UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE MESTRADO E DOUTORADO EM CIÊNCIAS DA COMPUTAÇÃO

JOSÉ AUGUSTO CÂMARA FILHO

UM GUIA PRÁTICO PARA APOIAR TAREFAS PREDITIVAS EM CIÊNCIA DE DADOS

FORTALEZA

2022

JOSÉ AUGUSTO CÂMARA FILHO

UM GUIA PRÁTICO PARA APOIAR TAREFAS PREDITIVAS EM CIÊNCIA DE DADOS

Dissertação apresentada ao Curso de do Programa de Mestrado e Doutorado em Ciências da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Sistemas da Informação.

Orientador: Prof. Dr. José Maria Da Silva Monteiro Filho.

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C173g Câmara Filho, José Augusto.
Um guia prático : para apoiar tarefas preditivas em Ciência de dados / José Augusto Câmara Filho. – 2022.
141 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2022.
Orientação: Prof. Dr. José Maria da Silva Monteiro Filho.

1. Guias práticos. 2. Problemas de predição. 3. Ciência de dados. I. Título.

CDD 005

JOSÉ AUGUSTO CÂMARA FILHO

UM GUIA PRÁTICO PARA APOIAR TAREFAS PREDITIVAS EM CIÊNCIA DE DADOS

Dissertação apresentada ao Curso de do Programa de Mestrado e Doutorado em Ciências da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Sistemas da Informação.

Aprovada em: 30/11/2022

BANCA EXAMINADORA

Prof. Dr. José Maria Da Silva Monteiro Filho (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. João Paulo do Vale Madeiro
Universidade Federal do Ceará (UFC)

Prof. Dr. José Gilvan Rodrigues Maia
Universidade Federal do Ceará (UFC)

À minha falecida mãe, por sempre acreditar em mim, mesmo quando eu já nem acreditava mais. Ela que em diversos momentos esteve sempre com seu pensamento em mim.

AGRADECIMENTOS

Ao meu pai, José Augusto Câmara, que nos momentos de minha ausência dedicados ao estudo sempre fez entender que o futuro é feito a partir da constante dedicação que temos em nosso presente.

A Jomábia Cristina Gonçalves dos Santos, que em seus incentivos diários me fez aprender mais, evoluir e enxergar a vida com outros olhos. Além de sempre se fazer presente nos dias felizes ou nos dias tristes, sendo uma rocha mesmo sem saber que estava sendo pra mim.

Ao Professor Doutor José Maria Da Silva Monteiro Filho Monteiro, por me orientar em minha dissertação de mestrado, além de ter me amparado nos mais diversos momentos difíceis que tive nesse período.

Aos professores do Departamento de Computação(DC), aos professores do Departamento de estatística e matemática aplicada(DEMA), e em especial ao Professor Doutor José Gilvan Rodrigues Maia da UFC Virtual por terem me proporcionado o conhecimento não apenas racional, mas a manifestação do caráter e afetividade no processo educacional de formação profissional, por todas as horas despedidas para se dedicaram a mim, não somente por terem me ensinado, mas por terem me feito aprender.

Aos amigos de laboratório do ÁRIDA e demais colegas de mestrado que durante esses 3 anos foram apoio incondicional para suportar as disciplinas, desafios de pesquisa e principalmente torcer para que todos conseguissem chegar até os nosso objetivos.

Aos meus colegas de trabalho da Dell Lead e do Grupo 3 Corações por apoiarem e terem paciência quando precisava dividir as atenções com pendências do mestrado.

E ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) por financiarem esses anos de pesquisa.

"O que significa isso, melhorar como pessoa? Melhorar como pessoa...eu acho que eu tenho só 28 anos, eu tenho uma vida toda pela frente, eu tenho muito pra aprender na minha vida e muito pra dar aqueles que realmente participam da minha vida. Você sente que ficou alguém no caminho que você teria uma dívida ou que não deu? Se existe alguém, sou eu mesmo."
((BYTES, 2020))

RESUMO

Atualmente, profissionais das mais diversas áreas de atuação precisam explorar seus repositórios de dados com a finalidade de extrair conhecimento e criar novos produtos ou serviços. Diversas ferramentas têm sido propostas com o objetivo de facilitar as tarefas envolvidas no ciclo de vida da Ciência de Dados. No entanto, tais ferramentas exigem de seus usuários conhecimentos específicos em diversas áreas da Computação e Estatística, tornando seu uso praticamente inviável por profissionais não especialistas em Ciência de Dados. Nesta dissertação, propomos um guia prático para apoiar tarefas preditivas, mais especificamente, regressão e classificação. Além disso, apresentamos uma ferramenta, denominada *DSAdvisor*, que seguindo o guia proposto busca auxiliar os usuários na execução das diversas atividades envolvidas em um problema de predição. A *DSAdvisor* visa encorajar usuários leigos a construir modelos de aprendizado de máquina para executar tarefas preditivas, extraindo conhecimento de seus próprios repositórios de dados. Por fim, para avaliar a *DSAdvisor*, aplicamos o questionário *System Usabilidade Scale*(*System Usability Scale* (SUS)) para mensurar aspectos de usabilidade de acordo com a avaliação subjetiva do usuário e o método *Net Promoter Score*(*Net Promoter Score* (NPS)) para mensurar a satisfação do usuário e a disposição de recomendar a ferramenta a outros usuários. Este estudo envolveu 20 respondentes que foram divididos em dois grupos, a saber, usuários especialistas e não especialistas. O método SUS obteve uma pontuação de 68,5 o que significa um produto "bom", e os resultados da utilização do NPS obtiveram um valor de 55% o que significa um NPS "muito bom".

Palavras-chave: guias práticos; problemas de predição; ciência de dados.

ABSTRACT

Currently, professdiveionals from thst diverse areas need to explore their data repositories in order to extract knowledge and create new products or services. Several tools have been proposed in order to facilitate the tasks involved in the Data Science lifecycle. However, such tools require their users to have specific (and deep) knowledge in different areas of Computing and Statistics, making their use practically unfeasible for non-specialist professionals in Data Science. In this paper, we propose a guideline to support predictive tasks, more specifically, regression and classification. In addition, we present a tool, called DSAdvisor, which following the stages of the proposed guideline and aims to encourage non-expert users to build machine learning models to solve predictive tasks, extracting knowledge from their own data repositories. To evaluate DSAdvisor, we applied the System Usability Scale (SUS) questionnaire to measure aspects of usability in accordance with the user's subjective assessment and the Net Promoter Score (NPS) method to measure user satisfaction and willingness to recommend it to others. This study involved 20 respondents who were divided into two groups, namely experts and non-expert users. The SUS method had a score of 68.5 which means a "good" product, and the results of using NPS get a value of 55% which means "very good" NPS.

Keywords: guidelines; predictive tasks; data science.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CD	Ciência de dados
CSV	*Comma-separated values*
ENN	*Edited Nearest Neighbor*
FN	Falso Negativo
FP	Falso Positivo
IM	Informática dos Materiais
IQR	intervalo interquartílico(IQR)
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Absoluto Médio
MC	*medcouple*
ML	*Machine Learning*
MSE	Erro Quadrático Médio
NPS	*Net Promoter Score*
OS	*Over-sampling*
RMSE	Raiz do Erro Quadrático Médio
ROC	*Receiver operating characteristic*
RUS	*Random Under-Sampling*
Smote	*Synthetic Minority Oversampling TEchnique*
SUS	*System Usability Scale*
US	*Under-sampling*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

Devido à grande quantidade de dados atualmente disponíveis, surge a necessidade, por parte de profissionais das mais diferentes áreas, de extrair conhecimento de seus próprios repositórios com a finalidade de criar novos produtos e serviços. Por exemplo, médicos cardiologistas precisam explorar grandes repositórios de sinais eletrocardiográficos para prever a probabilidade de morte súbita de um determinado paciente. Da mesma forma, os auditores fiscais podem explorar seus bancos de dados para prever a probabilidade de evasão fiscal. Mas também podemos citar exemplos mais próximos do nosso cotidiano, onde pequenos empresários precisam criar estratégias para alavancarem suas vendas, onde a principal fonte de dados é o histórico de vendas e o perfil de seus clientes.

Por outro lado, o volume e a variedade de dados excedem em muito a capacidade humana de análise manual. Neste contexto, foram desenvolvidos algoritmos complexos que permitem identificar padrões ocultos nesses conjuntos de dados. Adicionalmente, avanços nas tecnologias de computação paralela e em nuvem possibilitaram executar esses algoritmos em tempos cada vez menores, tornando viável sua utilização em produtos e serviços digitais. A convergência desses fenômenos impulsionou o desenvolvimento e a popularização da ciência de dados (PROVOST; FAWCETT, 2013).

Ciência de dados(Ciência de dados (CD)) é uma área multidisciplinar que envolve a extração de informação e conhecimento de grandes repositórios de dados (PROVOST; FAWCETT, 2013), o que inclui diferentes desafios, tais como: a coleta, integração, gestão, exploração e extração de conhecimento dos dados para a tomada de decisões, entender o passado e o presente, prever o futuro e criar novos serviços e produtos (OZDEMIR, 2016). Desta forma, a ciência de dados possibilita a obtenção de novos *insights* ocultos nesses conjuntos de dados. A Figura 1 mostra os estágios do ciclo de vida da ciência de dados: compreensão do negócio, compreensão de dados, preparação de dados, modelagem, avaliação e implantação.

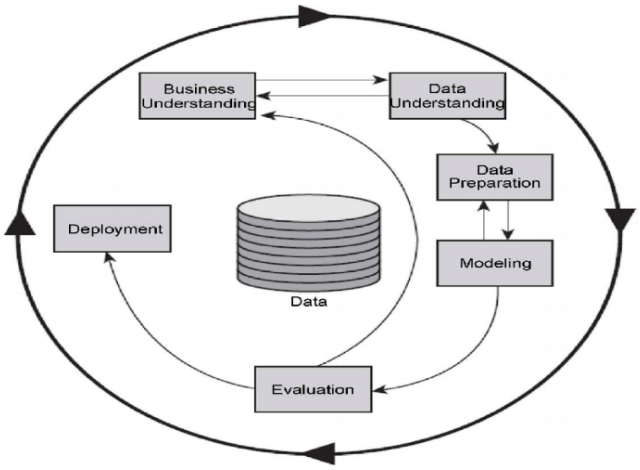


Figura 1 – Ciclo de Ciência de Dados. Adaptado de Chapman *et al.* (2019).

Para extrair conhecimento dos dados, devemos ser capazes de (CHERTCHOM, 2018):

1. Entender problemas ainda não resolvidos com o uso de técnicas de mineração de dados;
2. Entender os dados e suas inter-relações;
3. Extrair um subconjunto dos dados;
4. Criar modelos de aprendizagem de máquina para solucionar um determinado problema;
5. Avaliar o desempenho dos modelos criados;
6. Demonstrar como esses modelos podem ser usados na tomada de decisão.

A complexidade dessas tarefas explica por que apenas usuários altamente experientes conseguem dominar todo o ciclo de vida da ciência de dados (CHERTCHOM, 2018). Por outro lado, diferentes ferramentas têm sido propostas para apoiar as tarefas envolvidas no ciclo de vida da ciência de dados. No entanto, tais ferramentas exigem que seus usuários tenham conhecimentos específicos (e profundos) em diversas áreas, tais como Computação e Estatística, o que acaba tornando seu uso praticamente inviável por profissionais que não sejam especialistas em CD.

Uma das principais tarefas da CD consiste em criar modelos de aprendizagem de máquina para solucionar um determinado problema. A aprendizagem de máquina, do inglês *Machine Learning*(*Machine Learning* (ML)), tem sido utilizada com sucesso para solucionar algumas categorias de problemas bem específicas, como ilustra a Figura 2 e destacamos a seguir:

- Aprendizado Supervisionado (*Supervised Learning*)
  - Classificação
  - Regressão

- Aprendizado Não supervisionado (*Unsupervised Learning*)
  - Clusterização
  - Redução de dimensionalidade
- Aprendizado por Reforço (*Reinforcement Learning*)
  - Decisões em tempo real
  - Inteligência artificial para jogos

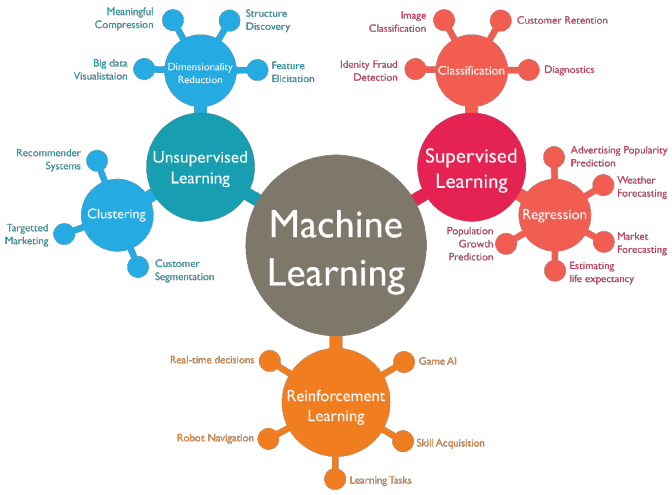


Figura 2 – Sub divisões da área de *Machine Learning*. Fonte: link.

Embora a habilidade de elaborar essas soluções esteja retida em alguns poucos profissionais, pesquisas recentes tentam contornar esse problema através da criação de ferramentas de ML que sejam fáceis e acessíveis para pessoas que não são formalmente cientistas de dados, as quais podemos nomear como "não especialistas- pessoas que não são formalmente treinadas em ML e podem estar ativamente criando soluções para atender às suas necessidades. O objetivo principal dessas ferramentas consiste em possibilitar que usuários "não especialistas"apliquem técnicas e métodos de ML para impulsionar seus negócios e resolver problemas repetitivos e/ou custosos (YANG *et al.*, 2018).

