS	67

1 Introdução

A evasão escolar é um dos problemas analisados mundialmente que tem despertado a atenção de pesquisadores acadêmicos e órgãos governamentais devido à relevância da educação no processo do desenvolvimento econômico e na atenuação das desigualdades sociais. Conforme a teoria do capital humano, a educação é vital para o progresso econômico, aumento da competitividade do país e melhoria do bem-estar social (SOUZA; TABOSA; SIMONSASSI, 2015). Nesse contexto, surge a importância de identificar antecipadamente que perfis de estudantes serão considerados como de evasores, pois essa informação pode auxiliar na tomada de decisões para que os profissionais de educação possam adotar medidas para reverter esta tendência de evasão.(JÚNIOR; NORONHA; KAESTNER, 2018).

Segundo a *Organization for Economic Cooperation and Development* (OECD)(2018), as despesas do setor público relativo à manutenção tanto do ensino primário quanto do superior, em média internacional, atingem investimentos de 11 % nos países que compõem a OECD, e esse valor varia em torno de 6 % a 17 %.

Entre 2011 e 2015, a percentagem média do investimento total do governo destinado à educação primária e terciária permaneceram relativamente estáveis nos países da OECD, em torno de 11 %. No Brasil, essa aplicação de capitais na educação chega a atingir aproximadamente 80% se comparada com outros setores, já para outros serviços, incluindo a educação, esse emprego financeiro aproximou-se de 110 % entre 2011 a 2015(OECD, 2018).

O Plano de Desenvolvimento da Escola (PDE), de origem nacional, assume plenamente, inclusive na denominação, a agenda do programa “Compromisso Todos pela Educação”, no qual foram definidas cinco metas segundo (SAVIANI et al., 2007):

- Todas as crianças e jovens de 4 a 17 anos deverão estar na escola;
- Toda criança de 8 anos deverá saber ler e escrever;
- Todo aluno deverá aprender o que é apropriado para sua série;
- Todos os alunos deverão concluir o ensino fundamental e o médio;
- O investimento necessário na educação básica deverá estar garantido e bem gerido.

Realizar uma análise prévia sobre a possível evasão de um aluno pode ajudar a atingir as premissas 1 e 4 do “Compromisso Todos pela Educação” desenvolvida pelo PDE. Para o Plano Nacional de Educação (PNE) assegurar a qualidade da educação nos cursos técnicos de nível médio corrobora com a redução da evasão escolar proposta pelo “Compromisso Todos pela Educação” (INEP, 2015).

A previsão da evasão escolar é um tema já estudado por diversos pesquisadores (LANES; ALCÂNTARA, 2018) (RAMOS et al., 2018) (SALES; LEITE; JOYE, 2012). No entanto, neste trabalho é apresentada uma abordagem diferente destes, pois busca compreender esse problema analisando dados socioeconômicos de alunos dos cursos do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) campus Fortaleza. Esses dados são utilizados para realizar os treinamentos e testes em Redes Neurais Artificiais (RNAs), gerando um modelo para prever se novos alunos interessados nos cursos do IFCE do campus Fortaleza serão futuros evasores.

O IFCE campus Fortaleza é uma das instituições brasileiras que tem enfrentado o desafio de reduzir a elevada taxa de evasão, pois aproximadamente 13% dos alunos abandonam seus cursos durante a trajetória acadêmica (PEÇANHA, 2019). Por isso, existe a necessidade de desenvolver propostas automatizadas para auxiliar na detecção antecipada da desistência escolar, permitindo que suas possíveis causas possam ser analisadas e tratadas prematuramente reduzindo possíveis custos oriundos deste processo.

A IoT permite a facilidade de interação entre tecnologias do mundo real e recursos virtuais. O desenvolvimento dessa tecnologia estimula o crescimento desse conceito em vários setores, como indústria, saúde, segurança, educação e outras áreas (NEHRU; CHAKRABORTY, 2019). Esse crescimento incentiva outros pesquisadores a desenvolver trabalhos que abordem o assunto. Entre os setores mencionados, o setor educacional apresentou atraso no uso dessas tecnologias. Somente recentemente, as instituições educacionais estão entendendo a relevância da IoT e adotando-a nas subáreas da educação, como em administração, ensino e políticas internas.

Existem vários tipos de dispositivos e sistemas de IoT aplicados na educação para ajudar na tomada de decisões, incluindo sistema de monitoramento de desempenho acadêmico (RAKSHITH; KENCHANNAVAR; THYAGARAJU, 2017), sistema de otimização de ensino e aprendizagem (BAGHERI; MOVAHED, 2016) e sistemas que permitem prever o sucesso do aluno no ensino médio (ROBLYER; MARSHALL, 2002). No entanto, cada estudo de caso passa por vários desafios que geralmente podem comprometer a entrega de uma estrutura de IoT.

O objetivo deste trabalho é reunir vários aspectos computacionais, com o diferencial de garantir a usabilidade e promover o uso intuitivo por usuários leigos de computadores que não conhecem as técnicas de análise de abandono. Outra contribuição indireta é apoiar a criação de grandes bancos com dados socioeconômicos, dentro de uma plataforma científica, com dados compartilhados de usuários que trabalham nesta área de estudo. Isso deve melhorar a estrutura aplicando seus algoritmos a uma base muito mais ampla. Assim, espera-se que essa plataforma, especificamente no campo da educação, contribua para a previsão do abandono escolar com dados socioeconômicos e possa proporcionar aos gestores da escola a oportunidade de intervir revertendo essa situação.

O principal desafio na implementação desse tipo de aplicativo é a combinação de três elementos que geralmente não são integrados: conhecimento em informática, análise de risco de desistência e alto poder computacional para fornecer previsões com tempo mínimo entre a requisição e os resultados. Esses três elementos podem ser integrados usando a Internet das Coisas (IoT) (JR et al., 2019). Vários projetos de pesquisa que abordam a previsão de evasão escolar podem ser encontrados, aplicando algoritmos diferentes e com soluções muito precisas, mas a maioria não possui uma interface apropriada para uso de profissionais da educação comum (MARTINHO; NUNES; MINUSSI, 2013; SANSONE, 2019).

Portanto, as principais contribuições deste trabalho são:

- Criação de uma plataforma simples para prever desistências com base em informações socioeconômicas;
- Torna possível prever a evasão nas universidades para ajudar os especialistas a reduzir a evasão no ensino superior;
- Criação de sistema de suporte a previsões que pode ser acessado a partir de diferentes dispositivos;
- Permite o uso de um sistema de IoT para prever desistências sem a necessidade de hardware específico e conhecimento avançado do usuário.

1.1 Motivação para a Dissertação

A motivação maior para a realização desta pesquisa é a de contribuir para o prognóstico precoce de evasão escolar no IFCE campus Fortaleza. Isso se fará através de um modelo de aprendizagem de máquina que normalmente apresenta bons resultados na predição de problemas com complexidade em suas amostras. Assim, será usado como base dados socioeconômicos para uma possível compreensão do perfil do aluno antes de ingressar no campus supracitado; informação esta que pode ser conferida através do entendimento das informações socioeconômicas.

Ainda que a gestão do IFCE campus Fortaleza tenha suporte no acompanhamento ao desempenho do aluno e ofereça algumas ações no que tange a permanência do mesmo, há certa dificuldade na previsão com a saída de novos discentes. Poucas são as opções de indicadores que expõem as várias facetas deste fenômeno nos cursos presenciais do IFCE campus Fortaleza. Informar à gestão do campus, ou mesmo passar *feedbacks*, de um possível afastamento dos novos alunos na pré-matricula, favorece uma tomada de decisão eficaz quanto a redução das chances de abandono dentro do sistema educacional.

1.2 Descrição do problema

No artigo apresentado por Nagy e Molontay (2018), a rede neural *Deep Learning* foi introduzida como um poderoso classificador na predição do risco de evasão escolar utilizando dados socioeconômicos. Este trabalho propõe uma nova abordagem, utilizando os dados socioeconômicos do sistema Q-Acadêmico do IFCE campus Fortaleza e aplicando técnicas de aprendizagem de máquina na predição do risco de abandono escolar. Como o instituto referenciado não classifica os alunos com risco de abandono, o método proposto usa o modelo treinado para realizar essa predição.

A vantagem deste modelo reside na sua capacidade de avaliar dados socioeconômicos automaticamente e destacar os discentes com perfil evasor usando apenas as informações fornecidas por eles.

1.3 Objetivos Geral e Específicos

Desenvolver uma ferramenta IoT capaz de fazer a predição de indivíduos com alto risco de evasão escolar, em cursos de Engenharias do IFCE campus Fortaleza.

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

- Estudar, a partir de revisão bibliográfica, como os fatores socioeconômicos impactam na evasão;
- Utilizar algoritmos de aprendizagem de máquina para mensurar a predição da evasão;
- Validar a metodologia através de simulações computacionais.
- Desenvolver a arquitetura IoT para compor modelo treinado

1.4 Produção científica

Durante este projeto de mestrado, os seguintes trabalhos científicos foram aceitos e publicados, a saber:

- FREITAS, Francisco Aislan da Silva; ALBUQUERQUE, Monck Charles Nunes de; SALES, Gilvandenys Leite; MEDEIROS, Cláudio Marques de Sá; CÂNDIDO, Antônio Leandro Martins; FILHO, Pedro Pedrosa Rebouças. O Uso de Redes Neurais Artificiais para Análise da Presencialidade em um Fórum LV. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2018. p. 1903.
- FREITAS, Francisco Aislan da Silva; MAIA, Paulo Henrique; BASILIO, Jezmael de Oliveira. Progame-Ambiente Educacional Gamificado com Elementos dos

Descritores de Matemática do Sistema Permanente De Avaliação Da Educação Básica Do Ceará. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2018. p. 1032.

- FREITAS, Francisco Aislan da Silva; MAIA, Paulo Henrique; BASILIO, Jezmael de Oliveira. PROGAME - Ambiente Gamificado para Simulação do Exame SPAECE. In: **Concurso de Apps do Congresso Brasileiro de Informática na Educação**. 2019.
- FREITAS, Francisco Aislan da Silva; SILVA, Alisson Oliveira; AMORIM, Laudenor; SALES, Gilvandenys Leite. A Física da Bailarina. In: **Concurso de Apps do Congresso Brasileiro de Informática na Educação**. 2019.
- CÂNDIDO, Antônio Leandro Martins; JÚNIOR, Corneli Gomes Furtado; FREITAS, Francisco Aislan da Silva. **Moderação inteligente de mensagens em ambientes virtuais de aprendizagem para alunos privados de liberdade**. In: SBIE 2019- Trilha 4 (). [s.n.], 2019

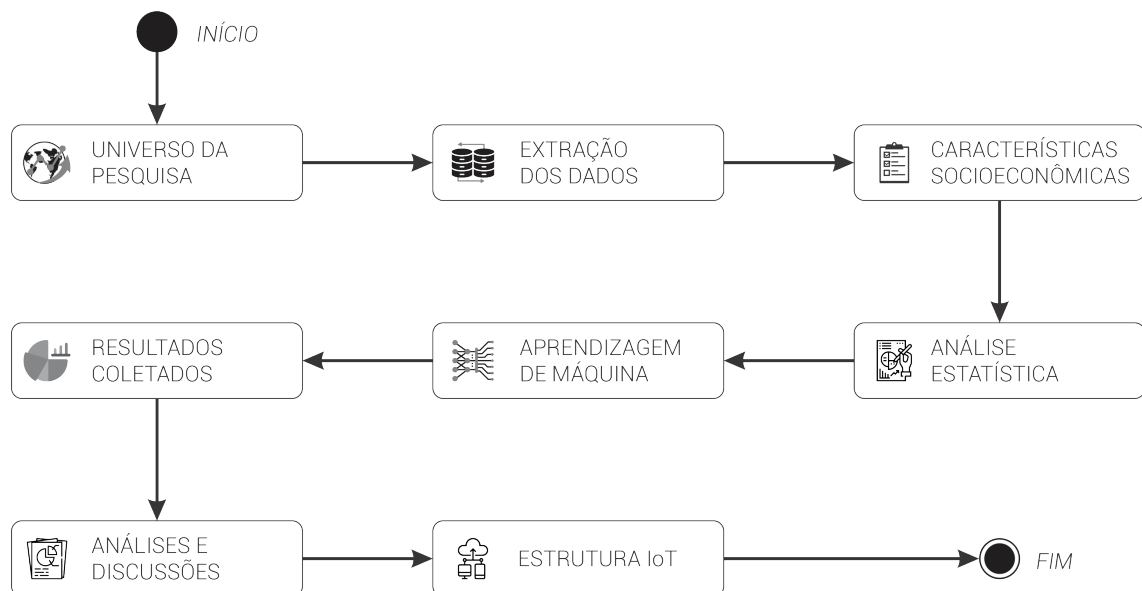
1.5 Estrutura da Dissertação

A dissertação é desenvolvida em quatro capítulos: Introdução, Metodologia, Resultados e Estrutura IoT. No *Capítulo 1* Introdução, é apresentado uma breve contextualização do problema, a motivação para dissertação, a descrição do problema, os objetivos, a produção científica e a estrutura da dissertação. No *Capítulo 2* metodologia, é abordado sobre o universo da pesquisa, a seleção dos dados, as definições dos atributos socioeconômicos, a análise estatística, os algoritmos de aprendizagem de máquina utilizados nas simulações computacionais. No *Capítulo 3* resultados, são apresentados os resultados da análise estatística e dos algoritmos de aprendizagem de máquina abordados nesta pesquisa comparando com outras configurações e com abordagens recentes. No *Capítulo 4* Estrutura IoT, são apresentadas os fluxos da abordagem proposta para detecção e classificação on-line de possíveis alunos que abandonam a escola, uma visão geral da arquitetura IoT e os resultados com o tempo dos treinamentos e testes em cada classificador utilizado neste trabalho.

2 Metodologia

Este capítulo tem por objetivo descrever o universo da pesquisa, o processo de extração dos dados, as características socioeconômicas abordadas nesta pesquisa, o percurso da análise estatística do conjunto de dados, os algoritmos de aprendizagem de máquina que compõem as simulações computacionais (Figura 1).

Figura 1 – Etapas do processo da metodologia: Universo da pesquisa, extração dos dados, características socioeconômicas, análise estatística, aprendizagem de máquina, resultados coletados, análises e discussões e estrutura IoT.



Fonte: Autor

2.1 Universo da Pesquisa

Quanto à população ou ao universo da pesquisa foi utilizada a base de dados de alguns cursos do ambiente acadêmico do IFCE campus Fortaleza. Essa plataforma é conhecida por Q-acadêmico¹. A plataforma hospeda diversos dados de cursos de diferentes modalidades de ensino. Tendo cursos e disciplinas que são do ensino médio técnico, graduação e pós-graduação na categoria presencial ou a distância. Foram escolhidos os cursos de graduação em engenharia do IFCE campus Fortaleza. Cursos utilizados para a coleta de dados:

- Engenharia de Telecomunicações
- Engenharia de Mecatrônica

¹ <https://qacademico.ifce.edu.br/>

- Engenharia Civil
- Engenharia da Computação

2.2 Extração dos dados

Com o objetivo de construir e adequar a base de dados e tornar compatível a análise com os modelos de aprendizagem de máquina, foi utilizada a linguagem de programação Python com a biblioteca denominada de Pandas para realizar análise e manipulação dos dados, Numpy para aplicar as funções matemáticas pertinentes a pesquisa, como também manipular com facilidade vetores e matrizes multidimensionais e, por fim, a biblioteca *sci-kit learning* com os algoritmos de aprendizagem de máquina para realizar a predição de cada aluno como um possível evasor.