Nesta dissertação, propomos um guia prático para apoiar tarefas preditivas, mais especificamente, regressão e classificação. Além disso, apresentamos uma ferramenta, denominada *DSAdvisor*, que seguindo o guia proposto busca auxiliar os usuários na execução das diversas atividades envolvidas em um problema de predição. A *DSAdvisor* visa encorajar usuários leigos a construir modelos de aprendizado de máquina para executar tarefas preditivas, extraindo conhecimento de seus próprios repositórios de dados. A *DSAdvisor* atua como um consultor para usuários não especialistas. Por fim, avaliamos a ferramenta *DSAdvisor* utilizando dois diferentes testes de usabilidade: *Net Promoter Score*(NPS) e *System Usability Scale*(SUS).

O restante desta dissertação está organizada da seguinte forma. O Capítulo 2 revisa os trabalhos relacionados dividindo-se entre diretrizes em ciências de dados, na qual exploramos diretrizes já existentes utilizadas em situações similares, e em ferramentas de ciência de dados. No Capítulo 3, apresentamos e discutimos o guia proposto com a finalidade de auxiliar os profissionais na solução de problemas preditivos. A implementação da *DSAdvisor* é discutida no Capitulo 4. No Capitulo 5 apresentamos os testes de usabilidade e satisfação do usuário aplicados com o objetivo de avaliar a ferramenta *DSAdvisor*, mais especificamente o *Net Promoter Score*(NPS) e o *System Usability Scale*(SUS), bem como a metodologia utilizada na aplicação desses testes. O Capitulo 6, apresenta os resultados obtidos nos testes NPS e SUS. Por fim, no Capítulo 7 apresentamos nossas conclusões e apontamos direções para trabalhos futuros.

2 TRABALHOS RELACIONADOS

Nesta seção, discutiremos os principais trabalhos relacionados a esta dissertação. Para um melhor entendimento, organizamos tais iniciativas em duas categorias: guias práticos e ferramentas de apoio.

2.1 Guias Práticos

Um guia prático é um roteiro que determina o curso de um conjunto de ações que compõem um processo específico, além de um conjunto de boas práticas para a execução dessas atividades. Algumas diretrizes foram propostas para orientar tarefas gerais de mineração de dados, buscando auxiliar o planejamento e a realização de experimentos, bem como a análise dos resultados obtidos.

2.1.1 Um Guia Prático para Auxiliar a Predição de Mudanças em Software

Em Melo *et al.* (2019), os autores apresentam um guia prático para apoiar o problema de predizer classes com tendência a mudanças em softwares desenvolvidos seguindo o paradigma da orientação a objetos. Esse guia é organizado em duas fases: projetar o conjunto de dados e aplicar a previsão para prever a mudança em softwares.

A primeira fase visa projetar e construir o conjunto de dados que será usado pelos algoritmos de aprendizado de máquina para prever classes sujeitas a mudanças. Esta fase, ilustrada na Figura 3, compreende as seguintes atividades: escolher as variáveis independentes, escolher a variável dependente e coletar as métricas selecionadas.

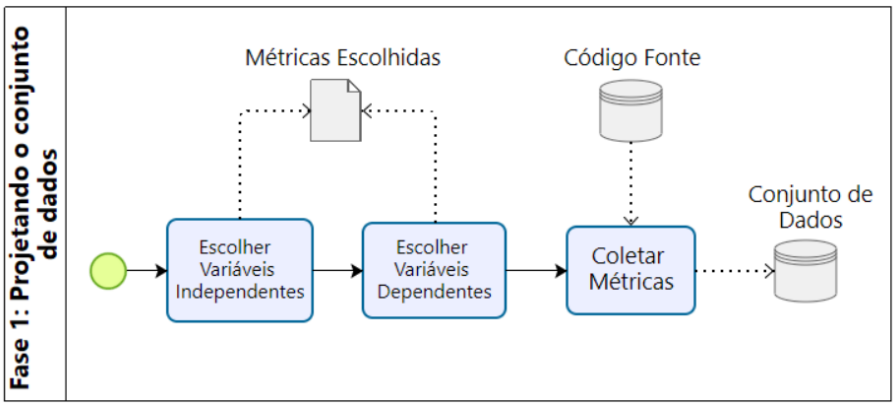


Figura 3 – Fase 1 - Projetar o conjunto de dados. Extraído de Melo *et al.* (2019)

A segunda fase do guia proposto visa construir modelos de predição de classes sujeitas a mudanças. Um modelo preditivo é construído a partir de dados históricos rotulados de forma supervisionada. Além disso, esta fase engloba as atividades relacionadas à análise das métricas de desempenho do modelo preditivo, a apresentação dos resultados e a garantia da reprodutibilidade dos experimentos. A Figura 5 ilustra as atividades que compõem a segunda fase do guia proposto.

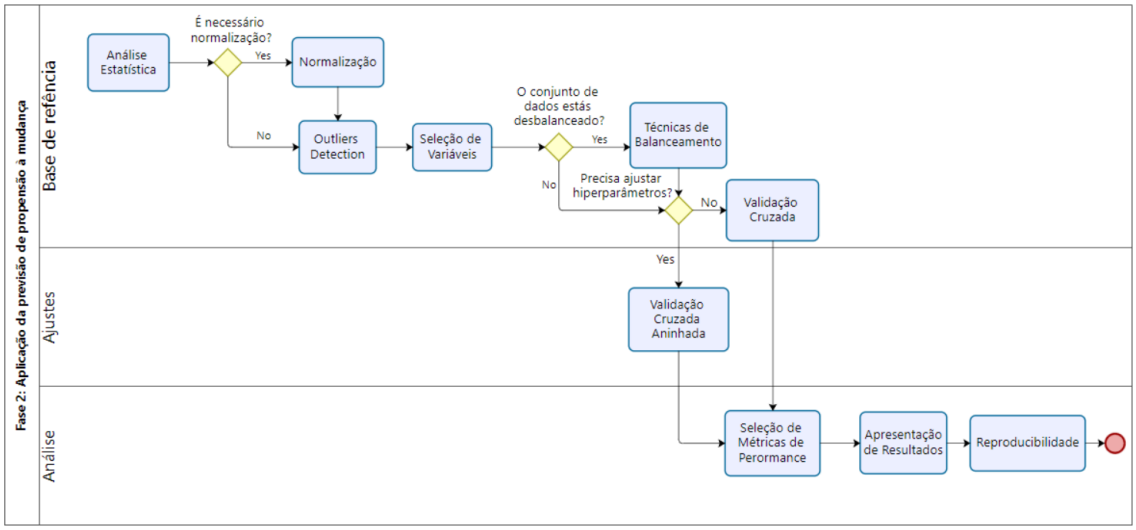


Figura 4 – Fase 2 - Aplicar a previsão de tendência à mudança. Extraído de Melo *et al.* (2019)

2.1.2 Um Guia Prático para Aplicar Aprendizado de Máquina na Ciência de Materiais

Wang *et al.* (2020) enfatiza o alto custo dos métodos tradicionais de tentativa e erro na pesquisa de materiais. Os cientistas de materiais têm confiado cada vez mais em métodos de simulação e modelagem para compreender e prever as propriedades dos materiais. A informática dos materiais(Informática dos Materiais (IM)) é um ramo resultante da ciência de materiais que usa computação de alto desempenho para analisar grandes bancos de dados de propriedades de materiais para obter percepções exclusivas.

Mais recentemente, o aprendizado de máquina foi adotado em IM para estudar a riqueza de dados experimentais e computacionais existentes na ciência dos materiais, levando a uma mudança de paradigma na forma como a pesquisa nesta área é conduzida. No entanto, existem muitos desafios ao se implementarem técnicas de ML, por exemplo, muitos cientistas de materiais experimentais carecem de *"know-how"* para iniciar a pesquisa baseada em dados, e faltam as melhores práticas recomendadas para a implementação de tais métodos na ciência dos materiais. O trabalho apresentado em Wang *et al.* (2020) propõe um projeto de ML, passo

a passo, para estudiosos da ciência dos materiais que desejam realizar pesquisas baseadas em dados.

Adicionalmente, os autores propuseram uma diretriz para aplicação dos métodos de ML na ciência dos materiais (Figura 5), iniciando com o carregamento e processamento de dados, divisão de dados, engenharia de atributos, ajuste de diferentes modelos de ML, avaliação do desempenho do modelo, comparação do desempenho entre modelos e visualização dos resultados. Ao longo deste processo, destacam-se alguns desafios e erros comuns encontrados durante um estudo empírico de ML em ciência dos materiais, bem como abordagens para superá-los ou resolvê-los.

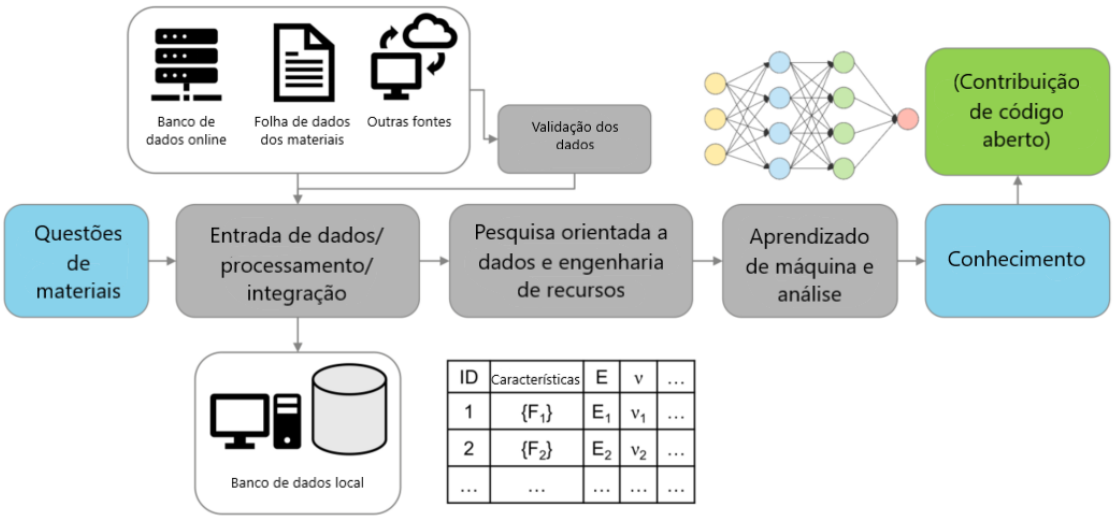


Figura 5 – Diretriz de aprendizado de máquina em ciência dos materiais. Extraído de Wang *et al.* (2020)

2.2 Ferramentas de Mineração de Dados

As ferramentas tradicionais de mineração de dados ajudam tanto cientistas quanto empresas a extrair padrões e conhecimento de seus dados, usando uma série de algoritmos e técnicas de aprendizado automático (RAMAMOHAN *et al.*, 2012). Essas ferramentas diferem entre si em relação a diferentes aspectos, tais como: usabilidade, linguagem em que foram desenvolvidas, suporte aos experimentos e tratamento dos dados. Entre as ferramentas de mineração de dados existentes, as mais populares são: KEEL, Knime, Orange, RapidMiner e WEKA (HASIM; HARIS, 2015).

2.2.1 Keel

A ferramenta KEEL(*Knowledge Extraction based on Evolutionary Learning*) facilita a análise de dados e extração de conhecimento por meio de técnicas de aprendizagem evolutiva, tais como: *evolutionary feature and instance selection*(CANO *et al.*, 2003), *evolutionary fuzzy rule learning and Mamdani rule tuning*(ALCALÁ *et al.*, 2006), *genetic artificial neural networks* (MARTÍNEZ-ESTUDILLO *et al.*, 2006), *Learning Classifier Systems* (BERNADÓ-MANSILLA; HO, 2005).

KEEL se destina, principalmente, a dois públicos distintos, pesquisadores e alunos, os quais possuem necessidades diferentes, como ilustrado na Figura 6. O uso mais comum dessa ferramenta por um pesquisador será a automação de experimentos e a análise estatística dos seus resultados. Rotineiramente, um projeto experimental inclui uma mistura de algoritmos evolutivos, análises estatísticas e técnicas relacionadas à Inteligência Artificial (IA). Já os estudantes possuem necessidades diferentes em relação aos pesquisadores, pois o objetivo não é mais fazer comparações estatisticamente sólidas entre algoritmos. Não há necessidade de repetir um experimento um grande número de vezes. Se a ferramenta for utilizada para fins didáticos em sala de aula, o tempo de execução deve ser curto e uma visão em tempo real da evolução dos algoritmos é necessária para o aluno compreender como ajustar os parâmetros dos algoritmos. Atualmente a versão disponível do KEEL contem as seguintes funcionalidades:

- Gerenciamento de Dados: Este módulo é composto por um conjunto de ferramentas que podem ser usadas para construir conjuntos de dados (*data sets*), exportar e importar dados em diferentes formatos para ou do formato específico da KEEL, edição e visualização de dados, aplicar transformações e particionamentos nos dados, etc.
- Projetar Experimentos (módulo *off-line*): O objetivo deste módulo consiste em auxiliar o projeto de experimentos sobre os conjuntos de dados selecionados, o que inclui a construção de modelos preditivos ou descritivos.
- Experimentos Educacionais (módulo *on-line*): Com uma estrutura semelhante ao módulo anterior, permite projetar experimentos que podem ser executados passo a passo para ilustrar o processo de aprendizagem de um determinado algoritmo, sendo bastante interessante para fins educacionais.

Figura 6 – Tela da ferramenta KEEL. Extraído de KEEL.ES

2.2.2 KNIME

KNIME(*Konstanz Information Miner*) é um ambiente modular que permite fácil integração de novos algoritmos, manipulação de dados de forma bastante simples e diferentes métodos de visualização. Sua interface é configurável permitindo a seleção de diversos métodos de aprendizado automático. Especificamente, pode-se selecionar fontes de dados, tarefas de pré-processamento de dados, algoritmos de aprendizado de máquina, bem como ferramentas de visualização. Para criar um fluxo de trabalho, o usuário conecta nós a portas de entrada ou de saída, de forma visual, apenas clicando e arrastando componentes visuais para uma área de trabalho Berthold *et al.* (2009).

No KNIME, o usuário pode modelar fluxos de trabalho, que consistem em nós que processam dados, transportados por meio de conexões entre esses nós. Um fluxo geralmente começa com um nó que lê dados de alguma origem, que geralmente são arquivos de texto, mas podem ser planilhas ou bancos de dados, como ilustrado na Figura 7. Os dados importados são armazenados em um formato proprietário e apresentados em forma de tabelas, cujas colunas podem ser de diferentes tipos: inteiro, string, imagem, molécula, etc.

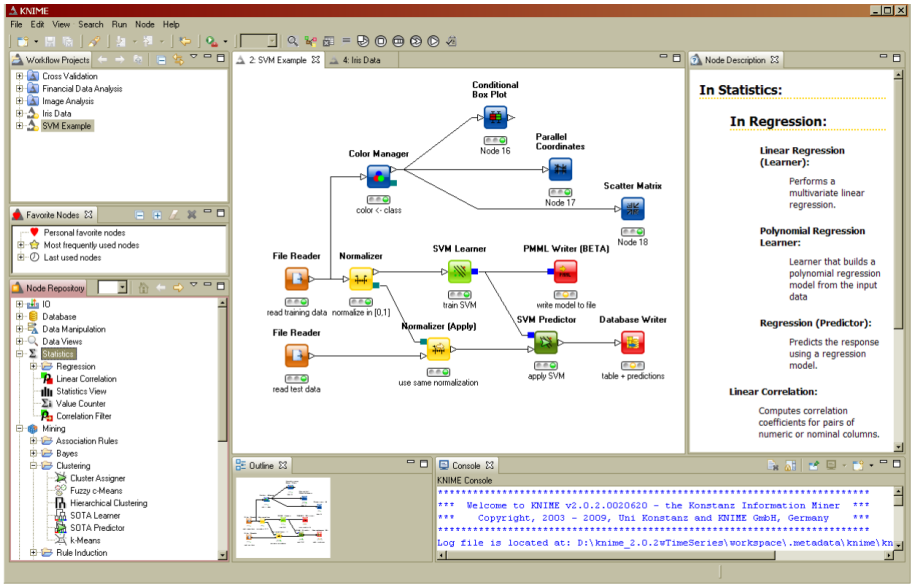


Figura 7 – Tela da ferramenta KNIME. Extraído de (BERTHOLD *et al.*, 2009)

Diferentemente de outras ferramentas, os nós do KNIME processam inteiramente toda uma tabela de entrada antes que os resultados sejam encaminhados para os nós sucessores. As vantagens são que cada nó armazena seus resultados permanentemente e, portanto, a execução do fluxo de trabalho pode ser facilmente interrompida em qualquer nó e retomada posteriormente.

2.2.3 Orange Data Mining Tool

A ferramenta Orange tem diferentes recursos que são visualmente representados por *widgets* (por exemplo, ler arquivo, discretizar, treinar classificador, SVM, etc.) (DEMŠAR *et al.*, 2013). Cada *widget* possui uma breve descrição na interface. A programação é realizada colocando *widgets* na tela para conectá-los às suas entradas e saídas assim como ilustra a Figura 8.

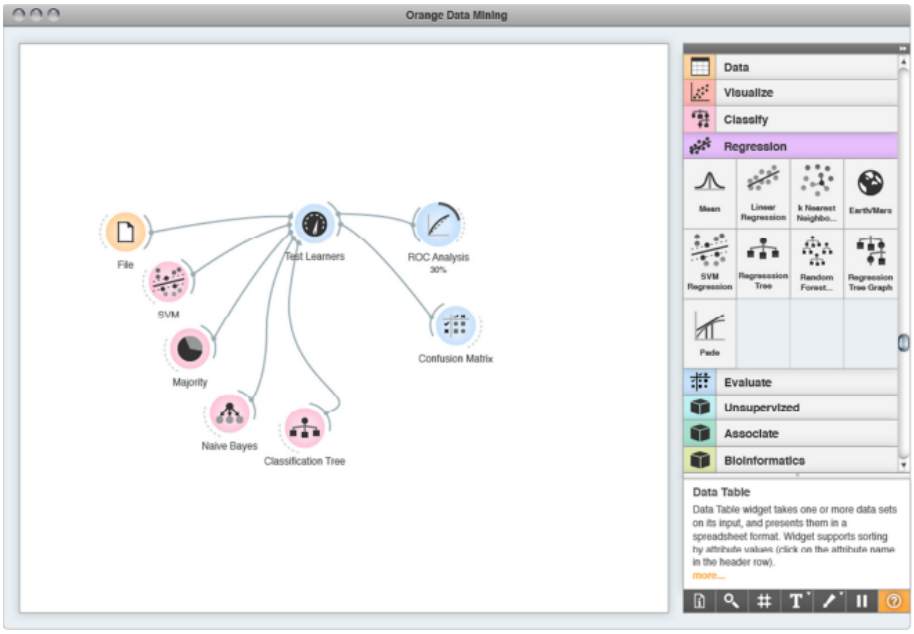


Figura 8 – Tela da ferramenta Orange Data Mining Tool. Extraído de Demšar e Zupan (2012)

Ao contrário de outras ferramentas que possuem implementações para dezenas de algoritmos diferentes, a Orange fornece implementações apenas para os algoritmos mais comumente utilizados. Porém, possibilita utilizar esses algoritmos de forma flexível e amigável. A ênfase da ferramenta está na exploração de dados. Assim, a Orange pode ser utilizada na indústria, na pesquisa científica e até para fins didáticos. Atualmente, o seu parceiro industrial mais notável é a Astra-Zeneca, uma gigante farmacêutica, que usa a Orange no desenvolvimento de medicamentos e patrocina o seu desenvolvimento (DEMŠAR; ZUPAN, 2012).

2.2.4 RapidMiner

RapidMiner é um sistema que apoia o desenvolvimento e a documentação de projetos de mineração de dados. Ele oferece não apenas um conjunto quase completo de operadores, mas também estruturas que expressam o fluxo de controle do processo. Nestes fluxos, as estruturas utilizadas assemelham-se às encontradas nas linguagens de programação. Porém, o RapidiMiner não exige que o usuário saiba programar. Visualmente, as estruturas são mostradas por caixas que estão vinculadas ou aninhadas Hofmann e Klinkenberg (2016).

O RapidMiner possui uma interface confortável, onde as análises são configuradas em um determinado fluxo, onde cada etapa (por exemplo, uma etapa de pré-processamento ou um procedimento de aprendizagem) é representada por um operador Land e Fischer (2012).

Esses operadores possuem portas de entrada e saída por meio das quais podem

se comunicar com outros operadores para receber ou enviar dados processados e/ou modelos gerados para outros operadores. Assim, um fluxo de dados é criado ao longo de todo o processo de análise, como pode ser ilustrado por meio da Figura 9.

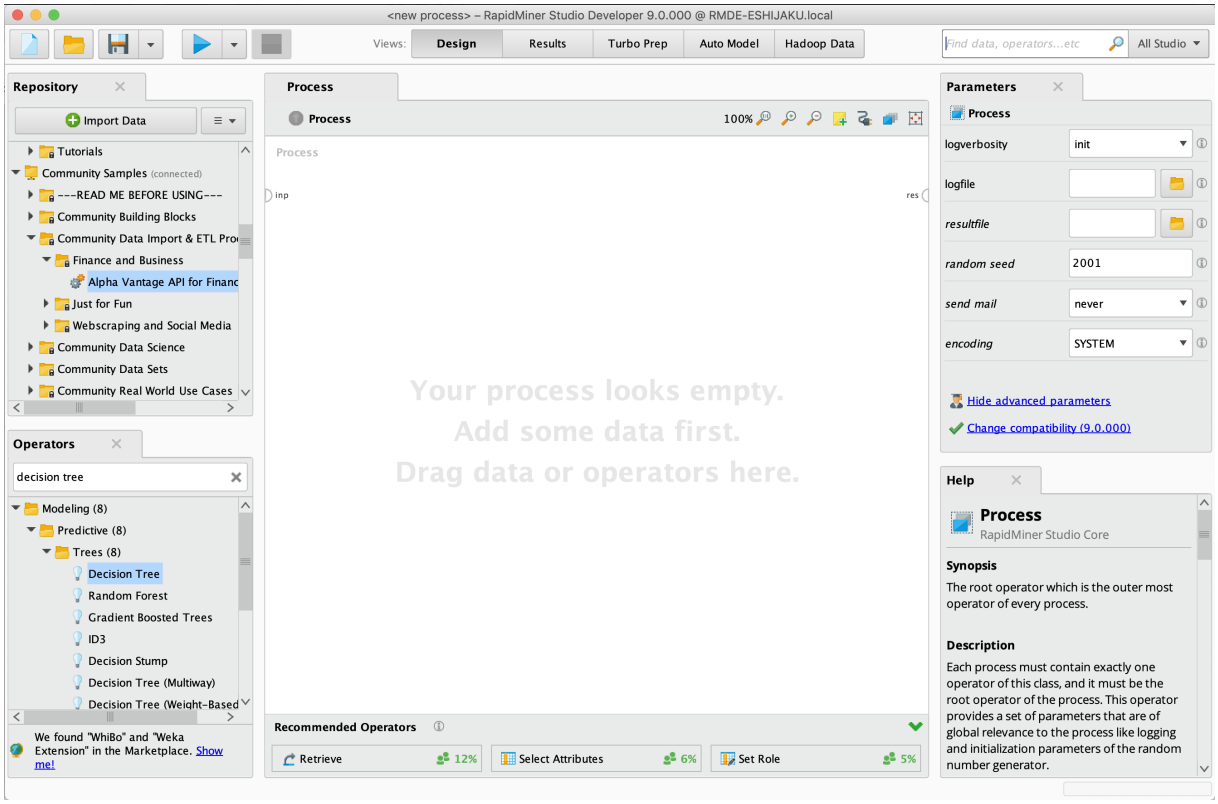


Figura 9 – Tela da ferramenta RapidMiner. Extraído de Land e Fischer (2012)

2.2.5 Weka

O *Waikato Environment for Knowledge Analysis* (WEKA) surgiu por meio da necessidade de um ambiente de trabalho unificado que permitisse aos pesquisadores fácil acesso às técnicas de ponta em aprendizado de máquina. Atualmente, o WEKA é reconhecido como um sistema de referência em mineração de dados e aprendizado de máquina Piatetsky-Shapiro (2005).

O projeto WEKA visa fornecer uma coleção abrangente de algoritmos de aprendizado de máquina e ferramentas de pré-processamento de dados para pesquisadores e profissionais. Permite que os usuários experimentem e comparem rapidamente diferentes métodos de aprendizado de máquina em novos conjuntos de dados. Sua arquitetura modular e extensível permite que processos sofisticados de mineração de dados sejam desenvolvidos a partir de uma ampla coleção de algoritmos e ferramentas de aprendizagem Hall *et al.* (2009). Weka oferece quatro opções operacionais: interface de linha de comando (CLI), *Explorer*, *Experimenter* e *Knowledge*

flow. A opção *"Explorer"* permite a definição de fonte de dados, preparação de dados, execução de algoritmos de aprendizado de máquina e visualização de dados como ilustra a Figura 10.

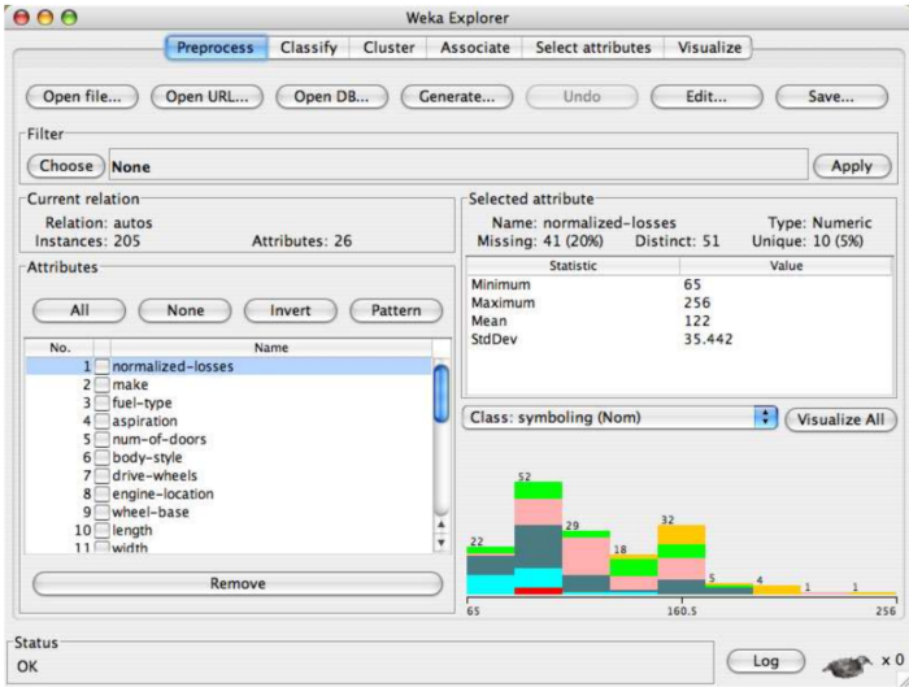


Figura 10 – Tela da ferramenta Weka. Extraído de Land e Fischer (2012)

Mesmo antes da popularização da ciência de dados, todas essas ferramentas foram desenvolvidas para auxiliar as tarefas de mineração de dados. Essas ferramentas possuem características distintas, tais como: usabilidade, tipo de licença, linguagem de programação em que foram desenvolvidas, se a ferramenta possui algum suporte à reprodutibilidade(repro), a curva de aprendizagem pelo usuário para melhor manipulação dos recursos da ferramenta(podendo variar em alta, intermediária e baixa, onde o melhor cenário é quando temos uma curva levemente acentuada, refletindo menores esforços de manipulação do usuário), dentre outras. As ferramentas mais populares são: KEEL, Knime, Orange, RapidMiner, Tanagra e Weka. A tabela 1 fornece uma comparação entre essas ferramentas e a DSAadvisor (ferramenta proposta nesta dissertação).