Para a coleta de dados foram realizados os seguintes passos:

- Realizar o backup da base de dados do Q-acadêmico em formato csv e depois restaurar o backup em uma máquina local.
- Identificar os principais atributos socioeconômicos ligados à evasão.
- Montar uma base com as informações extraídas do tópico anterior
- Selecionar apenas as amostras que fazem parte dos cursos de engenharia do IFCE-campus Fortaleza.

2.3 Características socioeconômicas

A base de dados utilizada neste trabalho foi coletada contendo informações socioeconômicas sobre os 1549 alunos dos cursos de Engenharias. Os alunos tiveram a sua identificação preservada, sendo a matrícula o atributo usado para relacionar os dados.

Os dados utilizados são: sexo, idade, raça, índice Desenvolvimento Humano por Bairro (IDHB), renda, escola de origem, procedência, distância entre o campus e a residência (DCR), turno e se houve ou não evasão. Nesse sentido, esse trabalho utiliza o atributo evasão como rótulo, transformando o problema em uma classe binária (Não-Evasão e Evasão).

O atributo procedência está relacionado se os alunos são de origem urbana ou rural. Nesta base de dados foram selecionados apenas alunos de zona urbana, tendo em vista que existem poucas amostras que são caracterizadas como rural.

Para o atributo turno, os cursos de Engenharia do IFCE-campus Fortaleza são ofertados diurnamente, sendo assim não havendo a necessidade de utilizar esse atributo na base de treinamentos e testes.

Entre as informações dos alunos selecionados, existem dois tipos de variáveis, as contínuas e as categóricas. As variáveis contínuas tiveram seus valores preservados. A Tabela 1 apresenta as variáveis contínuas utilizadas, bem como os valores máximos e mínimos que eles podem assumir. Por outro lado, as variáveis categóricas passaram por um processo de atribuição de valores, cada categoria assume um valor possível de ser aplicado a ela. O valor atribuído a cada categoria corresponde ao percentual de abandono observado para essa categoria específica. A única exceção é a variável Renda, em que o valor atribuído é a média de faixas salariais. A Tabela 2 apresenta as variáveis categóricas e os valores atribuídos a cada uma.

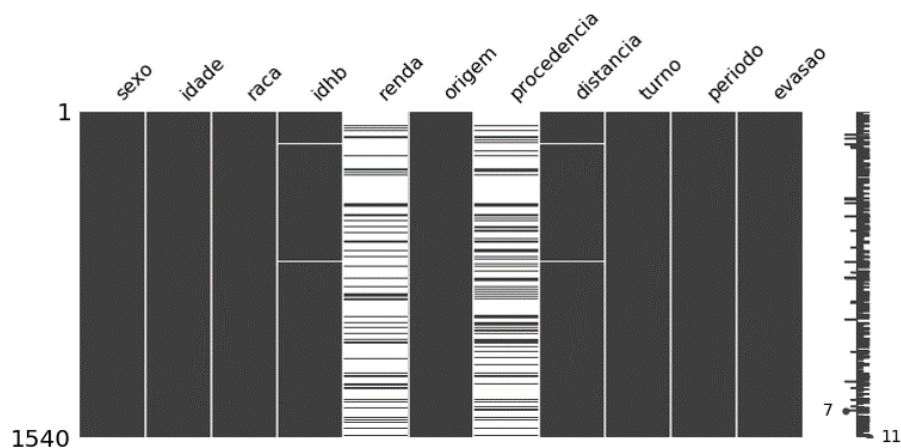
Tabela 1 – Variáveis contínuas com os correspondentes mínimos e máximos observados.

Variável Contínua	Mínimo	Máximo
Idade	17 anos	54 anos
IDHB	0,11	0,95
DCR	1,3 km	67,4 km

Fonte: Autor.

Antes de realizar qualquer interferência no conjunto de dados, é importante listar os elementos que compõem a fase de pré-processamento das informações. O primeiro elemento diz respeito aos *missing values*, ilustrado na Figura 2, caso tenha ocorrência no conjunto de dados, os mesmos devem ser tratados. Compreender, analisar e tratar os valores ausentes favorece um conjunto apto para realizar treinamentos e testes da RNA.

Figura 2 – Identificação dos valores ausentes nos atributos.



Fonte: Autor.

O gráfico apresentado na Figura 2, trata-se da visualização mais clara dos valores ausentes no conjunto de dados. As colunas são representadas pelos atributos precursores e a classe. As linhas são representadas pela quantidade de amostras, e as lacunas na cor branca

Tabela 2 – Variáveis categóricas com as respectivas categorias e os valores que podem ser assumidos pelas respectivas categorias. O valor atribuído é o percentual de abandono observado na categoria (o valor atribuído para cada categoria da variável renda corresponde a média salarial da categoria)

Variável Categórica	Categoria	Valor Atribuído
Sexo	Masculino	0,1551
	Feminino	0,1238
Raça	Branco	0,1649
	Pardo	0,1434
	Preto	0,1410
	Amarelo	0,2222
	Indígena	0
	Não se declaram	0,1307
Escola de Origem	Privada	0,1370
	Federal	0,1269
	Estadual	0,1607
	Municipal	0,005
	Filantrópica	0,2857
	Outras	0,2157
Renda	0 a 1 salário mínimo	0,5
	1 a 2 salários mínimos	1,5
	2 a 3 salários mínimos	2,5
	3 a 5 salários mínimos	4
	5 a 10 salários mínimos	7,5
	10 a 20 salários mínimos	15
	+ 20 salários mínimos	20

Fonte: Autor.

representam os valores ausentes da amostra. Ao final é exposto um pequeno gráfico de linhas em que é representado pela quantidade mínima e máxima de colunas que possuem os atributos preenchidos nas amostras.

Os atributos que possuem recorrência na ausência do preenchimento das amostras são: renda e procedência, para os mesmos seus valores são substituídos pelo o valor médio da variável. Ao final das conversões de cada atributo é gerado um vetor contendo 7 atributos previsores e a classe evasão representado na Tabela 3.

Os algoritmos de aprendizagem de máquina ou de redes neurais artificiais, normalmente apresentam melhores resultados quando são normalizados. Para que fosse possível realizar os treinamentos e testes, é aplicado no conjunto de dados a normalização entre -1 e 1, tornando mais viável o aprendizado da topologia que é proposta.

Tabela 3 – Conjunto de parâmetros para treinamento da posição a_1 até a_7 (sexo, idade, raça, IDHB, renda, escola de origem e dist. campus / casa) e a classe y (não-evasão e evasão)

<i>Posição</i>	<i>Nome</i>
a_1	Sexo
a_2	Idade
a_3	Raça
a_4	IDHB
a_5	Renda
a_6	Escola Origem
a_7	DCR
y	Não-Evasão
	Evasão

Fonte: Autor.

2.4 Análise Estatística da base de dados

O conjunto de dados contém 1549 registros, mas apenas 217 deles contém todas as informações preenchidas de cada atributo referenciado. Inicialmente são realizadas algumas análises sobre o conjunto de dados com todos os registros.

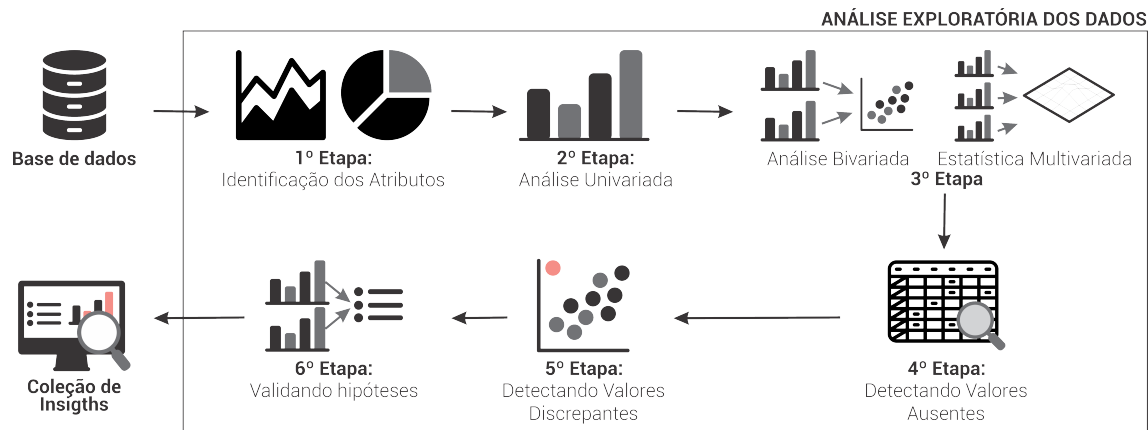
Para a análise estatística é usado o Spyder, interface gráfica para construção de algoritmos na linguagem Python, juntamente com a biblioteca Pandas desenvolvida para manipular e analisar os dados, para o processo de análises é seguido as etapas sugeridas por Tukey (1962) (Figura 3) . Utiliza-se o Pandas com a intenção de encontrar valores de tendência central, assim como valores de média, mediana, máximo, mínimo, coeficiente de variação e desvio padrão, essas informações auxiliam na interpretação da normalidade dos dados (COSTA, 2015).

Nessa análise estatística também é analisado o primeiro, o segundo e o terceiro quartil de cada atributo numérico, a fim de inferir a medida separatriz dos dados, auxiliando na compreensão do conjunto dos alunos que evadem e concluem. Para os atributos categóricos são construídas tabelas de contingência, a fim de inferir a frequência absoluta e a frequência relativa de cada atributo categórico, essas tabelas auxiliam na compreensão das categorias que são propensas à evasão.

Na tentativa de encontrar uma representação de dados quantitativos, agrupados em classes de frequência que permita distinguir a forma, o ponto central e a variação da distribuição, além de outros dados como amplitude e simetria na distribuição dos dados são gerados alguns histogramas de cada atributo, juntamente com gráficos de densidade.

Com o objetivo de pontuar os dados em um eixo vertical e horizontal com a intenção

Figura 3 – Etapas para o processo de análise exploratória dos dados: 1º Identificação dos atributos, 2º Etapa análise univariada, 3º Etapa análise bivariada multivariada, 4º Etapa detectando valores ausentes, 5º Etapa detectando valores discrepantes e 6º Etapa validando hipóteses



Fonte: Autor.

de exibir quanto uma variável é afetada por outra, são construídos alguns gráficos de dispersão. Nessa análise, pode-se identificar aparentemente o conjunto de dados e conferir se é linearmente separável.

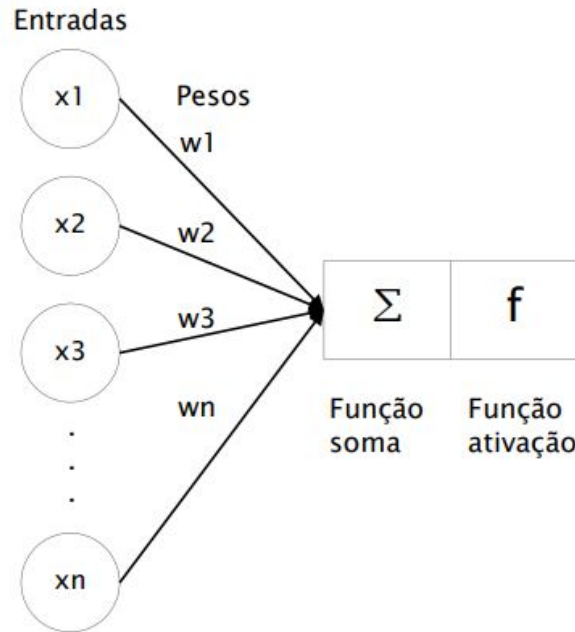
Para ilustrar se existe uma variação nos dados observados de uma variável numérica por meio de quartis, assim como identificar se existe *outlier*, averiguar o grau de dispersão entre os dados, a obliquidade nos dados e a amplitude interquartil são construídos gráficos *boxplots*. Por fim para identificar se os atributos possuem algum tipo de relação forte positivamente ou negativamente, é aplicado o cálculo estatístico de Pearson (1994) e gerado um gráfico para melhor compreensão.

2.5 Aprendizagem de máquina

Neste trabalho, diferentes técnicas de Aprendizado de Máquina foram utilizadas para realizar simulações computacionais no que tange a predição da evasão escolar.

Segundo Haykin (2001) o classificador *Single Layer Perceptron* (conhecido por perceptron simples), é capaz de identificar se o problema estudado é linearmente separável ou não. O processo de aprendizagem dessa rede neural, é inspirada em redes neurais biológicas, ao realizar as trocas de informações as redes possuem processo parecido. São fornecidos valores de entrada, a rede processa e retorna uma resposta, onde o neurônio é ativado somente se a o valor for maior que um limiar. O perceptron possui uma camada de entrada, onde são apresentados os dados, as ligações sinápticas é associadas aos pesos de cada ligação, o núcleo é propriamente o neurônio, onde possui uma função de soma e uma função de ativação e por fim a saída, essa estrutura é ilustrado na Figura 4.

Figura 4 – Neurônio artificial - Perceptron com a camada de entrada, os pesos sinápticos, a função soma e a função de ativação



Fonte: Autor.

$$soma = \sum_{i=1}^n x_i * w_i \quad (2.1)$$

$$f(x) = \begin{cases} 1 & \text{Se } x * w + b \geq 0 \\ 0 & \text{Senão} \end{cases} \quad (2.2)$$

$$erro = RespostaCorreta - RespostaCalculada \quad (2.3)$$

$$w(n + 1) = w(n) + \alpha * erro \quad (2.4)$$

A pesquisa desenvolvida por Bezerra et al. (2016) propôs uma análise nos dados dos Censos Escolares 2011 e 2012 das escolas públicas estaduais e municipais do estado de Pernambuco, utilizando o classificador árvore de decisão na análise da evasão escolar. Essa escolha se deve ao fato desse classificador apresentar facilidade na compreensão dos dados, por ser uma técnica que não exige preparação nas amostras e que pode ser utilizada juntamente com outras ferramentas na tomada de decisões.