Lista de ferramentas				
Ferramenta	**Usabilidade**	**Linguagem**	**Curva de aprendizagem**	**Reprodutibilidade**
DSAdvisor	Alta	Python	Baixa	Sim
KEEL	Baixa	Java	Alta	Sim
KNIME	Baixa	Java	Intermediária	Sim
RapidMiner	Baixa	Java	Intermediária	Sim
Orange	Alta	C++, Python	Baixa	Sim
Weka	Baixa	Java	Alta	Não

Tabela 1 – Características gerais de softwares de mineração de dados. Adaptado de Hasim e Haris (2015)

3 UM GUIA PRÁTICO PARA APOIAR TAREFAS PREDITIVAS

Neste capítulo, apresentaremos um guia prático que tem por finalidade auxiliar profissionais de diferentes áreas do conhecimento nas diversas atividades envolvidas na solução de problemas de predição, mais especificamente, regressão e classificação. O guia proposto nesta dissertação é organizado em três fases, sendo elas: análise exploratória, pré-processamento dos dados e criação de modelos preditivos.

3.1 Fase 1 - Análise Exploratória

A primeira fase do guia proposto tem por finalidade explorar os dados que serão utilizados na construção de um ou mais modelos preditivos. O principal objetivo desta fase consiste em entender, descrever e resumir os dados que serão utilizados. A Figura 11 ilustra esta primeira fase, a qual compreende as seguintes atividades: carga de dados, verificação do tipo de cada variável (atributo ou coluna), remoção de variáveis, definição dos códigos que serão utilizados para representar valores faltantes, exibição de estatísticas descritivas (média, mediana, desvio padrão, etc), classificação das variáveis em categóricas ou discretas, análise da distribuição de cada uma das variáveis contínuas e exame das correlações entre as variáveis.

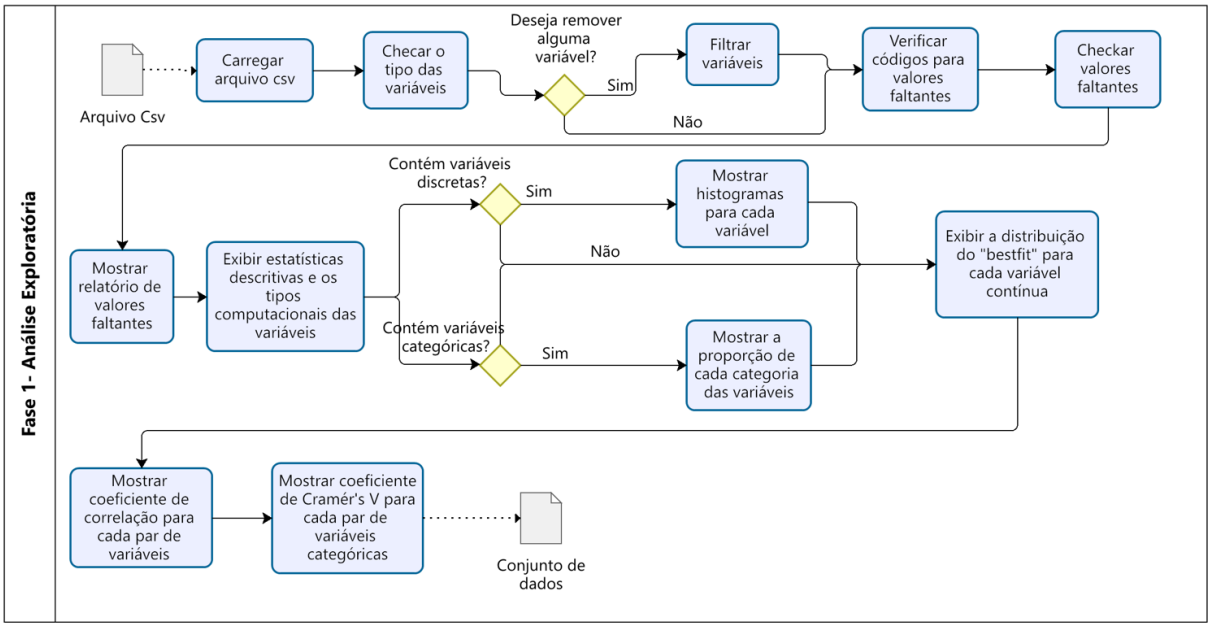


Figura 11 – Fase 1 - Análise Exploratória. Fonte: Autor.

3.1.1 Carregar os dados

Inicialmente, o cientista obtém os dados que serão utilizados a partir de um ou mais repositórios de origem. Em geral, esses dados são armazenados em arquivos do tipo *Comma-separated values* (CSV), txt, xlsx ou similares, e seguem uma disposição matricial (retangular), contendo linhas e colunas bem definidas. As colunas representam as variáveis de análise, podendo também serem chamadas de atributos, enquanto as linhas constituem as instâncias, também chamadas de observações. Para visualizar os dados, o cientista geralmente utiliza algum software de edição ou análise. Neste momento, é fundamental que o cientista de dados verifique se o arquivo está correto e completo.

3.1.2 Checar o tipo de cada variável

Logo após confirmar que o software selecionado realizou o carregamento dos dados com êxito, ou seja, os dados estão corretos e completos, é recomendado verificar o tipo de cada variável. Essas variáveis podem ser dos seguintes tipos:

- **Variável Quantitativa:** Seu valor é expresso através de números. Uma variável quantitativa pode ainda ser classificada como:
  - **Variável Contínua:** Onde o valor converge para um conjunto infinito de valores. Ex: Peso medido pela balança, pressão arterial medido pelo esfigmomanômetro;
  - **Variável Discreta:** Onde o valor converge para um conjunto finito ou enumerável. Ex:Número de filhos, número de garrafas vazias;
- **Variável Qualitativa:** É a variável que pode ter seus valores separados em diferentes categorias. Uma variável qualitativa pode ainda ser classificada como:
  - **Variável Nominal:** Quando não existe uma ordem para seus valores. Ex: Estado civil, cor dos olhos.
  - **Variável Ordinal:** Quando existe uma ordem para seus valores. Ex: Grau de instrução.

3.1.3 Filtrar variáveis

Realizada as etapas anteriores (carregar os dados e checar o tipo de cada variável), o cientista pode ter a necessidade de filtrar variáveis, ou seja, selecionar apenas um subconjunto dos atributos. Isso pode ser necessário por diferentes motivos, como, por exemplo:

– Determinadas variáveis podem não ser relevantes para o problema preditivo.
– Necessidade de reduzir o tempo demandado para o treinamento do modelo preditivo.

Desta forma, após a realização dessa tarefa, gera-se um novo conjunto de dados contendo apenas as variáveis selecionadas pelo cientista, descartando-se as demais. Esse novo conjunto de dados será utilizado nas tarefas seguintes.

3.1.4 Definir os códigos para os valores faltantes

Valores ausentes(ou dados ausentes) são definidos como valores não armazenados na variável em observação. O problema da falta de dados é relativamente comum em quase todos os tipos de pesquisas e pode ter um efeito significativo nas conclusões que podem ser tiradas dos dados Graham *et al.* (2009).

No presente trabalho tratamos diferentes códigos para representar a ausência de dados, tais como: "Nan", "None", "Empty String"(Caractere vazio), NULL ou ainda códigos específicos definidos de acordo com a necessidade do estudo como por exemplo códigos de amostras, nomes de observações que deseja-se evitar no estudo. Neste sentido, é fundamental que o cientista observe e reporte todos os códigos utilizados para representar valores faltantes no conjunto de dados por ele utilizado.

3.1.5 Checar valores faltantes

Checar os valores faltantes de um experimento é importante pois dados ausentes apresentam vários problemas. O primeiro, a ausência de dados reduz a eficiência de métodos estatísticos. Em segundo lugar, os dados perdidos podem causar um viés na estimativa dos parâmetros. O viés mencionado consiste na diferença que ocorre entre um valor estimado e seu valor real, uma vez realizada a análise. Terceiro, pode reduzir a representatividade das amostras. Quatro, pode complicar as análises do estudo Kang (2013).

Após a definição dos códigos utilizados no conjunto de dados para representar valores faltantes na seção anterior, é importante averiguar os seguintes pontos:

– Se existem valores faltantes no conjunto de dados;
– Quais variáveis possuem valores faltantes;
– O percentual de valores faltantes em cada variável;
– Se existem instâncias com elevada quantidade de valores faltantes.

Na literatura existem métodos que preenchem esse valores faltantes como a impu-

tação simples, imputação múltipla e a remoção de linhas do conjunto de dados que contenha valores faltantes no conjunto de dados.

A imputação de dados faltantes tem sido uma estratégia comum para a análise de dados com esse problema. Entende-se por imputação a técnica de preencher os dados faltantes com valores plausíveis. Um atrativo para a utilização de técnicas de imputação é o fato de, após a imputação dos dados, o investigador poder utilizar técnicas tradicionais de análise estatística para dados completos Tang *et al.* (2005).

Métodos simples como imputação pela média ou pela mediana, também conhecidos como métodos de imputação única, têm sido bastante usados devido à sua facilidade de implementação. Entretanto, existem desvantagens na utilização desses métodos, tais como a subestimação da variabilidade da variável imputada que gerará intervalos de confiança mais estreitos do que o esperado e a impossibilidade de levar em consideração a variabilidade que possa existir entre diferentes imputações Heijden *et al.* (2006).

Como alternativa à imputação única e com o objetivo de corrigir suas desvantagens, surgiu a imputação múltipla. A ideia da imputação múltipla é a de que cada dado ausente é imputado m vezes, gerando m bancos de dados completos. Os m bancos são analisados separadamente por uma técnica tradicional de análise estatística e finalmente os m resultados obtidos são combinados de maneira simples para a análise final

3.1.6 Mostrar relatório de valores faltantes

Por fim, recomenda-se exibir a quantidade de valores faltantes para cada variável, incluindo todos os códigos definidos anteriormente. Essa etapa é importante para avaliar como se encontra o conjunto de dados em relação aos valores faltantes, bem com para analisar a necessidade e a possibilidade de utilização de técnicas para tratar o problema de valores ausentes, tais como remoção de variáveis, remoção de instâncias ou imputação de valores.

3.1.7 Exibir estatísticas descritivas e os tipos computacionais das variáveis

Estatística descritiva é o ramo da estatística que visa aplicar diferentes técnicas com a finalidade de descrever, organizar e resumir um determinado conjunto de dados. Diferencia-se da estatística inferencial, que tem por objetivo usar os próprios dados para extrair afirmações sobre a população alvo do estudo (GUEDES *et al.*, 2005).

Visando descrever um determinado conjunto de dados, a estatística descritiva faz uso

de medidas de resumo, tais como as medidas de tendência central e as medidas de dispersão. As medidas de tendência central informam sobre a posição típica dos dados. Como exemplos de medidas de tendência central podemos citar: média, mediana e moda.

Já as medidas de dispersão visam determinar a variabilidade dos dados. Podemos listar como medidas de dispersão: a amplitude, quantis, variância, desvio padrão e coeficiente de variação (FARIAS, 2006).

O guia proposto recomenda que, para as variáveis numéricas, o cientista de dados compute e analise as seguintes estatísticas descritivas:

- Número de linhas;
- Média aritmética;
- Desvio padrão;
- Coeficiente de variação;
- Menor valor ou limite inferior;
- Valor máximo ou limite superior;
- Percentis (25%, 50%, 75%): Os percentis são medidas que dividem a amostra em 100 partes, cada uma com uma percentagem de dados aproximadamente igual. O k-ésimo percentil Pk é o valor que corresponde à frequência cumulativa de $(N * k)/100$. Onde nosso N utilizado é 25, 50, 75 referentes as porcentagens de 25%, 50%, 75%.

Já para as variáveis categóricas, recomenda-se verificar:

- Número de linhas;
- Número de valores distintos;
- Elemento mais frequente
- Frequência do elemento mais frequente.

3.1.8 Mostrar o histograma de cada variável numérica

Um histograma, também conhecido como distribuição de frequências, é a representação gráfica em colunas ou em barras (retângulos), sobre um eixo horizontal, de um conjunto de dados, referente a uma variável de interesse, previamente e tabulado e dividido em classes. A base de cada retângulo representa uma classe. A altura de cada retângulo representa a quantidade ou a frequência absoluta com que o valor da classe ocorre no conjunto de dados para classes uniformes ou a densidade de frequência para classes não uniformes (FARIAS, 2006). A construção de histogramas tem caráter preliminar em qualquer estudo e é um importante indicador

da distribuição de dados. Os histogramas podem indicar se uma distribuição se aproxima de uma função normal, assim como também podem indicar a mistura de populações quando se apresentam bimodais. Neste sentido, recomenda-se exibir um histograma para cada variável numérica com a finalidade de visualizar graficamente a proporção dos valores presentes nessas variáveis.

3.1.9 Mostrar a proporção de cada categoria presente em cada variável categórica

Para cada variável categórica indica-se exibir um gráfico de pizza, se esta possuir até três categorias, ou um gráfico de barras, se esta possuir mais de três categorias, com a finalidade de ilustrar a proporção de cada categoria presente na variável categórica. A razão para essa escolha é de que o gráfico de setores (gráfico de pizza) a partir de mais de 3 categorias pode apresentar distorções quanto as frações dos seus setores, não representando fielmente a divisão de cada setor, sendo assim mais adequado utilizar o gráfico de barras. A Figura 12 exibe exemplos dos gráficos citados.

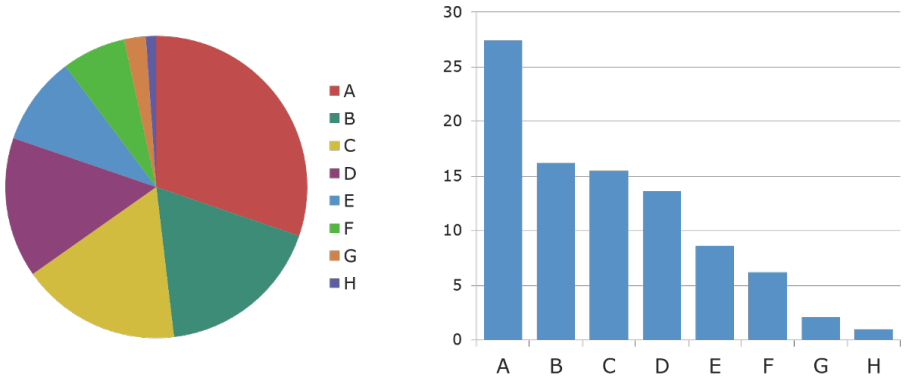


Figura 12 – Exemplo de gráficos. Fonte: Autor.

3.1.10 Exibir a distribuição do "Bestfit" para cada variável contínua

Uma distribuição estatística de um conjunto de dados é uma função que projeta todos os valores possíveis dos dados juntamente com a frequência com que estes ocorrem. Essa função pode descrever tanto o comportamento de uma variável aleatória contínua quanto de uma variável discreta (MORETTIN; BUSSAB, 2017). Portanto, há dois tipos de distribuição de probabilidade:

– Distribuição contínua: quando a variável que está sendo medida é expressa em uma escala contínua, como no caso de uma característica dimensional e

– Distribuição discreta: quando a variável que está sendo medida só pode assumir certos valores, como por exemplo os valores inteiros: 0,1,2, etc.

Como exemplos de distribuições discretas temos: Bernoulli e Binomial. Já como exemplos de distribuições contínuas podemos citar: Normal e Uniforme. Uma distribuição estatística demonstra a concentração de dados de uma variável e pode ser utilizada para modelar incertezas e descrever fenômenos físicos, biológicos, econômicos, dentre outros. Existem diversas distribuições estatísticas, mas, dentre elas, a distribuição normal tem uma importância particular uma vez que ela representa o comportamento de diferentes fenômenos comuns, como por exemplo, altura ou peso de uma população, a pressão sanguínea de um grupo de pessoas, o tempo que um grupo de estudantes gasta para realizar uma prova, dentre outros. Adicionalmente, a distribuição normal pode ser usada para aproximar distribuições discretas, como por exemplo a distribuição binomial.

Neste contexto, para avaliar a normalidade de uma determinada variável contínua, ou seja, se a distribuição dos dados da variável assemelha-se à distribuição Normal, recomenda-se a utilização de diferentes testes, tais como: K-quadrado de D'Agostino (D'AGOSTINO, 1970), Lilliefors (LILLIEFORS, 1967) e Shapiro-Wilk (SHAPIRO; WILK, 1965). Adicionalmente, pode ser relevante descobrir qual distribuição estatística mais se aproxima da distribuição dos dados de uma determinada variável, o que é conhecido como heurística de "*bestfit*". Neste caso, sugere-se a utilização do teste de Kolmogorov-Smirnov (SMIRNOV, 1948).

A análise de normalidade visa determinar a veracidade de uma hipótese (HIRAKATA *et al.*, 2019). Neste sentido, deseja-se verificar se uma determinada variável segue uma distribuição específica, no caso a distribuição normal, para aferição de normalidade. Neste caso, temos que nossa hipótese nula (H0) é que a variável segue uma distribuição normal. Já a hipótese contrária, chamada de hipótese alternativa (H1), é que a variável não segue uma distribuição normal. Visto isso, usamos os testes de D'Agostino (D'AGOSTINO, 1970), Lilliefors (LILLIEFORS, 1967), Shapiro-Wilk (SHAPIRO; WILK, 1965) para testar a hipótese que a variável numérica segue uma distribuição normal ou não, onde cada método retorna se a hipótese H0 é rejeitada ou aceita.

A heurística de "*Bestfit*"utiliza como entrada um conjunto de distribuições contínuas, onde cada distribuição neste conjunto será comparada com a distribuição dos dados da variável. Em seguida, seleciona-se aquela que mais se assemelha à distribuição da variável. Assim, a heurística de "*Bestfit*"possui as seguintes etapas: escolher as distribuições para comparação

com a distribuição da variável analisada, computar a semelhança entre a distribuição teórica e a distribuição da variável e aferir a distribuição teórica que mais se aproxima da distribuição da variável em análise.

3.1.11 Mostrar coeficientes de correlação para cada par de variáveis

Uma das principais contribuições da estatística para ampliar o entendimento humano sobre os fenômenos observados foi a capacidade de medir a relação entre diferentes variáveis. Os coeficientes de correlação auxiliam os cientistas a mensurar essa relação. Neste sentido, os coeficientes de correlação são métodos estatísticos utilizados para mensurar as relações entre variáveis e o que elas representam. Mais especificamente, a correlação busca entender como uma variável se comporta em um cenário onde outra está variando, tentando identificar se existe alguma relação entre a variabilidade de ambas. Embora não implique em causalidade, a correlação quantifica a relação entre a variabilidade de duas variáveis.

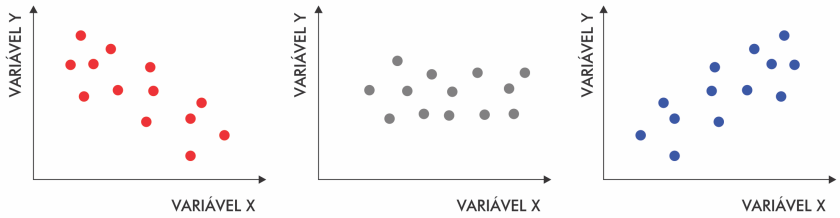


Figura 13 – Exemplo de correlações entre duas variáveis. Fonte: link

Existem diferentes coeficientes de correlação que podem ser utilizados para medir o grau de uma correlação. Um dos coeficientes de correlação mais conhecidos é o coeficiente de correlação de Pearson (sensível a uma relação linear entre duas variáveis). Contudo, existem outros coeficientes de correlação mais robustos que o coeficiente de correlação de Pearson, ou seja, mais sensíveis às relações não lineares.

Neste sentido, recomenda-se a utilização de diferentes coeficientes de correlação com base na distribuição dos dados. Os coeficientes de correlação de Spearman(SPEARMAN, 1961) devem ser exibidos para todos os pares de variáveis numéricas contínuas. Para cada par de variáveis que possuem distribuição normal, os coeficientes de correlação de Pearson (PEARSON, 1895) devem ser calculados e exibidos.

3.1.12 Mostrar o valor da medida V de Cramer para cada par de variáveis categóricas

O V de Cramer (Cramer's V) é uma medida que busca capturar a força da associação entre duas variáveis categóricas, com duas ou mais categorias. O valor desta medida varia de 0 a +1, onde um valor mais próximo de 1 indica uma forte associação (CRAMÉR, 1999). Portanto, se o conjunto de dados utilizado possui variáveis categóricas é recomendado exibir a medida V de Cramer para cada par de variáveis categóricas.

3.2 Fase 2 - Pré-processamento dos dados

O pré-processamento de dados é uma etapa essencial na solução de problemas preditivos. O principal objetivo da segunda fase do guia proposto, chamada de pré-processamento dos dados, consiste em preparar os dados para que estes possam ser utilizados na construção de modelos preditivos. Esta fase inclui atividades relacionadas à detecção de valores discrepantes, normalização de dados, escolha da variável independente, seleção de atributos, balanceamento de dados, e divisão dos conjuntos de treinamento e teste. A Figura 14 ilustra as atividades que compõem a segunda fase do guia proposto.

3.2.1 Escolher a variável dependente

Nesta etapa, a variável dependente deve ser definida pelo cientista de dados. Uma variável independente, normalmente representada pela letra x, caracteriza uma grandeza que está sendo manipulada durante um experimento e que não sofre influência de outras variáveis. Já a variável dependente, normalmente representada pela letra y, é aquela que sofre influência dos demais atributos (chamados variáveis independentes), ou seja, de forma direta ou indireta x exerce influência sobre y. Um modelo preditivo utiliza as informações das variáveis independentes para estimar o valor da variável dependente. Por exemplo, suponha que queremos prever a quantidade de sorvete que serão vendidos em um determinado dia. Logicamente, a estação do ano pode influenciar as vendas de sorvete. Neste caso, a quantidade de sorvetes vendidos será a variável dependente e a estação do ano uma variável independente. Uma vez definida a variável dependente, os demais atributos do conjunto de dados serão considerados variáveis independentes.

Um exemplo pratíco ilustrado na Figura 15 é através da relação entre a expectativa de vida e um índice de felicidade calculado em diversos países obtidos a partir de um levantamento

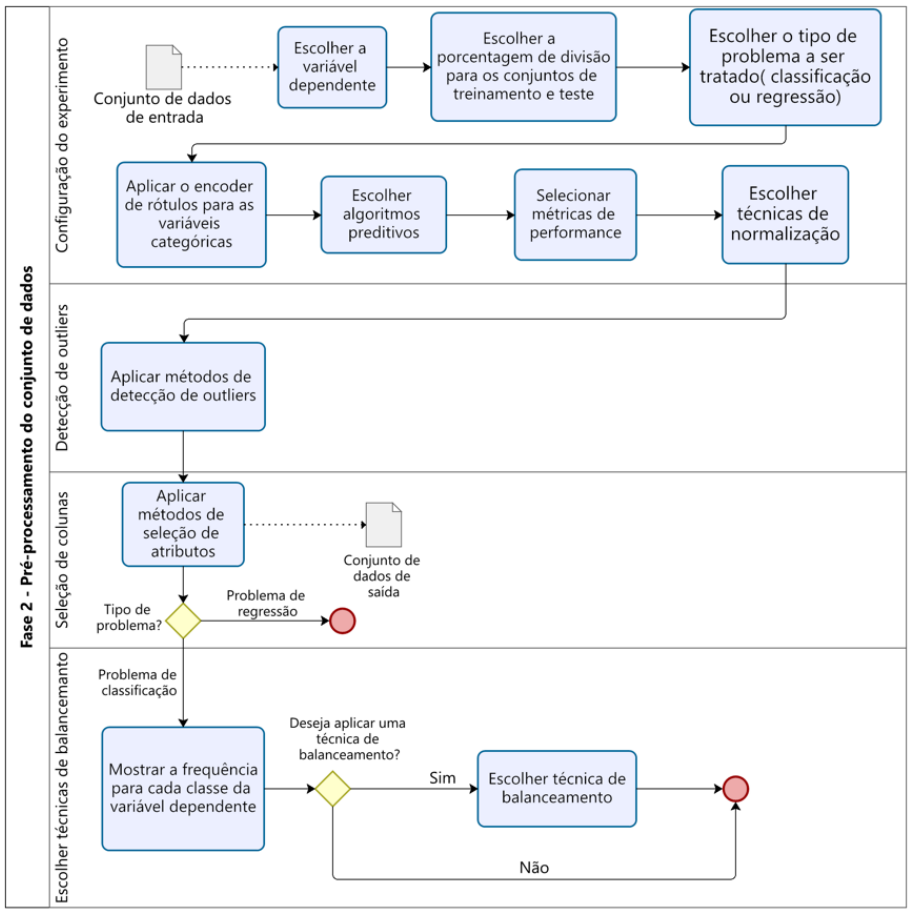


Figura 14 – Fase 2 - Pré-processamento dos dados. Fonte: Autor.

feito por Helliwell *et al.* (2020). A variável independente nesse exemplo é representada pelo índice de felicidade e a expectativa de vida age como variável dependente, dessa forma pode ser observada uma tendência de expectativa de vida maior em países com alto índice de felicidade, com uma força de correlação de 0,77.

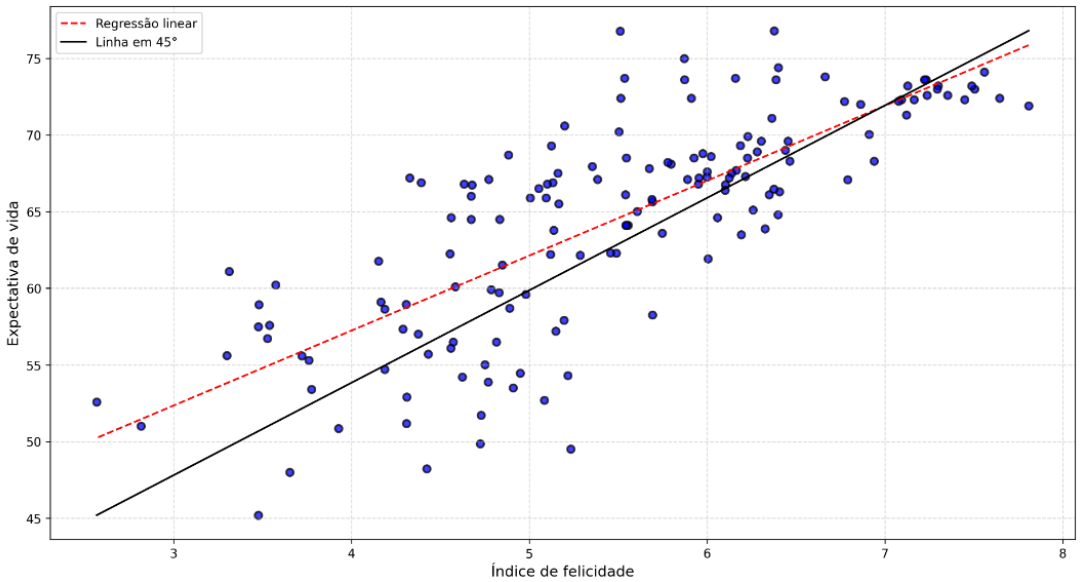


Figura 15 – Relação entre o índice de felicidade e expectativa de vida. Extraído de Helliwell *et al.* (2020).

3.2.2 Escolher a porcentagem de divisão para os conjuntos de treinamento e teste

Nesta etapa, o cientista de dados deve definir a proporção a ser utilizada para dividir o conjunto de dados original nos conjuntos de treino e teste, os quais podem ser entendidos da seguinte forma:

- Conjunto de treino: Conjunto de dados utilizado com a finalidade de gerar (treinar) modelos preditivos.
- Conjunto de teste: Conjunto de dados utilizado com a finalidade de avaliar o desempenho dos modelos preditivos.

Um proporção bastante utilizada na literatura é de 70% pra treino e 30% para o conjunto de teste. A escolha inadequada desta proporção pode conduzir o modelo preditivo a dois problemas bastante conhecidos: *overfitting* e o *underfitting*. O *underfitting* ocorre quando o modelo não se adapta bem ao conjunto de dados de treino e não é capaz de prever corretamente os dados do conjunto de teste. Já o *overfitting* ocorre quando o modelo se adapta exageradamente bem ao conjunto de dados de treino, mas não é capaz de prever corretamente os dados do conjunto de teste. A incapacidade de um modelo de capturar a verdadeira relação entre variáveis e o objeto a ser predito é o que chamamos de VIÉS (Bias em inglês). Então, quando o erro de viés é alto significa que o modelo não está aprendendo nada (*underfitting*). A variância é a sensibilidade de um modelo ao ser usado com outros datasets diferentes do treinamento. Se o modelo é muito sensível aos dados de treinamento, ou seja, identificou tão bem a relação entre

os dados de treinamento que quando colocado em teste irá errar justamente a variação que existe entre os datasets (*overfitting*). Na Figura 16 temos as diferenciações entre como os dados se comportam em cenários de *underfitting*, *overfitting* e o estado desejável para um modelo de aprendizagem de máquina.