A geração da árvore de decisão é um processo recursivo, em que será calculado a entropia, os pesos e o ganho da informação. Na Equação 2.5 está representado o cálculo

da entropia, onde ρ é a probabilidade (GAMA et al., 2004):

$$Entropia = - \sum_{i=1}^n \rho_i \log_2 \rho_i \quad (2.5)$$

$$Peso = \frac{N^o \text{ de amostras do filho}}{N^o \text{ de amostras do pai}} \quad (2.6)$$

$$Ganho \text{ da informação} = Entropia_{pai} - \sum peso_{filho} * entropia_{filho} \quad (2.7)$$

Calixto, Segundo e Gusmão (2017) propôs utilizar o classificador regressão logística por ter como vantagem: a função de estimar a probabilidade da variável dependente assumir um determinado valor em função dos dados conhecidos de outras variáveis. Essa é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias (MENARD, 2002).

$$odds \text{ Ratio} = \log \left(\frac{P(a|c)}{1 - P(a|c)} \right) \quad (2.8)$$

$$Prob(y = 1) = \frac{1}{1 + e^{-\Theta(\sum_{i=1}^m mixi + b)}} \quad (2.9)$$

Sorensen (2016) utiliza o classificador *Support Vector Machine* (SVM) em uma instituição da Carolina do Norte, o autor selecionou o SVM por ser um classificador que consegue lidar bem com grandes conjuntos de amostras, por tratar bem dados de alta dimensão e por ter um processo de classificação rápido (CORTES; VAPNIK, 1995).

$$f(x) = sign[\lambda \cdot y \cdot k(xi \cdot xj)] \quad (2.10)$$

$$k(xi \cdot xj) = \sqrt{\frac{\sqrt{(xi - xj)^2 + (yi - yj)^2}}{width}} \quad (2.11)$$

$$\lambda \rightarrow \nabla \Gamma = 0 \quad y = 1 \quad \lambda y = -1 \quad (2.12)$$

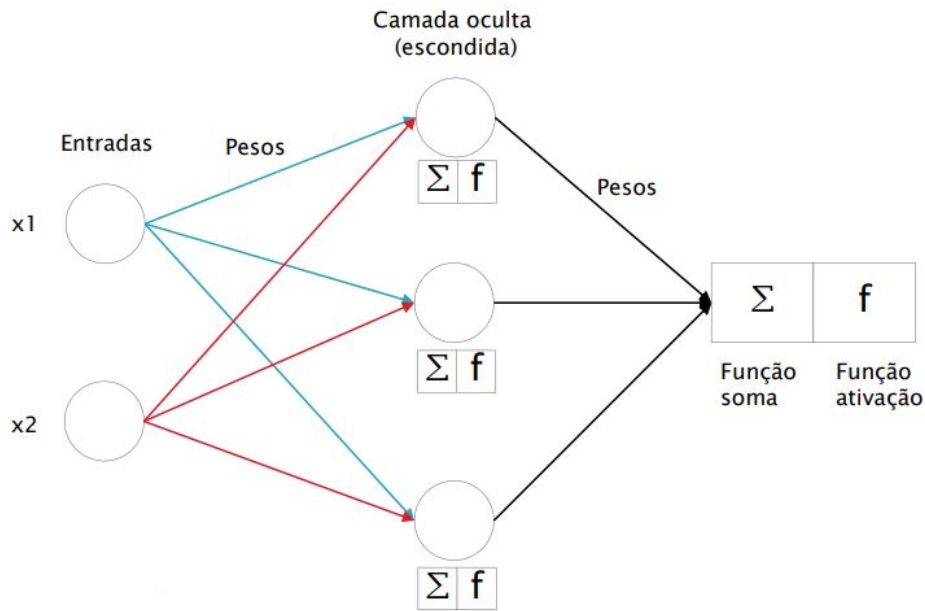
Aulck et al. (2016) afirma que a cada ano, aproximadamente 30 % dos alunos do primeiro ano das instituições de bacharelado nas universidades dos Estados Unidos da América não retornam pelo segundo ano, na tentativa de prever a evasão os autores propuseram utilizar a técnica de aprendizagem de máquina *k - nearest neighbors* (kNN),

tendo como vantagem o rápido tempo de predição, ou seja, o processo de determinar se uma nova classe a partir de um dado de teste é rápido, devido ao fato desse classificador armazenar todo conjunto de treinamento em memória tornando o processo de busca mais rápido (ALTMAN, 1992).

$$\text{Distância Euclidiana } (x,y) = \sqrt{\sum_i^n (x_i - y_i)^2} \quad (2.13)$$

Karamouzis e Vrettos (2008) sugere utilizar uma rede neural artificial *Multilayer Perceptron* (MLP) por ter bons resultados em problemas complexos. Esse modelo consiste em pelo menos três camadas de nós: uma camada de entrada, uma camada oculta e uma camada de saída (HAYKIN, 2001), é ilustrado na Figura 5. Exceto pelos nós de entrada, cada nó é um neurônio que usa uma função de ativação não linear. A rede MLP utiliza uma técnica de aprendizado supervisionado chamada *backpropagation* para treinamento. Suas múltiplas camadas e ativação não linear distinguem a MLP de um perceptron linear. Ele pode distinguir dados que não são linearmente separáveis e ao adicionar mais camadas ocultas é gerado uma rede de aprendizagem profunda (*Deep Learning*).

Figura 5 – Representação de uma rede MLP expondo a camada de entrada, as ligações dos pesos sinápticos, a camada oculta, os pesos sinápticos que interligam a camada de saída e a camada de saída com a função soma e ativação



Fonte: Autor.

$$f(x) = w_0 + K \cdot \sum_{i=1}^n w_i \cdot x_j \quad (2.14)$$

2.6 Métricas

A comparação entre os resultados coletados e as análises e discussões se dão na utilização das métricas de avaliação com base nos resultados da classificação da matriz de confusão. Os resultados da matriz de confusão incluem Verdadeiro Positivo (TP), Verdadeiro Negativo (TN), Falso Positivo (FP) e Falso Negativo (FN), alguns dos quais foram usados para as métricas de avaliação.

Accuracy (Acc) é uma medida entre os registros classificados corretamente e o número total de amostras. Acc é definido na Equação 2.15.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.15)$$

Recall (Rec) mede o número de amostras de uma classe que foi classificado e previsto para ser dessa classe. O *Recall* é definida na Equação 2.16.

$$Recall = \frac{TP}{TP + FN} \quad (2.16)$$

Precision (Prec) quantifica o número de previsões de uma classe que realmente pertencem a essa classe. Precisão é definida na Equação 2.17.

$$Precision = \frac{TP}{TP + FP} \quad (2.17)$$

O *F score* é uma medida que calcula a média harmônica de *recall* e *precisão*. Na Equação 2.18, é ilustrado como o *F score* é calculado.

$$F\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (2.18)$$

3 Resultados

Considerando as métricas Acc, Rec, Prec, $Fscore$ e os dados socioeconômicos, esta seção é dividida em 2 etapas, a primeira etapa é a avaliação estatística sobre a análise exploratória dos dados e o segundo passo é a avaliação dos classificadores de aprendizagem de máquina com trabalhos recentes relacionados.

3.1 Resultados da análise estatística dos dados

Realizando uma análise na base de dados, é possível identificar a distribuição dos dados e se existem tendências. Para os atributos numéricos desses alunos (idade, IDHB, renda e distância) são analisados a média, desvio padrão (DP), mínimo (Mín), 1º quartil (25 %), 2º quartil (50 %), 3º quartil (75 %) e máximo (Máx), todos alunos não-evasores representados na Tabela 4. Já para a representação dos alunos evasores, pode ser conferido na Tabela 5.

Tabela 4 – Dados estatísticos - Alunos não-evasores (Qnt, Média, DP, Mín, 25 %, 50 %, 75 % e Máx)

	Idade	IDHB	Renda	Distância
Qnt	1318	1318	229	1318
Média	24,58	0,46	3,38	12,81
DP	5,56	0,19	3,26	14,65
Mín	17,00	0,11	1,00	1,30
25%	21,00	0,31	1,50	4,60
50%	23,00	0,43	2,50	8,80
75%	27,00	0,63	4,00	13,20
Máx	54,00	0,95	20,00	67,40

Fonte: Autor.

Tabela 5 – Dados estatísticos - Alunos evasores (Qnt, Média, DP, Mín, 25 %, 50 %, 75 % e Máx)

	Idade	IDHB	Renda	Distância
Qnt	231	231	48	231
Média	25,65	0,45	3,56	12,63
DP	6,86	0,19	3,60	13,11
Mín	18	0,11	1,00	1,30
25%	21	0,29	1,50	5,40
50%	23	0,42	2,50	9,50
75%	28	0,63	4,00	12,70
Máx	59	0,95	20,00	67,40

Fonte: Autor.

Analisando o primeiro e o segundo quartil do atributo idade é possível identificar que os mesmos possuem uma faixa etária semelhante entre 17 e 23 anos, caracterizando-se como um perfil de concludente mais jovem. Já para o terceiro quartil, apresentam-se alunos com idades superiores a 27 anos, chegando até os 54 anos. Nesse perfil são identificados discentes que possivelmente tentaram reingressar no universo acadêmico e que, felizmente, concluíram essa etapa.

O atributo IDHB retrata o desenvolvido do bairro em que o aluno reside. Os valores para esse atributo começam com 0,11 que é considerado IDHB muito baixo, chegando a 0,95, classificado como um IDHB muito alto. Basicamente, a formação das turmas que concluíram o curso superior possui IDHB entre muito baixo e médio. Analisando o terceiro

quartil pode-se identificar que é formado por alunos com IDHB alto e muito alto. Com um desvio padrão de 0,19, sugere-se que existem muitos alunos com IDHB muito baixo, baixo e médio do que alunos com o IDHB considerado alto e muito alto. A análise do IDHB pode refletir como é caracterizada a turma dos alunos concludentes, neste caso, sendo uma turma heterogênea.

Para o atributo renda, apenas 229 alunos que concluíram o curso declararam sua renda. Dentre eles, é possível identificar renda mínima de até um salário mínimo e renda máxima de vinte ou mais salários mínimos, porém ao analisar cada quartil pode-se identificar que há uma variação entre as rendas declaradas. O primeiro e segundo quartil são formados por alunos extremamente pobres e alunos da baixa classe média. Mas mesmo com essa dificuldade financeira, os mesmos conseguiram concluir essa etapa da vida acadêmica. Já o terceiro quartil é formado por alunos que são da média classe média até alunos da alta classe alta. O desvio de mais de três salários mínimos nesse conjunto de dados, analisado em conjunto com os quartis, reforça que a maioria dos alunos que procuram estudar engenharia no IFCE-campus Fortaleza são de baixa renda e que felizmente nesse conjunto finalizaram essa etapa.

Sobre o atributo distância, nesse conjunto de alunos concludentes é caracterizado por morar próximo ao campus. O primeiro, o segundo e o terceiro quartil reforçam essa afirmação, sendo que até o segundo quartil a distância máxima é de 8,8 km, chegando até 13,2 km, que é onde começa o terceiro quartil. Para o desvio padrão apresentar o valor de 14,65, isso é resultante pela existência de alunos com residências próximas à 67,4 km. Em geral esse conjunto de alunos reside próximo ao campus e isso pode ser um fator motivador para conclusão do curso.

A idade média desses alunos evasores é de 25 anos, mas apresenta um desvio padrão de quase 7 anos, isso mostra que existem uma discrepância entre as idades desses alunos. Analisando o primeiro e segundo quartil, os alunos possuem idades entre 18 (mínimo) à 23 anos, podem ser considerados jovens discentes que evadiram o curso. Já para o terceiro quartil é composta pelos alunos que possuem entre 28 a 59 (máximo), essa representatividade é formada por alunos que possivelmente pretendiam se reinserir no mercado de trabalho, mas que infelizmente evadiram ao longo da caminhada acadêmica.

Os alunos evasores possuem IDHB entre 0,11 à 0,95, com uma média de 0,45 valor aproximado com a mediana que apresenta o valor de 0,42, mas ao analisar o primeiro e segundo quartil, pode caracterizar-se que uma parte dos alunos evasores são alunos com IDHB muito baixo e baixo, já a partir do terceiro quartil é exposto os alunos com IDHB alto e muito alto.

Sobre a renda dos alunos evasores, apenas 48 alunos de um total de 231 evasores declararam a renda familiar. A média salarial desses alunos é de 3,56 com um desvio padrão de 3,6 salários mínimos. Isso reflete que alguns alunos possuem uma renda muito baixa e

outros renda elevada. Assim, para os alunos concludentes, os alunos evasivos apresentaram renda mínima de 1 até 20 ou mais salários mínimos. Ao analisar o primeiro, segundo e terceiro quartil é possível identificar a caracterização desses alunos de extrema pobreza, média classe média e alta classe média que são os alunos que possui renda familiar igual ou superior a quatro salários mínimos. Basicamente, os alunos que evadiram são classificados como pobres, mas nesse cenário é exposto alunos que possuem uma renda alta e mesmo assim evadiram do curso.

Os alunos que evadiram o curso apresentam distâncias do campus a sua residência entre 1,3 a 67,4 km, com uma média de 12,63 km e um desvio padrão de 13,11 km. Ao analisar o primeiro, o segundo e o terceiro quartil, pode-se considerar que os discentes residem próximo ao campus. Essas distâncias superiores da apresentada no terceiro quartil pode ser consideradas como *outliers*, isso será visualizado nos gráficos de caixa.

Para analisar os atributos categóricos dos alunos concludentes e evadidos são apresentadas as tabelas de frequências (TF), com a frequência absoluta (FA), frequência relativa (FR) e a frequência total (FT). Essas TF são construídas para os atributos sexo, raça e origem. Na Tabela 6, pode ser conferido a TF dos alunos que concluíram e na Tabela 7 dos alunos evadidos.

Tabela 6 – TF: Atributo Sexo - Alunos não-evasores (Fa, Fr, M, F e FT)

Sexo	FA	FR (%)
Masculino	1063	80,65
Feminino	255	19,35
FT	1318	100

Fonte: Autor.

Tabela 7 – TF: Atributo Sexo - Alunos evasores (Fa, Fr, M, F e FT)

Sexo	FA	FR (%)
Masculino	195	84
Feminino	36	16
FT	231	100

Fonte: Autor.

As frequências relativas entre os alunos que concluíram e os alunos que evadiram são semelhantes, porém, os alunos do sexo feminino apresentaram frequência relativa inferior dentre todos os alunos que evadiram. Ao realizar uma comparação entre as duas tabelas, o percentual de alunos do sexo masculino, foram os que mais evadiram do curso.

Na Tabela 8 pode ser analisado a frequência do atributo raça para os alunos que concluíram o curso. Já na Tabela 9 é representado a frequência relativa dos alunos evadidos no atributo mencionado. Ao analisar essas tabelas, pode-se perceber que a frequência relativa entre pardos e brancos predomina nas duas tabelas, porém, vale ressaltar que os alunos que são considerados indígenas, possuem baixa frequência relativa na tabela dos alunos que concluíram e frequência relativa na tabela dos alunos que evadiram. Os alunos Que se declararam pardos apresentam uma frequência relativa bem próxima entre as duas classes, isto pode ser um indicativo para que o gestor observe com mais detalhes tais índices.

Tabela 8 – TF: Atributo Raça - Alunos não-evasores (Pardos, Brancos Pretos, Amarelos, Indígenas, Não declararam, FT, FA, FR)

Raça	FA	FR (%)
Pardos	681	51,67
Brancos	400	30,35
Pretos	67	5,08
Amarelos	14	1,06
Indígenas	3	0,23
Não declararam	153	11,61
FT	1318	100

Fonte: Autor.

Tabela 9 – TF: Atributo Raça - Alunos evasores (Pardos, Brancos Pretos, Amarelos, Indígenas, Não declararam, FT, FA, FR)

Raça	FA	FR (%)
Pardos	114	49,35
Brancos	79	34,20
Pretos	11	4,76
Amarelos	4	1,73
Indígenas	0	0,00
Não declararam	23	9,96
FT	231	100

Fonte: Autor.