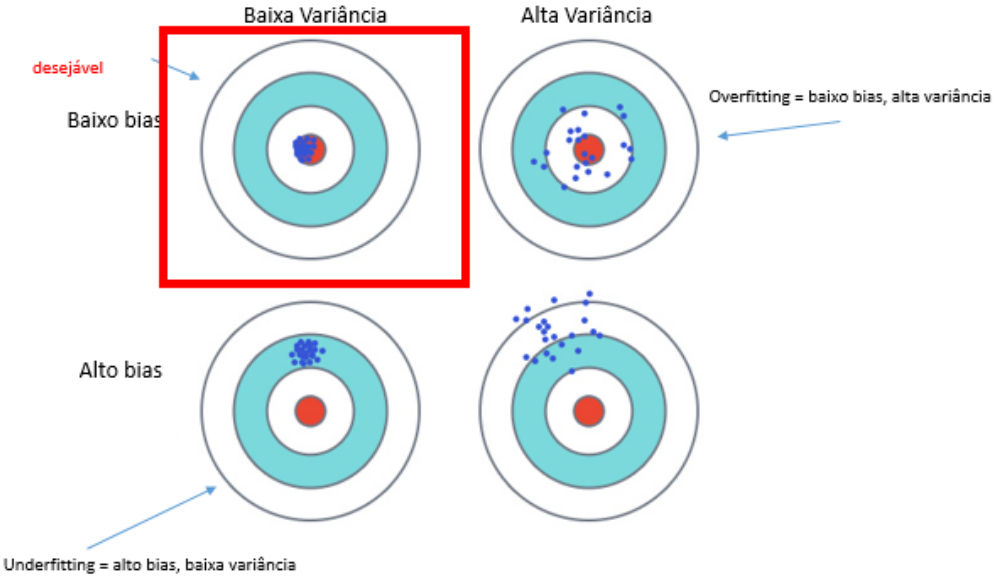


Figura 16 – Ilustração gráfica de Viés e Variância. Fonte: link.

Há diversas formas de dividir o conjunto de dados original nos conjuntos de treinamento e teste. Uma estratégia bastante utilizada é conhecida como *K-fold* (KUHN *et al.*, 2013). Nesta abordagem, as amostras são divididas aleatoriamente em k conjuntos do mesmo tamanho. Um modelo preditivo é treinado usando todos os dados exceto os do primeiro subconjunto, depois esse primeiro subconjunto é utilizado para avaliar o modelo, utilizando-se para isso uma determinada métrica. Essa estratégia é repetida até que todas os subconjuntos tenham sido utilizados para teste.

3.2.3 Escolher o tipo de problema a ser tratado (regressão ou classificação)

Após a identificação da variável dependente e das variáveis independentes, consiste em definir que tipo de problema preditivo será investigado. Quando a variável dependente é contínua, ou seja, assume valores reais, a tarefa preditiva é do tipo de regressão. Quando a variável dependente é categórica, a tarefa preditiva é do tipo de classificação. Definir se o problema investigado é uma regressão ou classificação é bastante importante, uma vez que

existem métodos, tratamento de dados e métricas de avaliação que são específicos para cada tipo de problema. Logo, identificar o tipo de problema investigado é um dos primeiros passos para saber como lidar com ele.

3.2.4 Aplicar o encoder de rótulos para as variáveis categóricas

O conjunto de dados utilizado pelo cientista pode conter variáveis categóricas. Contudo, alguns algoritmos preditivos não funcionam adequadamente com variáveis categóricas. Além disso, sabe-se que valores numéricos, em geral, produzem modelos preditivos mais eficazes. Neste contexto, uma abordagem bastante recomendada consiste em converter as variáveis categóricas em variáveis numéricas, o que é chamado de "*encoder*"(SHARMA *et al.*, 2020).

Existem diferentes estratégias para a codificação de variáveis categóricas, dentre elas destacam-se:

- Label / Ordinal encoding
- One-hot/dummy encoding
- Target encoding
- Frequency / count encoding
- Binary encoding
- Feature Hashing

Dentre essas estratégias, as mais populares são: *Label Encoder* e *OneHot Encoder*. A estratégia *Label Encoder* consiste em atribuir um valor numérico a cada rótulo distinto e substituir esse valor no conjunto de dados. Recomenda-se seu uso quando os rótulos têm propriedades diferentes. Como exemplo, considere uma variável categórica chamada "tamanho da camisa", a qual pode conter os seguintes valores: Pequeno, Médio e Grande. Neste caso, a estratégia *Label Encoder* irá substituir esses valores por 1, 2 e 3, respectivamente. É correto afirmar que existe uma ordem entre esses valores: pequeno (1) < médio (2) < grande (3). Em caso de não haver uma ordem entre os valores da variável categórica, recomenda-se utilizar a estratégia *OneHot Encoder*. Esse método cria uma nova coluna para cada categoria distinta de uma variável categórica. Em seguida, a variável categórica é removida do conjunto de dados. Assim, se uma determinada variável categórica "estado civil"possui três categorias: "casado", "solteiro"e "divorciado", então 3 novas colunas serão criadas: "casado", "solteiro"e "divorciado". Além disso, a coluna "estado civil"será removida do conjunto de dados. Adicionalmente, uma

instância com valor "divorciado"na variável categórica "estado civil"será representada pelos valores 0, 0 e 1 nas colunas "casado", "solteiro"e "divorciado", respectivamente.

3.2.5 Escolher algoritmos preditivos

Nesta etapa o cientista deve selecionar os algoritmos de regressão ou classificação que serão utilizados para construir os modelos preditivos. Para problemas de regressão recomenda-se utilizar métodos que estimam valores em um intervalo contínuo através de funções como a própria regressão linear ou suas variantes como a regressão polinomial. Esses algoritmos são comumente usados para previsões relacionadas a estoques de suprimentos e valores de ações. Já para problemas de classificação recomenda-se aplicar Regressão Logística, *Naive Bayes*, *Decision Tree*,*K-Nearest Neighbours* e *Multilayer Perceptron*.

Alguns algoritmos podem ser usados para classificação e regressão com pequenas modificações, como árvores de decisão e redes neurais artificiais.

Em alguns casos, é possível converter um problema de regressão para um problema de classificação. Por exemplo, os valores de uma ação, que são contínuos, podem ser convertidos em classes que representam faixas de valores:

- Classe 0: valores entre 0 e 100 e
- Classe 1: valores maiores que 100.

Desta forma, em vez de prever o valor de uma ação deseja-se prever a sua classe, ou seja, se valor da ação é menor ou igual a 100 ou maior que 100. Essa estratégia é chamada de discretização.

Por outro lado, em alguns casos, um problema de classificação pode ser convertido em um problema de regressão. Neste caso, um rótulo pode ser convertido em um intervalo contínuo.

3.2.6 Selecionar métricas de desempenho

Nesta etapa, o cientista de dados deverá definir as métricas que serão utilizadas para avaliar desempenho dos modelos preditivos. Uma métrica de desempenho é uma medida numérica que visa medir a eficácia de um modelo preditivo (SELIYA *et al.*, 2009). Logicamente, é importante escolher métricas que sejam adequadas ao tipo de problema (regressão ou classificação), ao algoritmo utilizado e às características do conjunto de dados (balanceado ou desbalanceado, por exemplo).

Em modelos de regressão uma das métricas mais utilizadas é a raiz do erro quadrático médio, comumente chamado na literatura de Raiz do Erro Quadrático Médio (RMSE), sigla em inglês, (KUHN *et al.*, 2013). O erro quadrático médio é uma medida baseada no valor ao quadrado da diferença entre o valor obtido pela predição do modelo e o valor observado de uma amostra, onde essa diferença é chamada de resíduo. O Erro Quadrático Médio (MSE) é a média de todos os erros e, consequentemente, o RMSE é a raiz quadrada do MSE. Podemos observar que quanto menor o valor desta métrica, mais próximas estão as predições dos valores observados, sendo assim, melhor é o modelo de regressão em relação a essa métrica. Outra métrica bastante utilizada para avaliar modelos de regressão é o coeficiente de determinação R^2, (KUHN *et al.*, 2013). O R^2 é definido como um número entre 0 e 1 que define o quão bem um modelo de regressão prevê os dados. Ele pode ser interpretado como a proporção de informação nos dados que podem ser explicadas pelo modelo. O R^2 é definido como uma métrica de correlação, ou seja, não é uma métrica que mede a acurácia, diferentemente do MSE. Quanto mais próximo o R^2 é de 1, melhor o modelo consegue explicar o conjunto de dados. Existe mais de uma forma de calculá-lo, uma simples e bastante utilizada na literatura é calcular o quadrado do coeficiente de correlação entre os valores previstos e valores observados.

No caso de modelos de classificação, existem diversas outras métricas que são comumente utilizadas para avaliar o desempenho dos modelos preditivos (KUHN *et al.*, 2013). A matriz de confusão é uma tabela que indica os erros e acertos do seu modelo, comparando com o resultado esperado. Nela temos os seguires rótulos:

- Verdadeiros Positivos: classificação correta da classe desejada;
- Falsos Negativos: erro em que o modelo previu a classe não desejada quando o valor real era da classe desejada;
- Falsos Positivos: erro em que o modelo previu a classe desejada quando o valor real era classe não desejada;
- Verdadeiros Negativos: classificação correta da classe não desejada.

A Figura 17 exibe como se estrutura a matriz de confusão entre o que é real e detectado.

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 17 – Matriz de Confusão. Fonte: link

A acurácia(*accuracy*) é definida pela porcentagem de acertos dentre todas as predições, porém essa métrica não avalia os tipos de erros que podem estar ocorrendo nas predições. Em outras palavras ela avalia o percentual de acertos, ou seja, ela pode ser obtida pela razão entre a quantidade de acertos e o total de entrada .A fórmula da acurácia segue abaixo:

$$accuracy = \frac{VP+VN}{VP+VN+FN+FP} \tag{3.1}$$

Em modelos de classificação binária (os quais possuem apenas duas classes), a precisão(*precision*) é definida como a porcentagem de casos positivos previstos corretamente entre todos os casos positivos previstos (corretamente ou incorretamente). Segue a fórmula da precisão abaixo:

$$precision = \frac{VP}{VP+FP} \tag{3.2}$$

Procurar modelos com melhor precisão é desejável quando o custo um de falso positivo é muito alto, como, por exemplo, na detecção de *spams* em caixa de e-mails ou quando se deseja fazer um investimento financeiro em uma carteira virtual não tão precisa na hora de escolher uma bolsa para aplicar a compra e posteriormente vender.

Enquanto o revocação(*recall*) é uma métrica definida como a porcentagem de verdadeiros positivos entre todas as amostras verdadeiramente positivas. Segue a fórmula da revocação abaixo:

$$recall = \frac{VP}{VP+FN} \tag{3.3}$$

Modelos com o maior *recall* são escolhidos normalmente quando a quantidade de falsos positivos é muito alto, como, por exemplo, na detecção de fraudes em transações bancárias ou quando um determinado classificador para avaliar a condição de pacientes classifica os pacientes doentes como saudáveis.

Uma das métricas que combina *recall* e precisão em um único valor é conhecida como $F1$-score que é dada por $2 \times recall \times precisão / (recall + precisão)$. O $F1$-score é preferível em relação à acurácia quando queremos equilibrar escolhas entre recall e precisão e os dados das classes são mais desbalanceados (SELIYA *et al.*, 2009).

Outras métricas para classificação podem usar as probabilidades das classes no lugar das classes previstas. As probabilidades tem possibilidade de mostrar mais informação sobre o modelo do que os valores das classes. A área abaixo da curva *Receiver operating characteristic* (ROC) (BRADLEY, 1997) é um valor entre 0 e 1 e seu valor é usado para avaliar o *trade-off* entre a taxa de positivos previstos corretamente e a taxa de falso positivos. Um modelo classificador que fornece uma grande área sob a curva é geralmente preferível a uma classificador com uma área menor sob a curva, quando a área sob a curva ROC é 1, indica que ele consegue distinguir as duas classes perfeitamente (SELIYA *et al.*, 2009).

3.2.7 Escolher técnicas de normalização

A normalização e a padronização são técnicas frequentemente aplicadas na etapa de preparação dos dados, com o objetivo de colocá-los em um intervalo de valores comuns, a fim de evitar que o modelo preditivo fique enviesado para as variáveis com maior ordem de grandeza. Logo, essa técnicas podem impactar diretamente no desempenho do modelo preditivo. Alguns algoritmos precisam que os dados estejam na mesma escala, tais como: KNN(*K-Nearest Neighbours*), Redes Neurais, Regressão Linear, Regressão Logística e SVM(*Support Vector Machine*).

Tanto a normalização quanto a padronização possuem o mesmo objetivo: transformar todas as variáveis na mesma ordem de grandeza. A diferença básica é que a padronização de uma determinada variável irá resultar em valores com uma média igual a 0 e um desvio padrão igual a 1. Já a normalização irá resultar em valores dentro do intervalo de 0 e 1, e caso tenha resultado negativo entre -1 e 1. Vale destacar que ao utilizar uma rede neural do tipo *Feedforward*, recomenda-se que os dados sejam normalizados entre 0,1 e 0,9, ao invés de 0 e 1 para evitar a saturação da função sigmóide (BASHEER; HAJMEER, 2000).

Neste sentido, recomenda-se que o cientista de dados avalie pelo menos duas técnicas diferentes: *z-score* e "normalização Min-Máx". A padronização *z-score* resultará em valores com uma média igual a zero e desvio padrão igual a q, enquanto a "normalização Min-Máx"resultará os valores das variáveis re-mapeados para o intervalo de zero a um. A seguir, iremos detalhar

essas duas estratégias.