As Tabelas 10 e 11 representam frequência do atributo origem dos alunos que concluíram e evadiram, respectivamente. Pode-se perceber que nas duas tabelas a maior frequência relativa são das escolas privadas e estaduais, porém, a categoria Filantrópica apresenta menor frequência relativa na tabela dos alunos concluídos, mas se for analisar a frequência relativa na tabela dos alunos evadidos esse percentual é maior, sugere-se que alunos oriundos de instituições Filantrópicas são mais propensos a evasão.

Tabela 10 – TF: Atributo Origem - Alunos não-evasores (Privada, Estadual, Federal, Municipal, Filantrópica, Outros, FT, FA e FR)

Origem	FA	FR (%)
Privada	573	43,47
Estadual	559	42,41
Federal	117	8,88
Municipal	19	1,44
Filantrópica	10	0,76
Outros	40	3,03
FT	1318	100

Fonte: Autor.

Tabela 11 – TF: Atributo Origem - Alunos evasores (Privada, Estadual, Federal, Municipal, Filantrópica, Outros, FT, FA e FR)

Origem	FA	FR (%)
Privada	91	39,39
Estadual	107	46,32
Federal	17	7,36
Municipal	1	0,43
Filantrópica	4	1,73
Outros	11	4,76
Frequência Total	231	100

Fonte: Autor.

Nas tabelas de contingências é possível identificar se os atributos categóricos podem influenciar na evasão escolar dos discentes. Inicialmente é analisado o atributo sexo com base nos dados da Tabela 12. O percentual de alunos do sexo masculino são os que mais evadem do curso de engenharia, essa informação reforça os dados coletados pelo censo IBGE (2010), em que corrobora que alunos do sexo feminino são mais propensas a conclusão do curso de nível superior.

Na Tabela 13 é possível identificar que os alunos da raça amarela são os que mais evadem do curso. Essa informação corrobora com os dados apresentados na pesquisa realizada pelo IBGE (2010), já que o percentual de conclusão de curso das pessoas dessa

Tabela 12 – Tabela de Contingência: Atributo Sexo (Masculino e Feminino), a Evasão (Sim e Não) e o Total

Sexo	Não	Evasão		%	Total
		%	Sim		
Masculino	1063	84,50	195	15,50	1258
Feminino	255	87,63	36	12,37	291
Total	1318	85,09	231	14,91	1549

Fonte: Autor.

raça é inferior das demais. A raça indígena ganha destaque nessa análise, em que não é evidenciado nenhuma evasão, porém, nos dados coletados pelo censo do IBGE (2010) não é exposto nenhum indígena que tenha concluído um curso de nível superior.

Tabela 13 – Tabela de Contingência: Atributo Raça (Branços, Pardos, Pretos, Amarelos, Indígenas e não declaram), a Evasão (Sim e Não) e o total)

Raça	Não	Evasão		%	Total
		%	Sim		
Branços	400	83,51	79	16,49	479
Pardos	681	85,66	114	14,34	795
Pretos	67	85,90	11	14,10	78
Amarelos	14	77,78	4	22,22	18
Indígenas	3	100	0	0,00	3
Não declaram	153	86,93	23	13,07	176
Total	1318	85,09	231	14,91	1549

Fonte: Autor.

Na Tabela 14 é possível analisar a contingência do atributo origem. Os alunos que são de origem Filantrópica são os que mais evadem, isso pode ser recorrente as dificuldades em que essas instituições enfrentam. Segundo Gouveia (2007), as instituições Filantrópicas enfrentam dificuldades de várias ordens, como identificar fontes de financiamento, elaborar propostas consistentes, captar recursos e gerir as organizações e seus projetos. Com todas essas dificuldades a evasão de discentes oriundos de instituições Filantrópicas, podem ser reflexos dessas dificuldades encontradas. Os alunos que menos evadem são de origem pública municipal, esse índice deve ser destacado pelos indicadores que são elencados pelo município, em que buscam fornecer um ambiente educativo, profissionais com boas práticas pedagógicas, avaliações de desempenho, uma gestão escolar democrática, formação e condições de trabalho para os professores, como também o acesso, permanência e sucesso na escola dos alunos (RIBEIRO; GUSMÃO, 2004).

Ao aplicar a equação do coeficiente de contingência para os atributos categóricos

Tabela 14 – Tabela de Contingência: Atributo Origem (Privada, Pública Federal, Pública Estadual, Pública Municipal, Filantrópica e Outros), a Evasão (Sim e Não) e o Total)

Origem	Não	Evasão		%	Total
		%	Sim		
Privada	573	86,30	91	13,70	664
Pública Federal	117	87,31	17	12,69	134
Pública Estadual	559	83,93	107	16,07	666
Pública Municipal	19	95,00	1	5,00	20
Filantrópica	10	71,43	4	28,57	14
Outros	40	78,43	11	21,57	51
Total	1318	85,09	231	14,91	1549

Fonte: Autor.

sexo, raça e origem, pode-se obter os resultados respectivamente 0,034, 0,047 e 0,069 para cada atributo supracitado. Na tabela de intensidade do coeficiente de contingência proposta por Costa (2015) é possível comparar o valor obtido da equação com a intensidade da relação entre as variáveis categóricas e a evasão. Nesse sentido todos os atributos categóricos apresentam intensidade fraca com a evasão, isso quer dizer que apenas um único atributo categórico não é suficiente para predizer a evasão escolar, mas o conjunto de atributos pode auxiliar a encontrar o perfil do aluno evasor.

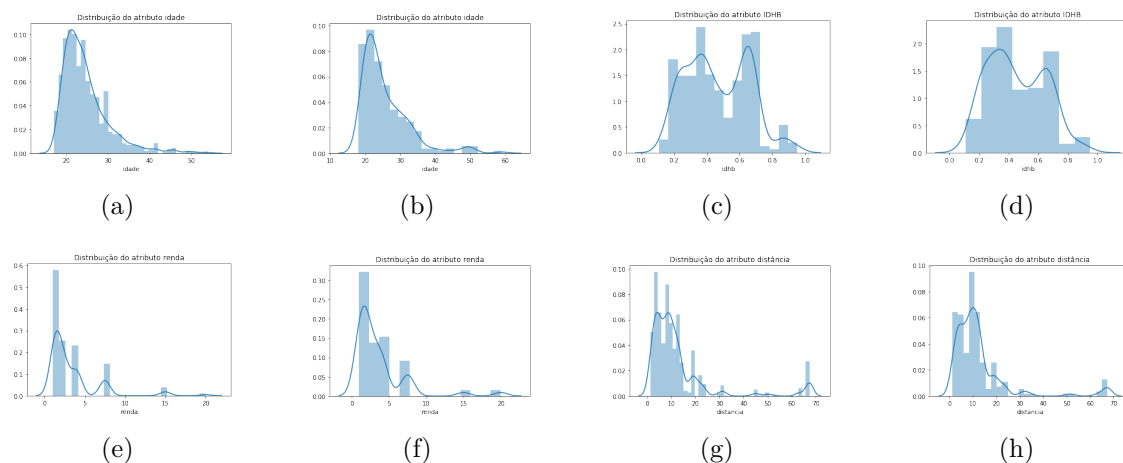
Ao final dessas análises entre os atributos categóricos e numéricos, sugere-se que o perfil mais propenso a evasão no curso de engenharia são o dos alunos de sexo masculino, entre 18 a 23 anos, sendo eles de raça amarela, com IDHB entre 0,11 a 0,38, renda familiar até 2,5 salários mínimos, oriundos de instituições filantrópicas e residindo com distâncias superior a 12,7 km.

Já para traçar o perfil menos propenso a evasão é caracterizados alunos do sexo feminino, com idades entre 21 a 27 anos, em que não declaram sua raça, tendo IDHB entre 0,63 a 0,95, sendo a renda familiar acima de 4 salários mínimos, oriundo da rede municipal de ensino e residindo próximo a instituição com distâncias entre 1,3 a 4,6 km.

Nesse momento, faz-se necessário analisar a amplitude, a frequência e a densidade dos atributos numéricos, ilustrado na Figura 6. Inicialmente, são analisadas essas informações dos alunos que não evadiram. Sobre o atributo idade, pode-se perceber que o gráfico está concentrado a esquerda, ou seja, a frequência de alunos que concluem o curso desse conjunto aponta para os mais jovens com idades entre 20 a 25 anos. Os alunos com idades superiores a 25 anos apresentam uma frequência inferior se comparados com idades abaixo dos 25. A amplitude desse atributo é de 37 anos. Com base nisso, pode-se perceber uma certa variação entre as idades dos alunos mais jovens (17 anos) em relação aos mais velhos

(54 anos).

Figura 6 – Histogramas dos alunos que concluíram (A (Idade), C (IDHB), E (Renda) e G (Distância)) e os que evadiram (B (Idade), D (IDHB), F (Renda) e H (Distância))



Fonte: Autor.

A densidade dos alunos que evadem o curso está distorcida à direita, com maior frequência entre os alunos que possuem idades de 20 a 30 anos, isso retrata que esses jovens são os que mais evadem do curso.

Ao analisar o histograma do atributo IDHB dos alunos que não evadiram, pode-se perceber que a distribuição desse atributo é bimodal, existindo dois picos que ganham destaque. A primeira concentração está entre os alunos com IDHB muito baixo e baixo, e a segunda concentração está entre os alunos com IDHB médio, com algumas frequências de alunos com IDHB alto e muito alto. A amplitude do atributo IDHB dos alunos que não evadiram é de 0,84. Esse valor apresenta uma variação muito extensa dentre esses alunos.

Sobre o IDHB dos alunos que evadiram do curso, pode-se perceber que sua densidade também é bimodal, ou seja existe uma grande concentração dos dados entre os alunos com IDHB muito baixo e baixo, já para os alunos que são classificados com IDHB alto e muito alto sua frequência é inferior comparada com as demais classes do IDHB. Isso retrata que a evasão escolar não está ligada apenas aos alunos com IDHB muito baixo e baixo.

O atributo renda dos alunos que não evadiram apresenta densidade multimodal, pois existem quatro picos que ganham destaque. A primeira concentração dos dados está entre os alunos que contém renda declarada de 1 a 4 salários mínimos, a segunda concentração é formada por alunos que possuem renda familiar aproximadamente de 10 salários e a terceira concentração são dos alunos que ganham mais de 15 salários, mesmo apresentando frequência inferior dos demais.

A renda familiar dos alunos que evadem o curso possui uma densidade multimodal.

O pico de evasão dos alunos ao averiguar a renda está entre os que possuem entre 1 a 4 salários mínimos, ou seja, os alunos que são considerados de extrema pobreza até os alunos da baixa classe média. O que pode gerar questionamento são os alunos que possuem renda familiar acima de 5 salários mínimos apresentar frequência de evasão escolar.

Para o atributo distância dos alunos que não evadiram, a densidade também é multimodal. Há existência de picos que retratam a proximidade do campus, em alguns casos um pouco mais distantes e em outros com distâncias superiores a 30 km. Basicamente os alunos que não abandonaram o curso residem próximo ao campus com distâncias entre 1,3 km a 18 km. Deste modo, pode-se perceber que quanto maior a distância menor será a frequência. Vale ressaltar que existem distâncias superiores a 60 km que apresentam frequências significativas ou consideradas *outliers*.

Os alunos que evadem, em sua grande maioria, residem próximo ao campus, perfil similar ao dos alunos que não evadiram. A densidade das distâncias dos alunos que evadiram é considerado multimodal, apresentando uma determinada frequência entre os alunos que residem até 9 km, como também alguns picos de frequência com distâncias entre 30 a 40 km e um pico de frequência entre 60 a 70 km. Os alunos que residem com distância superior a 9 km apresenta baixa frequência no histograma.

Até o momento pode ser perceptível que existem atributos com *outliers* (valores discrepantes). Afim de discutir sobre esse assunto e verificar a existência dos mesmos dentro do conjunto de alunos que não evadiram e os que evadiram do curso, na Figura 7 é ilustrado os *boxplots* dos atributos Idade, IDHB, Renda e Distância. Para os alunos que concluíram o curso, é notória a faixa etária das idades. Nesse conjunto estudado, os alunos que não evadem estão entre 19 a 35 anos. Os alunos que apresentam idades superiores a 35 anos são considerados *outliers*. Isso pode dificultar no treinamento do classificador, pois, o mesmo tentará encontrar a melhor função que represente essa classe. Os alunos que possuem mais de 34 anos, assim como os alunos que possuem menos de 21 anos serão removidos do conjunto de alunos que não evadiram, pois alunos com idades inferiores a 21 anos não tem como concluir o curso de engenharia em menos de 5 anos. Caso ele siga o padrão regular do calendário da instituição e a matriz curricular, isto é, sem adiantar disciplina.

Os alunos que evadem do curso possuem idades entre 20 a 30 anos, isso representa 50 % desse conjunto. Para os alunos que possuem idades superiores a 39 anos, são considerados *outliers*. Neste caso todos os valores discrepantes serão removidos. A base dos alunos evadidos conterà apenas alunos com idades inferiores a 39 anos.

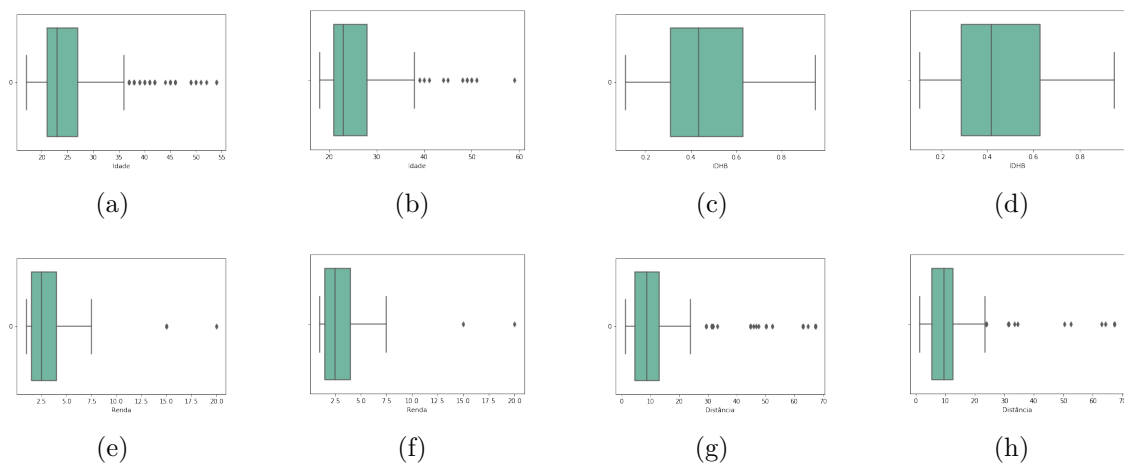
No que diz respeito aos alunos que não evadiram do curso, em relação ao atributo IDHB, mesmo esse atributo sendo multimodal, como estudado nos histogramas dos alunos que não evadiram, o mesmo não apresenta valores discrepantes. A metade desses alunos possui IDHB entre 0,31 a 0,63 que são classificados como discentes de IDHB baixo e médio.

Assim como no atributo IDHB dos alunos que não evadiram, o conjunto dos alunos evadidos não apresentam *outliers* nesse atributo, porém, a classificação desses alunos sobre o IDHB ao utilizar o primeiro, o segundo e o terceiro quartil refere-se ao IDHB muito baixo ao médio. Vale salientar que o IDHB é composto por escolaridade, longevidade e renda. Se o IDHB desses alunos está entre muito baixo e médio, pode-se subentender que as condições dos mesmos não são tão favoráveis para dar continuidade a vida acadêmica.

A renda familiar dos alunos que não abandonaram o curso está concentrada entre 1 a 4 salários mínimos. Esses discentes são pertencentes ao grupo de alunos considerados pobres a alunos da baixa classe média. Para esse atributo é encontrado alguns *outliers* a partir de 7,5 salários mínimos. Isso denota que alunos da classe alta iniciam e concluem o curso no IFCE-campus Fortaleza. Neste cenários serão removidos dessa base todos os alunos que possuem renda familiar acima de 7,5 salários mínimos.

A concentração da renda dos alunos que evadem está entre 1 a 4 salários mínimos, isso é reforçado com base no IDHB, pois, corrobora com o conceito do atributo. Existem alunos com renda familiar muito alta e eles são considerados valores discrepantes nesse conjunto, isso pode gerar questionamento. O que leva um aluno com boas condições financeiras a abandonar um curso de engenharia em uma instituição federal?

Figura 7 – *Boxplot* dos alunos que concluíram (A (Idade), C (IDHB), E (Renda) e G (Distância)) e os que evadiram (B (Idade), D (IDHB), F (Renda) e H (Distância))



Fonte: Autor.

Neste caso, alunos com renda familiar maior do que 7,5 salários mínimos serão removidos do conjunto de dados dos alunos evasores.

Os alunos que não evadiram do curso apresentam pelo histograma que residem próximo ao campus, mas para ser avaliado a existência de *outliers* é averiguado o *boxplot*. Ao analisar a metade dos dados pela divisão de cada quartil, pode-se perceber que os alunos residem próximo ao campus. Quando o aluno reside com distância superior a

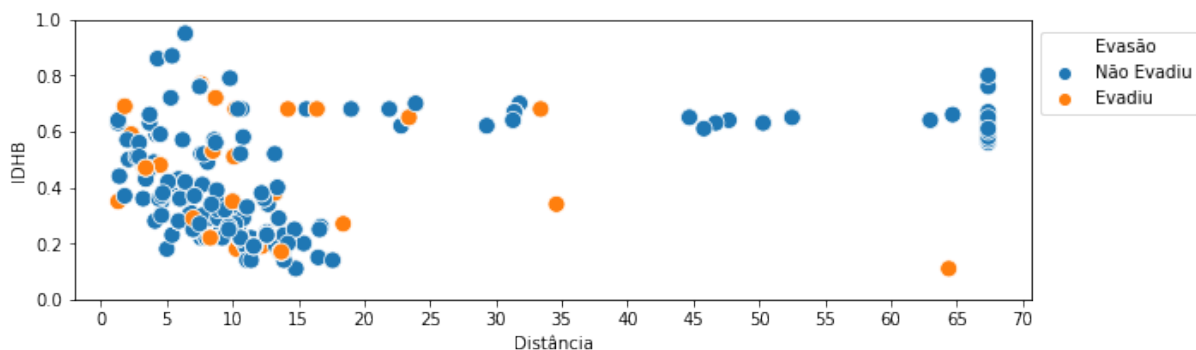
aproximadamente 25 km, esses discentes já são considerados *outliers*. Esse atributo é o que apresenta mais *outliers*, e isso pode tornar mais difícil o treinamento do classificador que é proposto neste trabalho. Neste caso, serão removidos os registros que apresentam distâncias superiores a 25 km.

Ao analisar metade do conjunto dos discentes que evadiram, pode-se perceber no *boxplot* que os discentes evadidos residem próximo ao campus, e essa distância fica entre 5,4 a 12,7 km. Para os alunos que residem com distâncias superiores a aproximadamente 23 km já são considerados *outliers*, essa comparação se tornar semelhante aos alunos que concluíram o curso. Para esses conjuntos serão removidos registros com distâncias superiores a 23 Km.

Analisar gráficos de caixa (*boxplot*) pode remeter diversas averiguações que são relevantes para o conjunto. Entender a distribuição dos dados e se existem valores discrepantes do normal do conjunto pode favorecer ainda mais na tomada de decisões no conjunto de dados. Nesse cenário exposto até o momento, os alunos que evadem e os alunos que concluem possuem perfis muito parecidos, até mesmo com relação a distribuição dos dados.

A fim de traçar uma linha que possivelmente separará os alunos evasivos dos não evasivos, são gerados gráficos de dispersão para averiguar tal fato. Inicialmente são analisados os atributos distância e IDHB, ilustrado na Figura 8. Como pode ser analisado na ilustração supracitada, aparentemente não há como traçar uma linha capaz de separar os alunos evasivos dos não evasivos. É notório o aglomerado de alunos com distâncias inferiores a 20 km e com IDHB entre 0,2 a 0,6. Nessa análise à torna-se inviável uma separação linear, impossibilitando a divisão entre os alunos.

Figura 8 – Gráfico dispersão: Distância (entre 0 à 70) x IDHB (entre 0 à 1), na cor azul os que não evadiram e em laranja os que evadiram

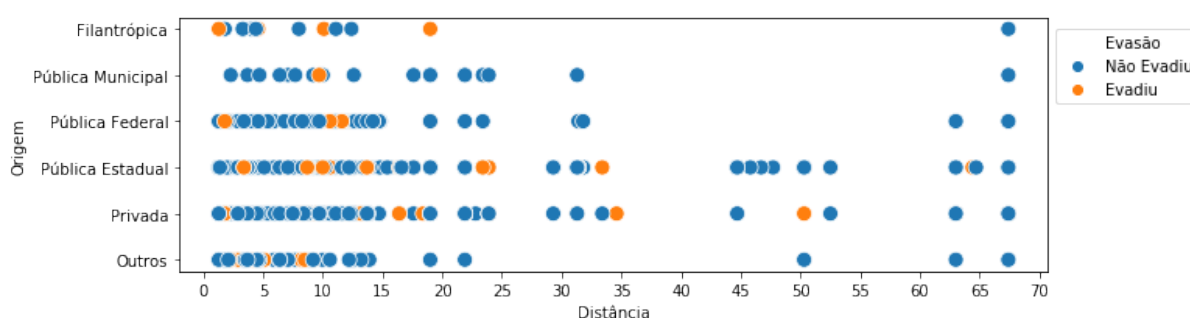


Fonte: Autor.

Ao gerar o gráfico de dispersão para os atributos distância e origem, ilustrado na Figura 9, torna-se inviável traçar uma linha separável entre as classes, tendo em vista

que muitos dos discentes evadidos estão bem próximos dos não evadidos. A incidência de alunos evasivos é decorrente de todas as origens, sugere-se que, indiferente da instituição de origem, os alunos podem evadir do curso.

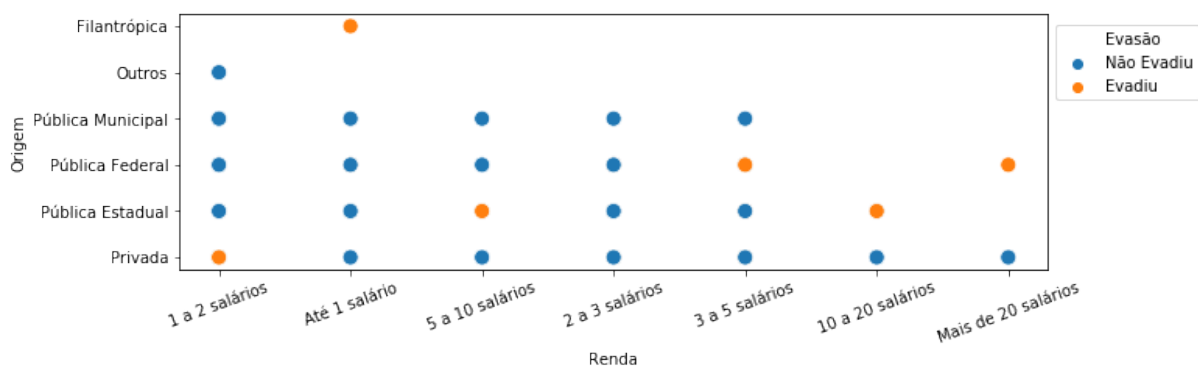
Figura 9 – Gráfico dispersão: Distância (entre 0 à 70) x Origem (Filantrópica, Pública Municipal, Pública Federal, Pública Estadual, Privada e Outros), na cor azul os que não evadiram e em laranja os que evadiram



Fonte: Autor.

Ao analisar o gráfico de dispersão dos atributos renda e origem, ilustrado na Figura 10, pode-se perceber que alunos que são oriundos da rede privada, mas que a renda familiar é entre 1 a 2 salários mínimos, apresentam evasão escolar. Os alunos que são de instituições filantrópicas e ao mesmo tempo possuem renda familiar de até 1 salário mínimo é recorrente na evasão escolar. Para os alunos que apresentam rendas superiores a 5 salários mínimos, oriundos de instituições federais ou estaduais, é possível surgir um questionamento do porquê da evasão desses discentes. Ao final, sugere-se que utilizar apenas a origem e a renda para tentar separar os discentes evasivos dos não evasivos não é possível.

Figura 10 – Gráfico dispersão: Renda (de 1 Salário mínimo à mais de 20 salários) x Origem (Filantrópica, Pública Municipal, Pública Federal, Pública Estadual, Privada e Outros)

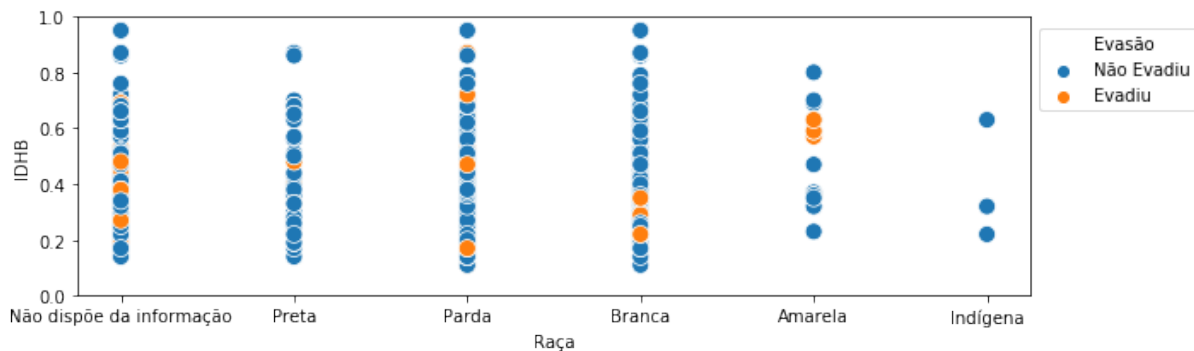


Fonte: Autor.

Na Figura 11 é analisado o gráfico de dispersão em uma dimensão 2, entre os atributos raça e IDHB, onde é possível perceber que para todas as raças, com exceção da

raça indígena é apresentado evasão escolar. Assim como os outros gráficos de dispersão apresentados até o momento, é impossível traçar um linear capaz de separar alunos concludentes dos evasivos utilizando apenas os atributos raça e IDHB.

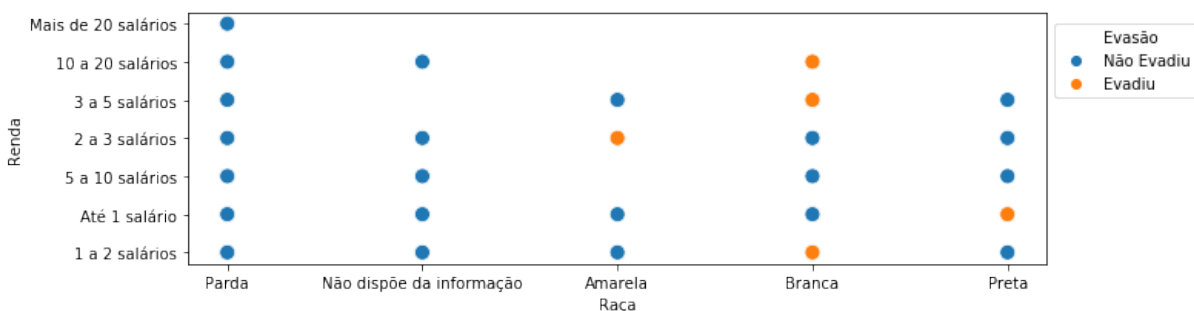
Figura 11 – Gráfico dispersão: Raça (Branços, Pardos, Pretos, Amarelos, Indígenas e não declaram) x IDHB (entre 0 à 1)



Fonte: Autor.

Construir um gráfico com os atributos raça e renda, como ilustrado na Figura 12, pode-se perceber que existem algumas raças que sobressaem no quantitativo da evasão escolar, como é o caso da raça branca, preta e amarela. Isso não descarta o fato que existem evasões nas outras raças, sendo as mesmas já discutidas anteriormente. O que é destacado nessa análise é a dificuldade de traçar uma linha capaz de separar as duas classes. Isso reforça que o problema da evasão escolar, não é um problema linear e que se faz necessário utilizar de outros meios para tentar prever esse fato.

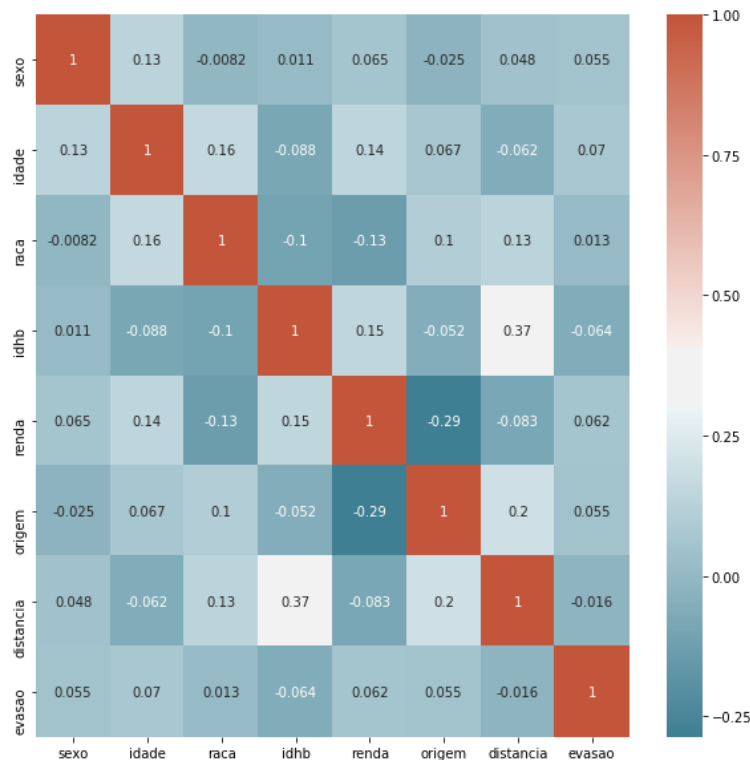
Figura 12 – Gráfico dispersão: Raça (Branços, Pardos, Pretos, Amarelos, Indígenas e não declaram) x Renda (de 1 Salário mínimo à mais de 20 salários)



Fonte: Autor.