Normalização Mín-Máx Jain e Bhandare (2011) realiza uma transformação linear nos dados originais. Suponha que min_A e max_A sejam os valores mínimo e máximo de um atributo A. A normalização Mín-Máx mapeia um valor, v_i, de A para v'_i no intervalo [$newmin_A$, $newmax_A$] calculando:

$$v'_i = newmin_A + (newmax_A - newmin_A)\left(\frac{v_i - min_A}{max_A - min_A}\right) \tag{3.4}$$

A normalização Min-Máx preserva os relacionamentos entre os valores de dados originais. Ele encontrará um erro "fora dos limites"se um caso de entrada futuro para normalização ficar fora do intervalo de dados original para *A* Skiena (2017).

Padronização com *Z-Score* Han *et al.* (2012) pode ser calculada usando a seguinte equação:

$$z = \frac{x - \mu_A}{\sigma_A} \tag{3.5}$$

Na equação 3.5, μ_A e σ_A são a média e o desvio padrão do traço A. O traço original e o traço normalizado são representados por x e z, respectivamente. Após a padronização, a média e o desvio padrão de todos os recursos tornam-se 0 e 1, respectivamente.

3.2.8 Aplicar métodos de detecção de outliers

Outliers são valores extremos que se desviam de outras observações nos dados (ou seja, uma observação que diverge de um padrão geral em uma amostra).

Outliers são um dos principais problemas enfrentados ao se construir um modelo preditivo. Eles podem influenciar negativamente o desempenho de um modelo preditivo, principalmente se forem decorrentes de erros na coleta dos dados, os quais podem ocorrer, por exemplo, devido a falhas em um sensor de temperatura ou erros de digitação. Neste caso, os *outliers* podem ser descartados. Portanto, é de fundamental importância utilizar métodos eficazes para a identificação de *outliers* (LIU *et al.*, 2004). Existem duas técnicas principais para detectar *outliers*: intervalo interquartil (uma abordagem paramétrica univariada) e *boxplot* ajustado (uma abordagem não paramétrica univariada). Técnicas univariadas se restringem a análise a somete uma variável, já ao fator de paramétrica é se caso conheça os parâmetros da distribuição

utiliza-se preferencialmente técnicas paramétricas e caso não se tenha precisão dos parâmetros das distribuições é recomendado utilizar técnicas não paramétricas. A seguir, descrevemos em detalhes as abordagens do intervalo interquartil e boxplot ajustado.

Intervalo interquartílico: O Intervalo interquartílico(intervalo interquartílico(IQR) (IQR)) é uma medida de dispersão estatística, frequentemente usada para detectar *outliers*. O IQR é o comprimento da caixa no *boxplot* (ou seja, Q3 - Q1). Aqui, *outliers* são definidos como instâncias abaixo de $Q1 - 1,5 * IQR$ ou acima de $Q3 + 1,5 * IQR$.

Boxplot ajustado: Em distribuições assimétricas, o *boxplot* usual geralmente sinaliza muitos pontos de dados regulares como remotos. O *boxplot* ajustado para assimetria corrige isso usando uma medida robusta de assimetria para determinar a cerca das curvas de assimetria Hubert e Vandervieren (2008). Para medir a assimetria de uma amostra univariada (x_1, ..., x_n) de uma distribuição unimodal contínua, usamos o *medcouple*(MC) e Q2 é a mediana da amostra (BRYS *et al.*, 2004). É definido como na equação 3.6:

$$MC = med_{x_i \leq Q_2 \leq x_j} h(x_i, x_j) \tag{3.6}$$

onde $h(x_i, x_j)$ é igual a:

$$h(x_i, x_j) = ((x_j - Q_2) - (Q_2 - Q_1))/Q_3 - Q_1 \tag{3.7}$$

Nesta nova abordagem, ajustamos a assimetria do *boxplot* padrão, incorporamos o *medcouple* (MC) a definição dos intervalos. Isso pode ser feito introduzindo as funções $h_l(MC)$ e $h_u(MC)$ aos valores de corte do intervalo interquartílico. Então, os *outliers* serão definidos como instâncias abaixo de $Q1 - h_l(MC) * IQR$ ou acima de $Q3 + h_u(MC) * IQR$. É importante frisar que é exigido que $h_l(0) = h_u(0) = 1,5$ para obter o *boxplot* padrão em distribuições simétricas.

Observe que, usando diferentes funções *hl* e *hu*, permitimos que a cerca seja assimétrica ao redor da caixa, de modo que o ajuste da assimetria seja realmente possível. Em (HUBERT; VANDERVIEREN, 2008), três modelos diferentes foram estudados, eles são:

– Modelo linear:

$$h_l(MC) = 1.5 + a * MC \qquad h_u(MC) = 1.5 + b * MC \tag{3.8}$$

– Modelo quadrático:

$$h_l(MC) = 1.5 + a_1 * MC + a_2 * MC\text{ˆ}2 \quad h_u(MC) = 1.5 + b_1 * MC + b_2 * MC\text{ˆ}2 \tag{3.9}$$

– Modelo exponencial:

$$h_l(MC) = 1.5e^{aMC} \qquad h_u(MC) = 1.5e^{bMC} \tag{3.10}$$

com $a, a1, a2, b, b1, b2 \in \mathbb{R}$.

3.2.9 Aplicar métodos de seleção de atributos

Durante o processo de tratamento dos dados, ou pré-processamento, um dos aspectos que devem ser observados é a ocorrência de atributos (variáveis independentes) redundantes. Um atributo redundante pode ser definido como um atributo que pode ter seu valores inferidos por meio de outros atributos já presentes nos dados. Ou seja, tem uma forte correlação com outros atributos e sua presença acaba não contribuindo para a melhoria do desempenho do modelo preditivo (KUHN *et al.*, 2013).

A seleção de atributos refere-se ao processo de obtenção de um subconjunto dos dados originais de acordo com um determinado critério. Neste sentido, atributos redundantes podem ser removidos, com a finalidade de melhorar o desempenho dos modelos preditivos. Existem diversas vantagens em remover atributos considerados redundantes, a redução do volume de dados pode se tornar extremamente vantajosa em casos que a memória computacional pode ser um problema, ou também a redução de tempo computacional usado para o treinamento dos modelos preditivos (CAI *et al.*, 2018).

Os métodos de seleção de atributos se enquadram em três categorias: *Filters*, *Wrappers* e métodos *Embedded/Hybrid*. Os métodos *Filters* levam menos tempo computacional para selecionar os melhores atributos. Como a correlação entre as variáveis dependentes é considerada ao selecionar as variáveis, isso leva à seleção de atributos redundantes (VENKATESH; ANURADHA, 2019) com a utilização de métodos como:

– Para variáveis numéricas contínuas:
  – Correlação de Pearson: É usado como uma medida para quantificar a dependência linear entre duas variáveis contínuas.
– Para variáveis numéricas categóricas:
  – Qui-Quadrado: É um teste estatístico aplicado as variáveis categóricas para avaliar a probabilidade de correlação ou associação entre eles usando sua distribuição de frequência.

Os métodos *Wrapper* são feitos como uma caixa preta, que são métodos de seleção

de variáveis de força bruta que avaliam exaustivamente todas as combinações possíveis dos dados de entrada para encontrar o melhor subconjunto. Podemos citar métodos como:

- *Forward Selection*: É um método iterativo no qual selecionamos nenhuma variável ao modelo. Para cada variável no modelo adicionamos uma variável que melhora o aprimoramento do algoritmo até que a adição de uma nova variável não melhore mais o desempenho do modelo.
- *Backward Elimination*: Diferente no método anterior começamos nossa iteração com todas as variáveis e a cada iteração observamos se com a eliminação de alguma variável o modelo possui algum aprimoramento. Se não houver nenhum melhoramento o algoritmo se encerra.
- *Recursive Feature Elimination*: É um algoritmo de otimização guloso que visa encontrar o subconjunto de variáveis com melhor desempenho. Ele cria modelos repetidamente e mantém de lado a variável de melhor ou pior desempenho em cada iteração. Ele constrói o próximo modelo com as variáveis da esquerda até que todos os atributos estejam esgotados. Em seguida, ele classifica as variáveis com base na ordem de eliminação.

Os métodos *Embedded/Hybrid* combinam as vantagens de ambas as abordagens, *Filters* e *Wrapper*. Uma abordagem híbrida usa a função de avaliação de desempenho do subconjunto de atributos e o teste de independência entre as variáveis (VEERABHADRAPPA; RANGARAJAN, 2010).

3.2.10 Mostrar a frequência para cada classe da variável dependente

Caso o cientista esteja solucionando um problema de classificação é importante verificar como se configuram as classes da variável dependente. Neste sentido, para cada classe presente na variável dependente deve-se computar a porcentagem de suas proporções.

3.2.11 Escolher técnica de balanceamento

Em problemas de classificação, dizemos que um conjunto de dados é desbalanceado quando as quantidades de instâncias em cada uma das classes presentes na variável dependente estão desequilibradas. É importante notar que a maioria dos algoritmos de classificação funcionam melhor quando os números de instâncias em cada classe são semelhantes ou próximos (LONGADGE; DONGRE, 2013). Em geral, quando temos dados desbalanceados dizemos que a classe majoritária domina a classe minoritária. Neste sentido, um determinado classificador

pode ser mais inclinados para a classe majoritária, o que pode comprometer o seu desempenho (KOTSIANTIS *et al.*, 2006).

Visando enfrentar esse problema, diversas estratégias foram propostas, as quais podem ser classificadas em três categorias: textitOver-Sampling (YAP *et al.*, 2014), *Under-Sampling* (LIU *et al.*, 2008) e *Hybrid Methods* (GULATI, 2020).

Over-sampling(*Over-sampling* (OS)): As abordagens de *Over-sampling* buscam aumentar a proporção da classe minoritária, criando novas instâncias da classe minoritária. Essas abordagens são bastante populares. A seguir, iremos detalhar alguns métodos de *Over-sampling*:

- *Random Over Sampler*: este método busca aumentar a proporção da classe minoritária criando instâncias sintéticas de forma randômica ou ainda repetindo amostras originais selecionadas aleatoriamente. O principal problema deste método é que a variabilidade das amostras não muda (LI *et al.*, 2013).
- *Synthetic Minority Oversampling TEchnique(Synthetic Minority Oversampling TEchnique (Smote))*: este método aumenta o número de amostras da classe minoritária gerando novas instâncias sintéticas para igualar com a classe majoritária (CHAWLA *et al.*, 2002).

Under-sampling(*Under-sampling* (US)): *Under-sampling* é uma das estratégias mais simples para lidar com o problema do desbalanceamento de dados, pois consiste em reduzir os dados da classe majoritária para equilibrar com a classe minoritária (FERNÁNDEZ *et al.*, 2018). Existem diferentes modelos de subamostragem, como *Edited Nearest Neighbor*(*Edited Nearest Neighbor* (ENN))(GUAN *et al.*, 2009), *Random Under-Sampling*(*Random Under-Sampling* (RUS))(BATISTA *et al.*, 2004) e *SMOTE-Tomek* (ELHASSAN; ALJURF, 2016), que são os mais populares. A ideia básica das técnicas mencionadas pode ser visualizada na Figura 18 [1].

[1] https://blog.strands.com/unbalanced-datasets

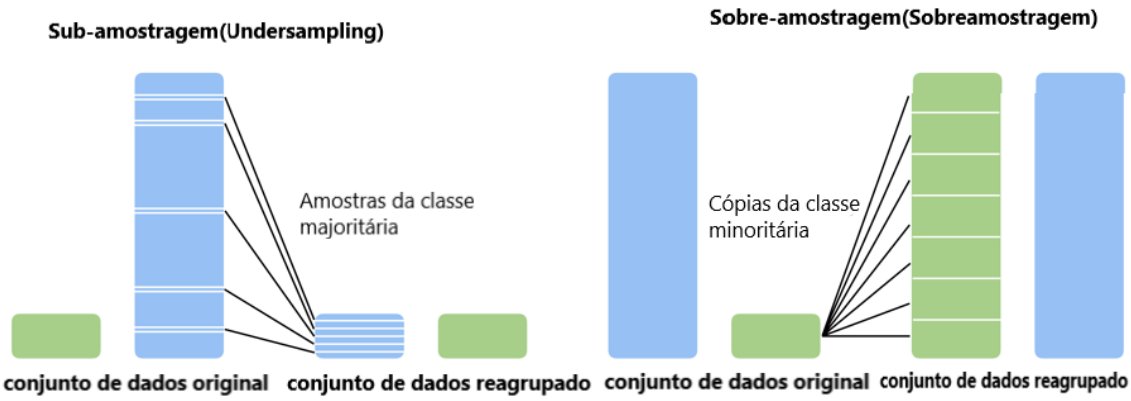


Figura 18 – Técninas de *Undersampling* e *Oversampling* (1). Fonte: Autor.

3.3 Fase 3 - Construção de Modelos Preditivos

A última fase do guia proposto tem por objetivo gerar modelos preditivos, analisar seus resultados e assegurar que os experimentos possam ser reproduzidos. A Figura 19 ilustra as atividades que compõem a terceira fase do guia proposto.

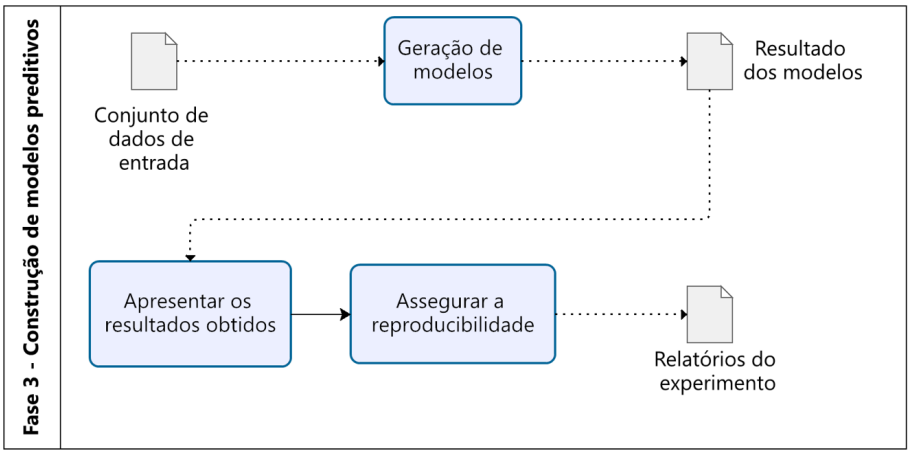


Figura 19 – Fase 3 - Construção de modelos preditivos. Fonte: Autor.