Predizer se o aluno é um possível evasor, é uma tarefa complexa. Ao aplicar a estatística descritiva, a construção e a análise de gráficos, a análise da centralidade e discrepância dos dados reforçaram que essa predição não é uma tarefa fácil. Afim de tentar identificar a correlação entre todos os atributos é gerado um gráfico de correlação de Pearson ilustrado na Figura 13.

Figura 13 – Correlação de Pearson entre os atributos, sexo, idade, raça, IDHB, renda, origem, distância e evasão



Fonte: Autor.

Ao analisar os atributos distância e IDHB são os que apresentam maior relação positiva 0,37, seguidos dos atributos distância e origem 0,2. Para as relações negativas (inversas), os atributos renda e origem apresentam relação de -0,29, seguidos dos atributos raça e renda com relação de -0,13. Para os atributos raça e IDHB a correlação entre eles é de -0,1. Os demais atributos apresentaram correlações mais fracas das mencionadas. Isso denota que apenas um único atributo, não é capaz de prever se o aluno irá evadir, mas a combinação desses atributos poderá traçar um perfil desse aluno.

Ao utilizar os métodos tradicionais, não foi possível separar alunos evasores de alunos concluintes, sendo assim na próxima seção serão utilizados algoritmos computacionais, a fim de apresentar um modelo capaz de prever com o mínimo de erros possíveis se o aluno evadirá ou não do curso. Após todas as remoções dos *outliers*, o conjunto de dados contém 1071 registros, sendo eles 869 e 202, alunos que concluíram e alunos evasores, respectivamente.

3.2 Resultados dos classificadores de Aprendizagem de Máquina

Esta seção apresenta os resultados obtidos usando diferentes classificadores. O conjunto de dados utilizado consiste em 1071 registros socioeconômicos, dos quais 869

referem-se a alunos que concluíram o curso e 202 corresponde a alunos que evadiram. A base foi dividida para que a mesma quantidade de dados de treinamento de cada classe fosse distribuída igualmente. Em seguida, 80 % dos registros socioeconômicos dos alunos concluintes e desistentes foram utilizados e o restante das amostras foi utilizado para testes. O conjunto de dados foi normalizado entre -1 e 1.

O hardware usado para teste no servidor é um computador com um processador Intel Core-i7 com 2,5 GHz, 12 GB de RAM e sem GPU (Unidade de processamento gráfico).

A parametrização dos classificadores utilizados, bem como os resultados obtidos, são apresentados nas seções a seguir.

Todas as técnicas de aprendizagem de máquina deste trabalho passaram por 50 treinamentos e validação cruzada com 10-*folds*. Para o classificador perceptron é utilizado taxa de aprendizagem entre 0,01 a 0,5. O classificador árvore de decisão é treinado com o parâmetro *criterion entropy* e *gini*. Já o classificador SVM foram definidos *kernel: linear, poly, Radial Basis* e *sigmoid*, para o parâmetro de regularização é atribuído uma variação de 0,1 a 10. O hiperparâmetro número de vizinhos do KNN variou de 3 a 11, selecionado apenas os números ímpares. Os classificadores MLP e *Deep Learning* foram treinados com função de ativação *sigmoid*, tangente hiperbólica e relu, a função de otimização para ambos é o Adam e a taxa de aprendizagem variou de 0,1 a 0,5, em relação a quantidade de neurônios, a rede neural MLP inicialmente começou com 5 sugestão do Haykin (2001) a 65 e a rede *Deep Learning* foi treinado inicialmente com 5 a 250 neurônios entre 1 a 4 camadas ocultas. Este trabalho apresenta apenas os melhores resultados de cada configuração.

Na Tabela 15 é ilustrado os resultados da predição usando o conjunto de dados tratado com os classificadores selecionados neste trabalho. A identificação da não linearidade do problema é dada pelos resultados obtidos do classificador perceptron, enfatizando na complexidade da predição. É evidente que a Árvore de Decisão obteve os melhores resultados em todas as métricas analisadas. Já o classificador *Deep Learning* obteve o segundo melhor resultado com uma diferença de 1,57 %, 1,58 %, 0,02 % e 3,13 % em Acc, *Fscore*, Pre e Rec respectivamente. Além disso, a Árvore de Decisão e o *Deep Learning* foram os únicos classificados que obtiveram resultados superiores a 90 % em todas as métricas.

Regressão Logística e SVM com *kernel sigmoid* apresentaram resultados ruins e, portanto, não são as melhores opções para lidar com os dados apresentados. Além disso, o KNN e o MLP apresentaram bons resultados, atingindo mais de 90 % em várias métricas.

Segundo Bezerra et al. (2016) a evasão escolar na educação pública brasileira é um problema de grandes proporções. Um em cada quatro brasileiros deixa a escola prematuramente, antes de terminar o ensino médio. A pesquisa desenvolvida por Bezerra

Tabela 15 – Resultado da predição de alunos com risco de abandono escolar usando dados socioeconômicos aplicando técnicas de aprendizagem de máquina, considerando *Recall* (Rec), *Precision* (Pre), *F score* e *Accuracy* (Acc) como métricas de avaliação

Classificador	Configuração	Rec(%)	Pre(%)	F score(%)	Acc(%)
Pecerptron		60,23 $\pm$ 0,02	56,79 $\pm$ 0,03	57,92 $\pm$ 0,02	56,25 $\pm$ 0,02
Árvore de Decisão	Entropy	100,0 $\pm$ 0,33	98,67 $\pm$ 0,22	99,33 $\pm$ 0,30	99,34 $\pm$ 0,30
Regressão Logística		58,33 $\pm$ 0,23	55,60 $\pm$ 0,03	56,93 $\pm$ 0,33	55,87 $\pm$ 0,23
SVM	Sigmoid	64,77 $\pm$ 0,01	59,38 $\pm$ 0,02	61,96 $\pm$ 0,02	60,23 $\pm$ 0,02
KNN	k = 3	84,82 $\pm$ 0,05	92,16 $\pm$ 0,05	88,34 $\pm$ 0,05	88,80 $\pm$ 0,05
MLP		88,65 $\pm$ 0,09	93,51 $\pm$ 0,09	91,02 $\pm$ 0,09	91,28 $\pm$ 0,09
Deep Learning		96,87 $\pm$ 0,07	98,65 $\pm$ 0,07	97,75 $\pm$ 0,07	97,77 $\pm$ 0,07

Fonte: Autor.

et al. (2016) faz um recorte do problema, ao analisar a evasão escolar no último ano do ensino fundamental nas escolas públicas estaduais e municipais do estado de Pernambuco, com base nos dados dos Censos Escolares 2011 e 2012. Utilizando Árvore de Decisão os resultados mostraram que fatores como idade, turno das aulas e região geográfica das escolas influenciam fortemente a evasão. O modelo proposto pelos autores atinge 69 % na acurácia dos alunos evadidos. Sobre as características em comum com este trabalho é destacado os atributos idade e sexo e os resultados com o mesmo classificador sobressai em 30,34 %. A acurácia obtida com esse classificador em relação a base de estudo deste trabalho reforça a viabilidade deste modelo.

Os pesquisadores Calixto, Segundo e Gusmão (2017) propõe um trabalho que tem como objetivo a construção de um estudo para identificar as variáveis concernentes à evasão escolar, utilizando os dados do censo educacional de 2014, 2015 e 2016 dos estados de Ceará e Sergipe. A metodologia *CRoss – Industry Standard Process for Data Mining* (CRISP-DM) foi usada para entender, preparar e modelar os dados. Na separação dos dados foi utilizada a ferramenta *Statistical Package for the Social Sciences* (SPSS). As análises se deram por meio do SPSS e do *RapidMiner*, ferramentas desenvolvidas para preparação de dados, visualização de resultados, validação e otimização de modelos (HOFMANN; KLINKENBERG, 2013). Os autores selecionaram a técnica Regressão Logística que tem como vantagens prever condições futuras, dar suporte quantitativo ao julgamento dos gerentes, apontar falhas no pensamento gerencial e fornecer novos *insights*. O modelos criados apresenta acurácia em torno de 87 %. Além disso, a idade, etapa de ensino, modalidade de ensino, existência de laboratórios e localização da escola se destacaram como variáveis influentes na evasão escolar. Dentre essas variáveis é comum a este trabalho apenas o atributo idade, e utilizando esse classificador na base em estudo os resultados foram ruins com uma diferença negativa de 31,13 %.

Sorensen (2016) afirma que em uma era de análise de dados dos estudantes, o

potencial total de uso desses dados ainda não foram totalmente alcançados quando o assunto é direcionar recursos para os alunos na redução da evasão. O estudo explora a utilidade das técnicas de aprendizado de máquina com dados administrativos de larga escala para identificar o risco de evasão de alunos. Usando dados de registros longitudinais de alunos do Departamento de Instrução Pública da Carolina do Norte, essa pesquisa avalia técnicas modernas de previsão, com foco em métodos de classificação baseados em máquinas de vetores de suporte. Esses métodos incorporam medidas preditivas da 3ª à 8ª série, incluindo desempenho acadêmico, indicadores comportamentais e características socioeconômicas e demográficas. Os atributos selecionados pelo o autor foram: sexo, raça, idade, almoço gratuito na escola, nível de educação dos pais, nível de leitura, nível de escrita, média na disciplina de matemática, frequência escolar, indisciplina, uso de eletrônicos e se o aluno trabalha. No estudo desenvolvido por Sorensen (2016) o algoritmo SVM por ser um classificador que consegue lidar bem com grandes conjuntos de amostras, por tratar bem dados de alta dimensão e por ter um processo de classificação rápido, alcançando acurácia de 90,8 %, sendo esse resultado bem expressivo. Ao comparar as variáveis em comum com esta pesquisa é destacado os atributos sexo, raça e idade, porém os resultados obtidos foram pífios, sendo o melhor resultado do SVM proposto 60,23 % na acurácia, utilizando *kernel sigmoid*.

Aulck et al. (2016) declaram que é investido milhões na educação norte americana, e que na ocorrência do abandono escolar esse investimento é desperdiçado. Os autores utilizaram as características socioeconômicas sexo, raça, origem e se o aluno possui residência. A técnica de aprendizagem de máquina aplicada para tentar prever a evasão escolar é o algoritmo KNN, onde obteve acurácia de 64,60 %. As características em comum com este trabalho são: sexo e raça, o atributo origem não é contemplado pois se trata se o discente é hispânico ou não, essa associação não faz menção a origem escolar, como é o caso da abordagem deste trabalho. Em relação ao resultado o modelo KNN treinado nesta pesquisa apresentou resultados mais expressivos do que o referido autor, com uma acurácia de 88,80 %, este modelo apresenta características que podem ser consideradas satisfatórias no que tange a predição da evasão escolar, já que em outras métricas também obteve bons resultados.

Karamouzis e Vrettos (2008) o percentual de abandono dos estudantes da graduação é um problema significativo e crescente no ensino superior. Os alunos estão abandonando as faculdades por vários motivos e os administradores das escolas estão se esforçando para reduzir as taxas de evasão. Prever o risco de potencial evasão na graduação é de grande valor para as universidades e um enorme potencial na utilidade de intervenções direcionadas. O autor considera o comportamento das Redes Neurais Artificiais como classificadores promissor, direcionando-os para o desenvolvimento, treinamentos e testes de uma MLP para prever as características de evasão de alunos da graduação. A rede foi desenvolvida com 11 neurônios na camada oculta usando a *backpropagation* como

otimizador. As características selecionadas foram raça, sexo, opção do curso na faculdade, idade, escola de origem, deficiente, atendimento especial, endereço do aluno, cidade de origem, idade em que concluiu o ensino médio. Com essas características a rede proposta por Karamouzis e Vrettos (2008) obteve 68 % de acurácia. As características em comum são: raça, sexo, idade e escola de origem, com o modelo proposto neste trabalho obteve acurácia de 91,28 %, tornando-o esse modelo mais viável a utilização no auxílio a tomada de decisões.

Nagy e Molontay (2018) afirma que prever o desempenho do aluno, prevenir falhas e identificar os fatores que influenciam o abandono do aluno são questões que atraíram muito interesse de pesquisa recentemente. Nesse estudo, é empregado e avaliado o classificador *Deep Learning* para identificar alunos em risco e prever a desistência em cursos de nível superior com base nos dados disponíveis no momento da matrícula, selecionando atributos que envolvem o desempenho do estudante no ensino médio e detalhes pessoais como sexo, idade e endereço. O modelo proposto por Nagy e Molontay (2018) obteve acurácia de 73,5 %, já o modelo proposto neste trabalho obteve acurácia de 97,77 % utilizando *Deep Learning*.

Este trabalho também comparou o método proposto com diferentes métodos atualizados, porém o conjunto de dados utilizado é o selecionado do sistema Q-Acadêmico do IFCE campus Fortaleza, essa comparação pode ser conferida na Tabela 16.

Tabela 16 – Comparação dos classificadores com métodos recentes usando conjuntos de dados diferentes e acurácia (Acc) como métrica de avaliação

Classificador	Método	Acc (%)
Árvore de decisão	Método proposto	99,34
	Bezerra et al. (2016)	69,00
Regressão Logística	Método proposto	55,87
	Calixto, Segundo e Gusmão (2017)	87,00
SVM	Método proposto	60,23
	Sorensen (2016)	90,08
KNN	Método proposto	88,80
	Aulck et al. (2016)	64,60
MLP	Método proposto	91,28
	Karamouzis e Vrettos (2008)	68,00
Deep Learning	Método proposto	97,77
	Nagy e Molontay (2018)	73,50

Fonte: Autor.

O que ganha destaque deste trabalho, são os resultados obtidos nos classificadores utilizados para prever o risco de evasão escolar dos discentes. Ao realizar uma comparação do método proposto em relação aos métodos dos referidos autores, os resultados desta

pesquisa na métrica acurácia sobressai em relação em boa parte das propostas, isso deve ao fato deste trabalho realizar uma análise exploratória dos dados, compreendendo assim, a normalidade dos dados, sua distribuição e o que caracteriza o perfil do aluno evasor. Utilizar os atributos sexo, idade, raça, IDHB, renda, Origem e DCR foram características que reforçaram ainda mais os resultados obtidos, já que as outras propostas não utilizaram as mesmas entradas e em alguns casos apenas parte delas. Na Tabela 17 é possível observar as características selecionada de cada pesquisa citada.

Tabela 17 – Relação dos atributos previsores de cada autor referenciado neste trabalho como objeto comparativo entre os atributos selecionados

Método	Atributos
Método proposto	Sexo Idade Raça IDHB Renda Escola Origem DCR
Bezerra et al. (2016)	Idade Turno Região geográfica
Calixto, Segundo e Gusmão (2017)	Idade Etapa de ensino Modalidade de ensino Existência de laboratório Localização da escola
Sorensen (2016)	Sexo Raça Idade Almoço na escola Escolaridade dos pais Nível de Leitura Nível de escrita Média em matemática Frequência Eletrônicos em sala Aluno trabalha
Aulck et al. (2016)	Sexo Raça Escola de origem Aluno residente Aluno hispânico
Karamouzis e Vrettos (2008)	Sexo Raça Opção do curso Idade Escola de origem Aluno deficiente Atendimento especial Distância casa campus Cidade de origem Idade de conclusão do ensino médio
Nagy e Molontay (2018)	Sexo Idade Distância casa campus

Fonte: Autor.

Neste trabalho foram selecionadas as características da ficha de pré-matrícula do Sistema Q-Acadêmico do IFCE, os atributos previsores são: Sexo possuindo duas categorias, masculino e feminino; Idade sendo um atributo com valor mínimo e máximo de 17 e 54 respectivamente; Raça com 6 categorias, Brancos, Pardos, Pretos, Amarelos, Indígenas e Não declaram; Índice de Desenvolvimento Humano por Bairro sendo esse atributo numérico com valor mínimo de 0,11 e máxima de 0,95; Renda sendo um atributo categórico com 7 categorias, 0 a 1 salário mínimo, 1 a 2 salários mínimos, 2 a 3 salários mínimos, 3 a 5 salários mínimos, 5 a 10 salários mínimos, 10 a 20 salários mínimos e + 20 salários mínimos; Escola de origem possui 5 categorias, Privada, Estadual, Municipal, Filantrópica e Outras; e por último o atributo Distância Campus Residência que é uma característica numérica tendo como valor mínimo 1,3 km e máximo de 64,4 km.

Bezerra et al. (2016) desenvolveu sua pesquisa selecionando apenas 3 características nos dados dos Censos Escolares 2011 e 2012 das escolas públicas estaduais e municipais do estado de Pernambuco, sendo elas: idade como um atributo numérico com valor mínimo de 14 anos e máxima de 19 anos; Turno sendo um atributo categórico com apenas 3 categorias, manhã, tarde e noite; e por último o atributo região geográfica com as categorias, Órgão Regional INEP, Ensino médio tradicional, Permanência diária na escola, Início das aulas até 13h20 e Início das aulas após 13h20.

Calixto, Segundo e Gusmão (2017) avançou em suas pesquisas usando 5 características previsoras, sendo elas: Idade sendo uma característica numérica; Etapas de ensino sendo uma característica categórica com as categorias, 1º ano do ensino médio, 2º ano do ensino médio e 3º ano do ensino médio; Modalidade de ensino também é uma característica categórica com duas categorias, tradicional e educação de jovens e adultos; Existência de laboratório de informática é uma característica categórica sendo um atributo dicotômico, sim e não; e por fim localização da escola também sendo uma característica dicotômica, mas neste caso, os valores são rural e urbana.

Sorensen (2016) dentre os pesquisadores citados desenvolveu sua pesquisa com mais características das demais, totalizando em 11 atributos, sendo eles: Sexo característica dicotômica com as categorias, masculino e feminino; Raça sendo uma característica categórica com as categorias, Preto, Branco Amarelo e outra raça; Almoço gratuito na escola sendo uma característica binária; Escolaridade dos pais é uma característica categórica com 6 níveis, sendo eles: nível 3 a nível 8, o nível 3 sendo o mais baixo e o nível 8 sendo o mais alto; Nível de Leitura também é uma variável categórica com 6 categorias, nível 3 a nível 8, o nível 3 sendo o mais baixo e o nível 8 sendo o mais alto; Nível de escrita possui as mesmas categorias da característica anterior; Média em matemática é um atributo numérico do tipo real; Frequência é um atributo numérico do tipo real que representa a presença total do aluno no referido ano; Eletrônico em sala é uma característica dicotômica, os valores são sim e não; Aluno trabalha é uma característica semelhante a anterior.

Aulck et al. (2016) utilizou apenas 5 características, sendo elas: Sexo; Raça; Escola de origem; Aluno residente; Aluno hispânico. As características que são diferentes das demais citadas é aluno residente que é binária, assim como aluno hispânico. Já o atributo escola de origem difere pelas categorias que compõem esse atributo.

Karamouzis e Vrettos (2008) usando 10 atributos desenvolveu sua pesquisa, porém, as características Sexo, Raça, Idade, Escola de origem, Distância casa campus e cidade de origem, já foram mencionadas nas pesquisas anteriores. Já para o atributo Opção do curso é do tipo numérico e possui apenas dois valores 1 e 2, o atributo Aluno deficiente é um atributo dicotômico com os valores sim e não, assim como atendimento especial. O atributo mais distinto dos demais é a característica idade que concluiu o ensino médio que é do tipo numérico, declarando apenas a idade em anos.

Nagy e Molontay (2018) utilizou apenas 3 características, sendo elas: Sexo, Idade e Distância casa campus, as suas descrições já foram mencionadas em pesquisas anteriores.

Utilizando o classificador árvore de decisão, é obtido na acurácia 99,34 % no método proposto e no trabalho desenvolvido por Bezerra et al. (2016) obteve 69 %, porém os autores utilizaram apenas três características: idade, turno e região geográfica. Pode-se subentender que as características selecionadas não foram suficiente para obter uma acurácia aproximada do método proposto deste trabalho, contudo, o resultado expressivo desse classificador no método proposto pode ressaltar a relevância das características selecionadas.

O classificador regressão logística segundo a proposta desenvolvida por Calixto, Segundo e Gusmão (2017) obteve acurácia de 87 %, utilizando as características: idade, etapa de ensino, modalidade de ensino, existência de laboratório e localização da escola. Já no método proposto nesta pesquisa não obteve resultados superiores ao dos referido autores, atingindo apenas 55,87 % na acurácia. Vale ressaltar que existe apenas uma característica em comum. Fatores como modalidade de ensino e existência de laboratório na escola pode ser algo que implique diretamente na evasão escolar e componha um conjunto de fácil classificação.

Para o classificador SVM, a pesquisa desenvolvida por Sorensen (2016) obteve 90,08 % na acurácia, resultado esse bem superior ao método proposto neste trabalho que obteve 60,23 %. Pode-se observar que Sorensen (2016) utilizou os atributos: sexo, raça, idade, se possui almoço na escola, o nível de educação dos pais, o nível de leitura e escrita do aluno, a média do aluno em matemática, a frequência, a utilização de eletrônicos em aula e se o aluno trabalha. Fatores como nota em disciplinas e a possibilidade do aluno trabalhar podem ter sido características que reforçaram a evasão escolar do discente e o autor apresentar bons resultados.

Aulck et al. (2016) considerou relevante utilizar apenas os atributos: sexo, raça e

origem em que obteve 64,60 % de acurácia com o classificador KNN, resultado esse bem inferior ao método proposto neste trabalho, sendo com o mesmo classificador obteve 88,80 %, porém, vale ressaltar que neste trabalho foram utilizadas outras características sendo elas usadas em conjunto apresentaram melhores resultados do que o referido autor.

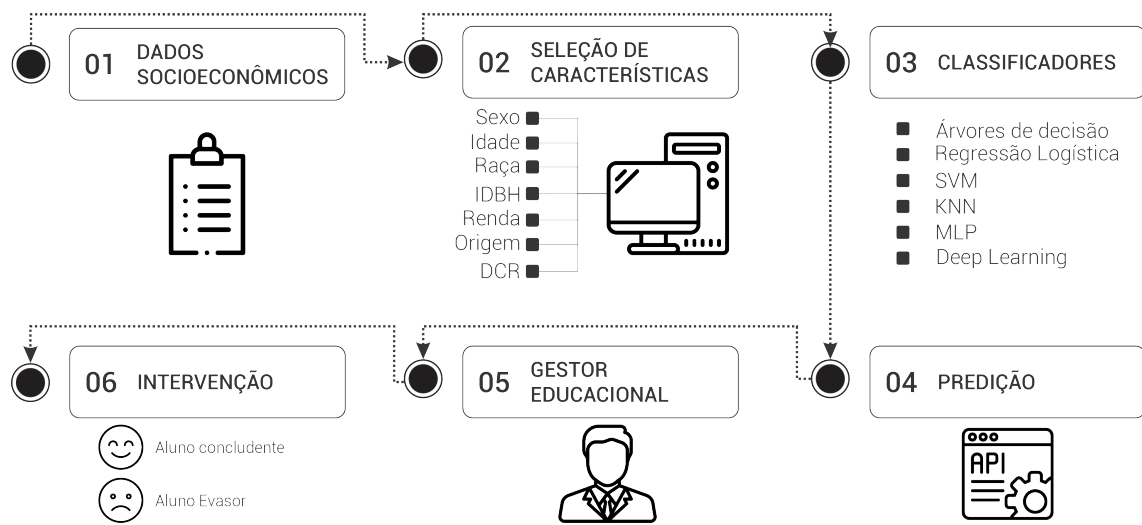
Segundo Karamouzis e Vrettos (2008) as características: raça, sexo, opção do curso, idade, escola de origem, se o aluno é deficiente, se o discente necessita de atendimento especial, endereço do aluno, cidade de origem e idade que concluiu o ensino médio seriam suficientes para prever o risco de evasão escolar, utilizando o classificador MLP obteve 68 % de acurácia, já no método proposto neste trabalho obteve 91,28 %, compreender a distribuição dos dados e as características que podem montar um modelo capaz de prever o risco de evasão escolar, são esforços valiosos para alcançar bons resultados no modelo, essas ações são destaques que não são abordadas no trabalho desenvolvido por Karamouzis e Vrettos (2008), mas que são tratados neste trabalho.

Já para Nagy e Molontay (2018) utilizar apenas as características sexo, idade e endereço seriam suficientes para prever a evasão escolar com o classificador *Deep Learning*, porém os mesmos conseguiram atingir 73,50 % na acurácia, no método proposto neste trabalho com o mesmo classificador é obtido 97,77 % na métrica acurácia. Esse resultado pode ser destacado pela análise exploratória dos dados, a remoção de amostras que foram consideradas *outliers* e outros tratamentos que foram citados no decorrer da pesquisa, além do mais, a quantidade de característica selecionada pode ter sido um fato impactante nessa predição.

4 Estrutura IoT

Nesta seção, será apresentada a estrutura IoT seguindo a metodologia utilizada para desenvolver o sistema proposto para prever abandono escolar utilizando de dados socioeconômicos. A figura 14 apresenta o diagrama contendo as etapas a serem desenvolvidas de acordo com a metodologia adotada.

Figura 14 – Fluxograma da abordagem proposta para detecção e classificação on-line de possíveis alunos que abandonam a escola.



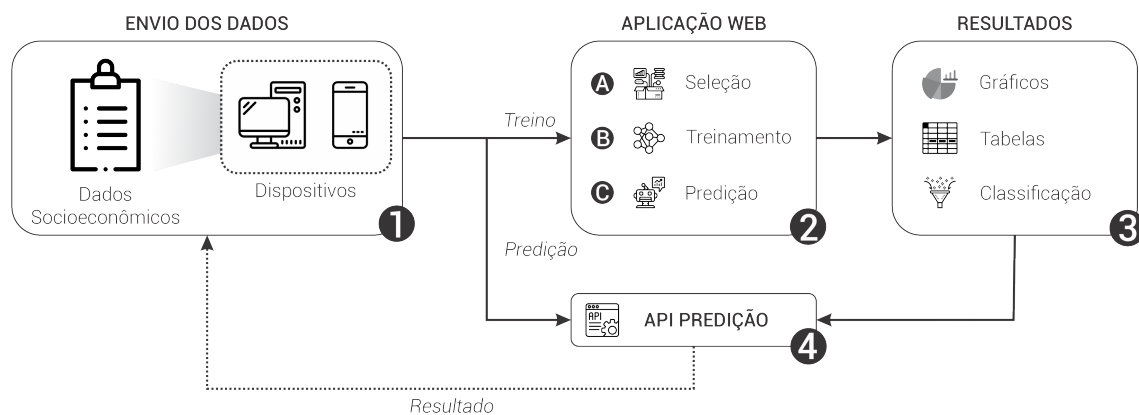
Fonte: Autor.

Para a seleção e o processamento dos dados extraídos dos registros socioeconômicos, foi criada uma plataforma de IoT. Essa plataforma foi desenvolvida usando duas estruturas principais: uma para processamento de dados e previsão de desistência e outra para criação de interface, armazenamento de dados e comunicação entre dispositivos. O primeiro sistema foi desenvolvido usando a linguagem Python e é responsável por processar os registros socioeconômicos disponíveis no aplicativo da web e, posteriormente, prever possíveis desistências. Além disso, esse sistema só entra em operação uma vez solicitado pelo aplicativo Web, que passa as informações dos registros socioeconômicos e dos classificadores que o usuário deseja aplicar. Esse sistema realiza o pré-processamento dos dados enviados, ou seja, atribui os valores a cada atributo categórico como já mencionado e o normaliza em um vetor de característica entre -1 e 1. Por fim, o sistema passa por esse vetor de atributos para o modelo carregado e mostra a previsão feita ao usuário que a solicitou.

A segunda parte do sistema foi desenvolvida usando Java e é a estrutura responsável pelo gerenciamento do serviço Web. Esta parte gerencia as informações dos registros socioeconômicos já disponíveis na plataforma entre os diferentes dispositivos conectados. Para armazenar essas informações, foi utilizado o PostgreSQL, que é um software livre

e armazena não apenas as informações do registro socioeconômico, mas também as configurações dos melhores classificadores utilizados e as últimas consultas realizadas. Essas configurações são disponibilizadas por meio de uma interface fácil de usar, na qual o usuário pode modificar as configurações do classificador ou enviar novos registros socioeconômicos, conforme necessário. A figura 16 mostra uma visão geral da metodologia de processamento adotada.

Figura 15 – Visão geral do processamento de registros socioeconômicos na estrutura da IoT.



Fonte: Autor.

A interface de acesso faz parte do sistema desenvolvido em Java e usa as estruturas Java Persistence API (JPA), Hibernate e Java Server Faces (JSF). Após o login na plataforma, o usuário será direcionado para uma tela com duas opções: ajustar as configurações do sistema ou analisar um registro socioeconômico. Se a opção de analisar o registro socioeconômico for selecionada, o usuário terá a opção de carregar o registro socioeconômico que deseja analisar ou analisar qualquer registro presente no banco de dados. Após selecionar o registro a ser analisado, a resposta será exibida na tela contendo a probabilidade de evasão do aluno e o tempo necessário para a análise.

Para descrever o processo, a API chama o sistema de previsão e envia os dados socioeconômicos e um código de hash exclusivo de cada usuário. O sistema de previsão executa o pré-processamento dos dados, seleciona os atributos e os normaliza entre -1 e 1, faz as previsões e, finalmente, apresenta os resultados obtidos no dispositivo do usuário.

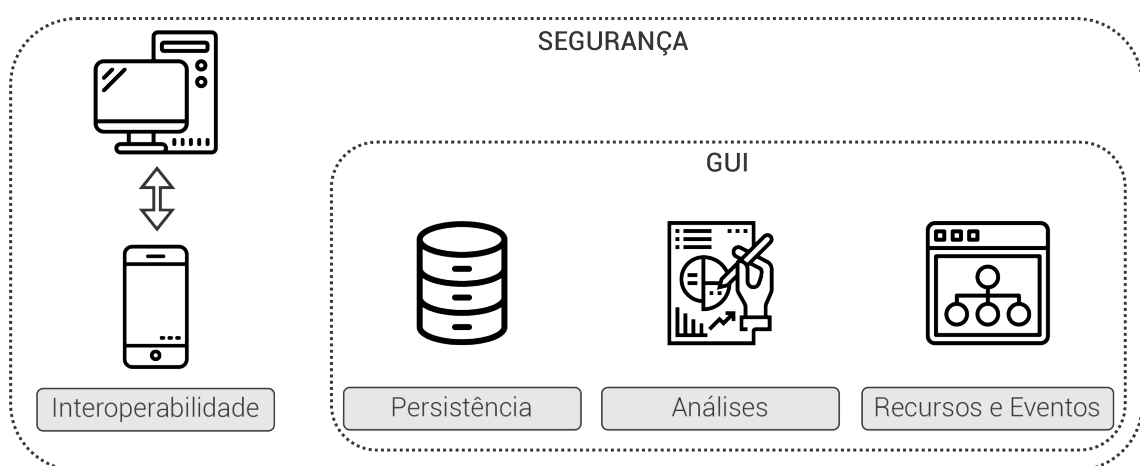
Se, após o *login*, o usuário optar por alterar as configurações, ele será levado para uma nova tela com as opções de configuração dos classificadores e, se for um usuário com permissões mais altas, também poderá alterar as variáveis que serão extraídos dos registros socioeconômicos e passados para os classificadores. Essa opção, embora não seja aconselhada e não esteja disponível para usuários inexperientes, permite que testes futuros sejam realizados para adicionar novas variáveis ao vetor de recurso e podem tornar o sistema ainda mais robusto. Além disso, o usuário também pode fazer *upload* de uma nova

base de registros socioeconômicos para a base existente e treinar um novo modelo a partir do novo conjunto de registros socioeconômicos. Isso torna o sistema mais robusto e flexível.

O objetivo desta plataforma é fornecer aos educadores uma ferramenta simples de aprendizado de máquina para prever desistências que podem ser acessadas por vários dispositivos e sem a necessidade de conhecimento técnico na área. Além disso, todas as previsões e treinamentos são realizados no servidor, reduzindo drasticamente o tempo necessário para receber uma resposta. À medida que mais pessoas usam essa ferramenta, o sistema poderá integrar diferentes bases de registros socioeconômicos entre seus usuários, permitindo que o sistema evolua.

A arquitetura do IoT Framework foi projetada e dividida em quatro partes fundamentais. A primeira parte é a interoperabilidade, que é um recurso que permite a comunicação e a integração com vários dispositivos. A segunda parte é o armazenamento de informações do sistema, isto é, dados de registros socioeconômicos, configurações do classificador e informações do usuário; essa etapa é conhecida como persistência e foi implementada usando o PostgreSQL. A terceira parte é responsável por executar o processamento de dados socioeconômicos e prever um possível abandono, chamado analytics. A última parte, Recursos e Eventos, é responsável pela interação do usuário com o sistema, onde o usuário pode fazer o upload dos registros socioeconômicos, escolher quais classificadores serão utilizados, além da configuração de cada um, entre outras opções. A figura. 16 apresenta o design do sistema em camadas. Em todas essas camadas, a segurança é uma prioridade e está presente em cada uma. A camada de interoperabilidade é a única que não possui uma interface.

Figura 16 – Camadas da arquitetura IoT



Fonte: Autor.

Além dos conceitos mencionados, alguns outros aspectos foram levados em consideração na implementação do sistema: disponibilidade, interoperabilidade e escalabilidade.

A disponibilidade é um recurso crucial de qualquer sistema baseado na Web,

permitindo que a plataforma seja acessada a qualquer momento continuamente. Se o sistema parar, as solicitações feitas serão enfileiradas para processamento adicional quando o sistema se tornar disponível novamente.

A escalabilidade permite que um grande número de usuários acesse a plataforma simultaneamente. Esse recurso depende da modularização do sistema, permitindo que o processamento seja realizado em paralelo e com mais eficiência. Se o sistema não fosse modularizado, não haveria maneira de gerenciar um grande volume de solicitações feitas por vários usuários, pois a parte da web do sistema e a parte de processamento de dados não seriam dissociadas e o processamento de informações não poderia ser paralelo.

A interoperabilidade, como mencionado anteriormente, é essencial para permitir a comunicação e a troca de informações entre diferentes dispositivos. Esse é um recurso fundamental de todos os dispositivos conectados à Internet, pois é necessária uma comunicação eficiente entre eles. O sistema da web lida com uma variedade de dispositivos difíceis de contar e, portanto, são necessárias ferramentas para lidar com esses diferentes dispositivos. Felizmente, ferramentas como HTML (*HyperText Markup Language*) 5 e CSS (*Cascade Style Sheets*) 3 podem lidar com essa situação de maneira rápida e fácil. Também deve ser observado que o acesso ao banco de dados por dispositivos externos e desconhecidos implica uma necessidade ainda maior de segurança do sistema. Para isso, a troca de informações e dados foi implementada usando um arquivo JSON e um código de hash exclusivo por usuário, permitindo que uma ampla variedade de dispositivos diferentes se conectasse e reduzindo os riscos de segurança que o sistema poderia ter.

Em relação as métricas, existe uma que pode ser considerada crítica em aplicativos de aprendizado de máquina é o tempo, não apenas durante a fase de treinamento, mas também no teste. Em aplicativos de IoT, essa variável pode se tornar o principal desafio devido à quantidade de informações transmitidas entre diferentes dispositivos. A tabela 18 apresenta o treinamento e o tempo médio de teste da amostra. Como mostrado, a Árvore de Decisão, que obteve o melhor resultado em outras métricas, tem um tempo de treinamento baixo e tempo de teste.

Tabela 18 – Comparação do treinamento dos classificadores e tempo médio de teste da amostra (τ) em segundos.

Classificador	Tempo do treinamento	τ
Decision Tree	$9,693 \pm 0,126$	$0,154 \pm 0,004$
LR	$0,325 \pm 0,025$	$0,003 \pm 0,000$
SVM (RBF)	$0,278 \pm 0,030$	$0,003 \pm 0,000$
KNN	$0,035 \pm 0,009$	$0,008 \pm 0,001$
MLP	$80,619 \pm 4,053$	$0,009 \pm 0,005$
DNN	$159,063 \pm 18,88$	$0,0149 \pm 0,010$

5 Conclusão

Esta dissertação buscou estabelecer um modelo de predição do risco de evasão escolar dos alunos dos cursos de engenharia do IFCE campus Fortaleza utilizando algoritmos de aprendizagem de máquina. Tendo como questão de pesquisa a seguinte indagação “É possível definir um modelo computacional utilizando algoritmos de aprendizagem de máquina que consiga prever a evasão escolar com dados socioeconômicos de forma automatizada, capaz de reduzir os dados gerados por essa possível perda e auxiliar o gestor na tomada de decisões?”

Para tal, realizou-se uma seleção das características que pudessem relacionar os perfis dos alunos com a evasão escolar, e que essas características também pudessem responder à questão de pesquisa, considerando que existem poucas características socioeconômicas relevantes no Sistema Q-Acadêmico para sinalizar a evasão.

O uso dos algoritmos de aprendizagem de máquina e seus respectivos modelos mostrou-se satisfatória para a mensuração da predição da evasão escolar, objeto deste estudo. Durante as fases e procedimentos desta dissertação, buscou-se desenvolver um modelo que permitisse mensurar o risco do aluno evadir, a partir da sua ficha de cadastro no Sistema Q-Acadêmico.

A ideia de ter uma máquina que possa mensurar esses valores, ajudaria muito aos gestores no momento da tomada de decisões, pois o gestor poderia verificar se o seu julgamento estaria dentro dos parâmetros indicados pelo computador e se, em algum momento, estaria sendo contraditório.

Nessa perspectiva, espera-se que o gestor possa acompanhar e gerenciar o risco de evasão dos alunos pela predição retornada pelo modelo. Ao mesmo tempo, essa predição permite conhecer um pouco mais do perfil do estudante, possibilitando ao gestor realizar, se necessário, uma intervenção.

Como indicado neste estudo, buscou contemplar os modelos utilizados de aprendizagem de máquina para aplicação na avaliação do risco de evasão dos alunos. O uso dos algoritmos de aprendizagem de máquina nos permite dar uma maior flexibilidade no momento de determinar essa predição, já que é possível mensurar o risco de evadir dos mesmos nos cursos de engenharia do IFCE campus Fortaleza.

Referências

- ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, Taylor & Francis Group, v. 46, n. 3, p. 175–185, 1992. Citado na página 37.
- AULCK, L. et al. Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*, 2016. Citado 5 vezes nas páginas 36, 54, 55, 56 e 58.
- BAGHERI, M.; MOVAHED, S. H. The effect of the internet of things (iot) on education business model. In: IEEE. *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. [S.l.], 2016. p. 435–441. Citado na página 24.
- BEZERRA, C. et al. Evasao escolar: Aplicando mineraçao de dados para identificar variáveis relevantes. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2016. v. 27, n. 1, p. 1096. Citado 7 vezes nas páginas 35, 52, 53, 55, 56, 57 e 58.
- CALIXTO, K.; SEGUNDO, C.; GUSMÃO, R. P. de. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 1447. Citado 6 vezes nas páginas 36, 53, 55, 56, 57 e 58.
- CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 36.
- COSTA, G. G. d. O. *Estatística Aplicada à Educação com Abordagem Além da Análise Descritiva*. first. [S.l.]: Ciência Moderna, 2015. Volume 1. Citado 2 vezes nas páginas 33 e 44.
- GAMA, J. et al. Concept drift in decision-tree learning for data streams. In: *Proceedings of the Fourth European Symposium on Intelligent Technologies and their implementation on Smart Adaptive Systems, Aachen, Germany, Verlag Mainz*. [S.l.: s.n.], 2004. p. 218–225. Citado na página 36.
- GOUVEIA, F. Ongs enfrentam desafios e ocupam espaço da ação pública. *Ciência e Cultura*, Sociedade Brasileira para o Progresso da Ciência, v. 59, n. 2, p. 6–8, 2007. Citado na página 43.
- HAYKIN, S. *Redes Neurais Artificiais: Princípios e Práticas*. 2ª edição. [S.l.]: Editora Bookman, Porto Alegre, 900p, 2001. Citado 3 vezes nas páginas 34, 37 e 52.
- HOFMANN, M.; KLINKENBERG, R. *RapidMiner: Data mining use cases and business analytics applications*. [S.l.]: CRC Press, 2013. Citado na página 53.
- IBGE, I. Censo demográfico 2010. *IBGE: Insituto Brasileiro de Geografia e*, 2010. Citado 2 vezes nas páginas 42 e 43.
- INEP, M. Plano nacional de educação pne 2014-2024: linha de base. *Brasília: INEP*, 2015. Citado na página 23.

- JR, C. M. D. et al. Deep learning iot system for online stroke detection in skull computed tomography images. *Computer Networks*, Elsevier, v. 152, p. 25–39, 2019. Citado na página 25.
- JÚNIOR, J. G. de O.; NORONHA, R. V.; KAESTNER, C. A. A. Método de seleção de atributos aplicados na previsão da evasão de cursos de graduação. *Revista de Informática Aplicada*, v. 13, n. 2, 2018. Citado na página 23.
- KARAMOUZIS, S. T.; VRETTOS, A. 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 6 vezes nas páginas 37, 54, 55, 56, 58 e 59.
- LANES, M.; ALCÂNTARA, C. Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2018. v. 29, n. 1, p. 1921. Citado na página 24.
- MARTINHO, V. R.; NUNES, C.; MINUSSI, C. R. Prediction of school dropout risk group using neural network. In: IEEE. *2013 Federated Conference on Computer Science and Information Systems*. [S.l.], 2013. p. 111–114. Citado na página 25.
- MENARD, S. *Applied logistic regression analysis*. [S.l.]: Sage, 2002. v. 106. Citado na página 36.
- NAGY, M.; MOLONTAY, R. Predicting dropout in higher education based on secondary school performance. In: IEEE. *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*. [S.l.], 2018. p. 000389–000394. Citado 5 vezes nas páginas 26, 55, 56, 58 e 59.
- NEHRU, R.; CHAKRABORTY, S. The education of things (eot) for smart learning through iot intervention: A case study based analysis. In: SPRINGER. *International Conference on Intelligent Computing and Communication Technologies*. [S.l.], 2019. p. 529–536. Citado na página 24.
- OECD, I. Education at a Glance 2018 (Summary in Slovak). In: . [s.n.], 2018. ISBN 9789264303386. Disponível em: <https://www.oecd-ilibrary.org/education/education-at-a-glance-2018/summary/slovak{_}ca3ef445>. Citado na página 23.
- PEARSON, K. Pearson, karl and fisher, ra on statistical tests-a 1935 exchange from nature. *American Statistician*, AMER STATISTICAL ASSOC 732 N WASHINGTON ST, ALEXANDRIA, VA 22314-1943 USA, v. 48, n. 1, p. 2–11, 1994. Citado na página 34.
- PEÇANHA, P. N. Disponível em <<http://resultados.plataformanilopecanha.org/2019/>>. Acesso em Agosto de 2019, v. 12, 2019. Citado na página 24.
- RAKSHITH, M.; KENCHANNAVAR, H. H.; THYAGARAJU, G. Prediction techniques in internet of things (iot) environment: A comparative study. *International Journal on Recent and Innovation Trends in Computing and Communication*, v. 5, n. 7, p. 581–590, 2017. Citado na página 24.
- RAMOS, J. L. C. et al. Um estudo comparativo de classificadores na previsão da evasão de alunos em ead. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2018. v. 29, n. 1, p. 1463. Citado na página 24.

- RIBEIRO, V. M.; GUSMÃO, J. Indicadores da qualidade na educação/ação educativa, unicef, pnud, inep-mec (coordenadores). *São Paulo: Ação Educativa*, 2004. Citado na página 43.
- ROBLYER, M.; MARSHALL, J. C. Predicting success of virtual high school students: Preliminary results from an educational success prediction instrument. *Journal of Research on Computing in Education*, Taylor & Francis, v. 35, n. 2, p. 241–255, 2002. Citado na página 24.
- SALES, G. L.; LEITE, E. A. M.; JOYE, C. R. Gerenciamento da aprendizagem, evasão em ead online e possíveis soluções: Um estudo de caso no ifce. *RENOTE*, v. 10, n. 3, 2012. Citado na página 24.
- SANSONE, D. Beyond early warning indicators: High school dropout and machine learning. *Oxford Bulletin of Economics and Statistics*, Wiley Online Library, v. 81, n. 2, p. 456–485, 2019. Citado na página 25.
- SAVIANI, D. et al. O plano de desenvolvimento da educação: análise do projeto do mec. *Educação & Sociedade*, Centro de Estudos Educação e Sociedade-Cedes, 2007. Citado na página 23.
- SORENSEN, L. C. “big data” in educational administration: An application for predicting school dropout risk. *Educational Administration Quarterly*, SAGE Publications Sage CA: Los Angeles, CA, p. 0013161X18799439, 2016. Citado 7 vezes nas páginas 36, 53, 54, 55, 56, 57 e 58.
- SOUZA, J. d.; TABOSA, F. J.; SIMONSASSI, A. Principais fatores que impactam na reprovação e evasão dos alunos dos cursos tecnológicos no instituto federal de ciência e tecnologia do ceará. *XI Encontro de Economia do Ceará-IPECE*, 2015. Citado na página 23.
- TUKEY, J. W. The future of data analysis. *The annals of mathematical statistics*, JSTOR, v. 33, n. 1, p. 1–67, 1962. Citado na página 33.