3.3.1 Geração de modelos

Essa primeira etapa visa a construção de modelos com a utilização do conjunto de dados gerado da Fase 2 e das escolhas feitas nas etapas anteriores, tais como a escolha dos algoritmos, métricas, variável dependente e porcentagem dos conjuntos de treinamento e de teste. A Figura 20 ilustra as atividades envolvidas na geração dos modelos preditivos, tais como o particionamento dos conjuntos de treino e teste, aplicação de técnicas de balanceamento dos dados e ajustes dos valores de hiperparâmetros.

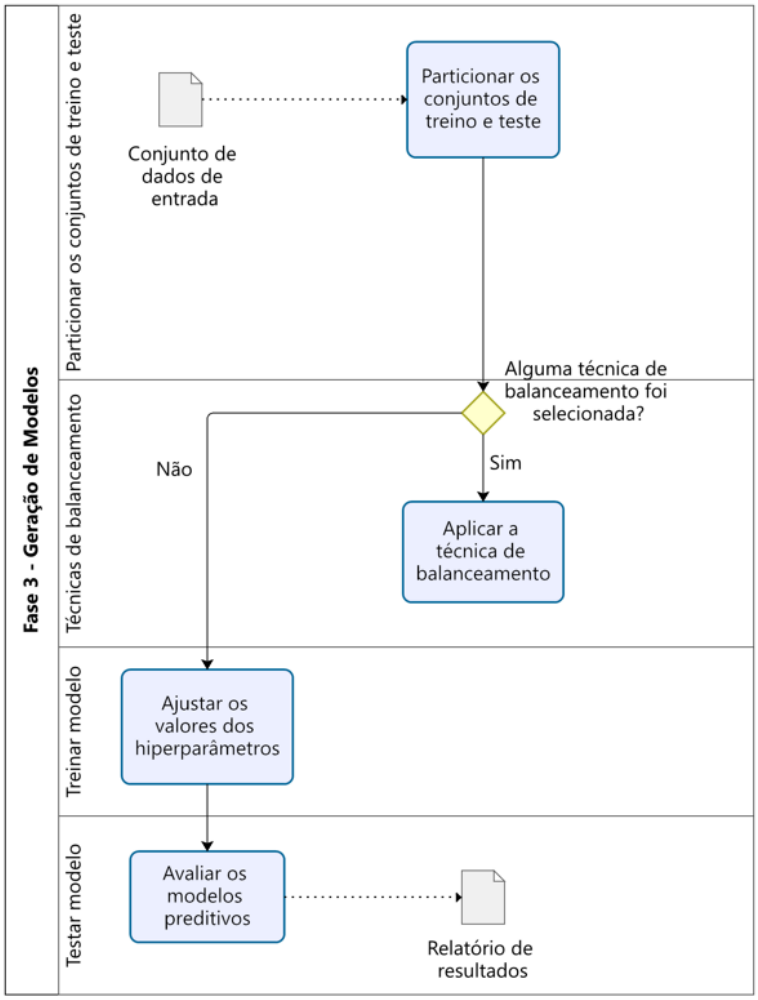


Figura 20 – Detalhamento da etapa de "geração de modelos". Fonte: Autor.

3.3.1.1 Particionar os conjuntos de treino e teste

Nessa etapa, o cientista deve aplicar a estratégia de particionamento selecionada na etapa anterior, chamada "Escolher a porcentagem de divisão para os conjuntos de treinamento e teste". Para a execução dessa tarefa é essencial que o cientista fique atento à execução do método de particionamento, o que é realizado, geralmente, com o suporte de um software.

3.3.1.2 Aplicar técnicas de balanceamento de dados

Nesta etapa, caso o cientista tenha verificado que o conjunto de dados por ele utilizado está desbalanceado, técnicas de balanceamento devem ser avaliadas. Sugere-se a utilização de pelo menos uma técnica de *Over-Sampling* e uma de *Under-Sampling*. Porém, quanto maior a diversidade de técnicas avaliadas, maior a probabilidade de se obter modelos preditivos com melhores desempenhos. Vale destacar ainda que, o cientista deve sempre avaliar

a geração de modelos preditivos usando o conjunto de dados original (desbalanceado).

3.3.1.3 Ajustar os valores dos hiperparâmetros

Todo algoritmo de predição tem um conjunto de hiperparâmetros, os quais controlam o processo de aprendizagem e determinam o seu desempenho (HUTTER *et al.*, 2019). Nesta etapa, o cientista de dados deve aplicar técnicas que possibilitem encontrar valores para os hiperparâmetros do algoritmo utilizado que proporcionem modelos preditivos com melhores desempenhos. A "otimização"de hiperparâmetros pode levar em conta um ou vários objetivos, tais como redução do tempo de processamento ou do consumo de recursos computacionais, tais como memória ou espaço em disco.

Uma das técnicas mais simples para a "otimização"de hiperparâmetros é chamada de *Grid Search*. Este método nada mais é que uma busca exaustiva feita com valores específicos de hiperparâmetros de um algoritmo (estimador). A combinação de valores com a melhor avaliação de desempenho é escolhida e considerada "ótima". A técnica *Grid Search* avalia um espaço de busca bastante elevado ou até ilimitado, o que pode levar a tempos de execução que sejam inviáveis. Logo, é necessário limitar o espaço de busca a ser avaliado. Uma outra técnica de "otimização"de hiperparâmetros, chamada *Random Search*, percorre uma combinação aleatória de hiperparâmetros tanto em domínios discretos, contínuos e mistos. O método *Random Search* supera o *Grid Search* ao realizar o ajuste de hiperparâmetros em termos de utilização de recursos computacionais e tempo de execução (LIASHCHYNSKYI; LIASHCHYNSKYI, 2019).

A validação cruzada é um procedimento de reamostragem usado para avaliar modelos de aprendizado de máquina em uma amostra de dados limitada. A validação cruzada *K-fold* envolve a divisão aleatória do conjunto de observações em "k"grupos, ou dobras, de tamanho aproximadamente igual. A primeira dobra é tratada como um conjunto de validação e o método se ajusta nas dobras $k-1$ restantes. Por exemplo, usando o erro quadrático médio como função de pontuação, MSE_1, é então calculado nas observações na dobra retida. Este procedimento é repetido k vezes; cada vez, um grupo diferente de observações é tratado como um conjunto de validação. Este processo resulta em k estimativas do erro de teste, MSE_1, MSE_2, ..., MSE_k. A estimativa da validação cruzada K-Fold é calculada pela média desses valores (JAMES *et al.*, 2013).

3.3.1.4 Avaliar os modelos preditivos

Nessa etapa, o cientista de dados deve avaliar o desempenho dos modelos preditivos gerados. Para comparar o desempenho de modelos diferentes, é indicada a utilização de métricas já consolidadas. Para problemas de classificação sugere-se a utilização das seguintes métricas:

– Matriz de Confusão(*Confusion Matrix*): Esta matriz indica quantos exemplos existem em cada grupo: falso positivo(Falso Positivo (FP)), falso negativo(Falso Negativo (FN)), verdadeiro positivo(Verdadeiro Positivo (VP)) e verdadeiro negativo(Verdadeiro Negativo (VN)). É interessante visualizar a contagem destes grupos tanto em números absolutos quanto em porcentagens da classe real

$VerdadeiroPositivo(TP)$	$FalsoNegativo(FN)$
$FalsoPositivo(FP)$	$VerdadeiroNegativo(TN)$

– Curva Roc(*Roc Curve*): A curva AUC - ROC é uma medida de desempenho para os problemas de classificação em várias configurações de limite. ROC é uma curva de probabilidade e AUC(Área abaixa da Curva) representa o grau ou medida de separabilidade. Diz o quanto o modelo é capaz de distinguir entre as classes. Quanto maior a AUC, melhor será o modelo em prever 0 classes como 0 e 1 classe como 1. Por analogia, quanto maior a AUC, melhor será o modelo em distinguir entre pacientes com a doença e sem doença.

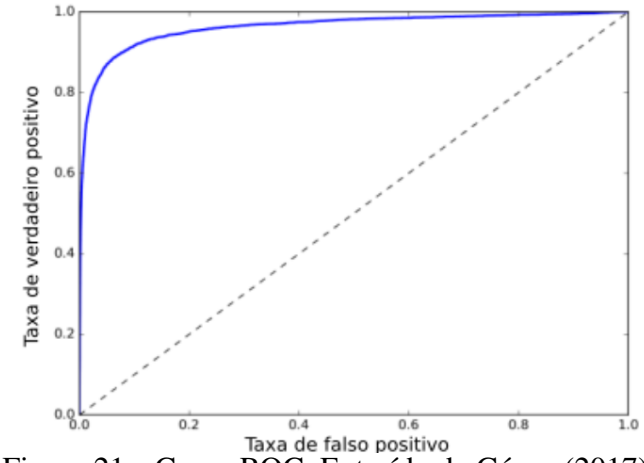


Figura 21 – Curva ROC. Extraído de Géron (2017).

– Acurácia(*Accuracy Score*): exibe quantos de nossos exemplos foram de fato classificados corretamente, independente da classe. Por exemplo, se temos 100 observações e 90 delas foram classificados corretamente, nosso modelo possui uma acurácia de 90%. A acurácia

é definida pela fórmula abaixo:

$$accuracy = \frac{VP+VN}{VP+VN+FN+FP} \tag{3.11}$$

– Precisão(*Precision Score*): é definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e o total de exemplos classificados como positivos, conforme a fórmula abaixo:

$$precision = \frac{VP}{VP+FP} \tag{3.12}$$

– Revocação(*Recall Score*): é definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e a quantidade de exemplos que são de fato positivos, conforme a fórmula abaixo:

$$recall = \frac{VP}{VP+FN} \tag{3.13}$$

Já para problemas de regressão o usuário deve utilizar as seguintes métricas:

– Erro Quadrático Médio(MSE): o Erro Quadrático Médio consiste na média do erro das previsões ao quadrado. Essa métrica apresenta valor mínimo 0, sem valor máximo e quanto maior esse número, pior o modelo. Abaixo segue a fórmula:

$$MSE = \frac{1}{n}\sum n = 1n(\hat{y}_i - y_i)^2 \tag{3.14}$$

– R-Quadrado: o R^2 ou Coeficiente de Determinação é uma métrica que visa expressar a quantidade da variança dos dados. O valor do seu R-Quadrado varia de 0 a 1 e geralmente é representado em porcentagem. A seguir temos como é calculado o R2:

$$R^2 = 1 - \frac{\sum_{n=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{n=1}^{n}(y_i - \bar{y}_i)^2} \tag{3.15}$$

– Raiz do erro quadrático médio(RMSE): é a raiz quadrada da média dos erros quadrados. O efeito de cada um dos erros no RMSE é proporcional ao tamanho do quadrado do erro; portanto, erros maiores têm um efeito desproporcional no RMSE tornando-o sensível a outliers ou anomalias. O RMSE pode ser definido como:

$$RMSE = \sqrt{\frac{1}{n}\sum n = 1n(\hat{y}_i - y_i)^2} \tag{3.16}$$

– Erro Absoluto Médio(Erro Absoluto Médio (MAE)): consiste na média das distâncias entre valores preditos e reais. Diferentemente do MSE e do RMSE, essa métrica não

"pune" tão severamente os outliers do modelo. Essa medida apresenta valor mínimo 0 e não apresenta valor máximo, abaixo segue a fóruma do MAE:

$$MAE = \frac{1}{n}\sum n = 1n|\hat{y}_i - y_i| \tag{3.17}$$

– Erro Percentual Absoluto Médio(Erro Percentual Absoluto Médio (MAPE)): essa medida exprime uma porcentagem, obtida através da divisão da diferença entre predito (ŷ) e real pelo valor real (y). Assim como o MSE e o MAE, quanto menor o valor, mais preciso seria o modelo de regressão. Abaixo segue a fórmula:

$$MAE = \frac{1}{n}\sum n = 1n|\frac{|\hat{y}_i - y_i|}{|y_i|} \tag{3.18}$$

3.3.2 Apresentar os resultados obtidos

Nesta etapa, o cientista deve organizar e apresentar os resultados obtidos, ou seja, os algoritmos, as características, as técnicas de pré-processamento aplicadas e os hiperparâmetros dos modelos que apresentaram os melhores desempenhos. O objetivo principal desta etapa é auxiliar o cientista a encontrar o modelo mais adequado para o problema investigado. Neste sentido, recomenda-se a produção de um artefato (um relatório, por exemplo) contendo todos os resultados alcançados, um resumo das análises realizadas pelo cientista, bem como as conclusões obtidas.

3.3.3 Assegurar a reprodutibilidade

Na última etapa da terceira fase do guia proposto, o cientista deve assegurar a reprodutibilidade das atividades executadas com a finalidade de fornecer credibilidade ao estudo realizado e possibilitar que este seja executado por outros usuários. Em (OLORISADE *et al.*, 2017) destaca-se que muitos trabalhos não conseguem assegurar a reprodução de seus experimentos. Com base nessa dificuldade recomenda-se a utilização de estratégias, ferramentas e artefatos que possam facilitar a reprodução dos experimentos realizados. Neste sentido, após a execução de um determinado conjunto de experimentos deve-se armazenar todas as decisões tomadas pelo cientista de dados.

4 DSADVISOR: UMA FERRAMENTA PATA APOIAR TAREFAS PREDITIVAS

Seguindo o guia prático apresentado no capítulo anterior, construímos uma ferramenta denominada *DSAdvisor*, a qual tem por finalidade direcionar o usuário na execução das principais atividades que compõem a solução de um problema preditivo, bem como orientar a interpretação dos resultados obtidos. A *DSAdvisor* foi desenvolvida em *Python* (ROSSUM, 1995) utilizando bibliotecas como *Flask* (GRINBERG, 2018), *scikit-learn*, *seaborn*, *matplotlib*, *seaborn*, *scipy*, *numpy*, *pandas* dentre outras. A Figura 22 ilustra a tela inicial da *DSAdvisor*.

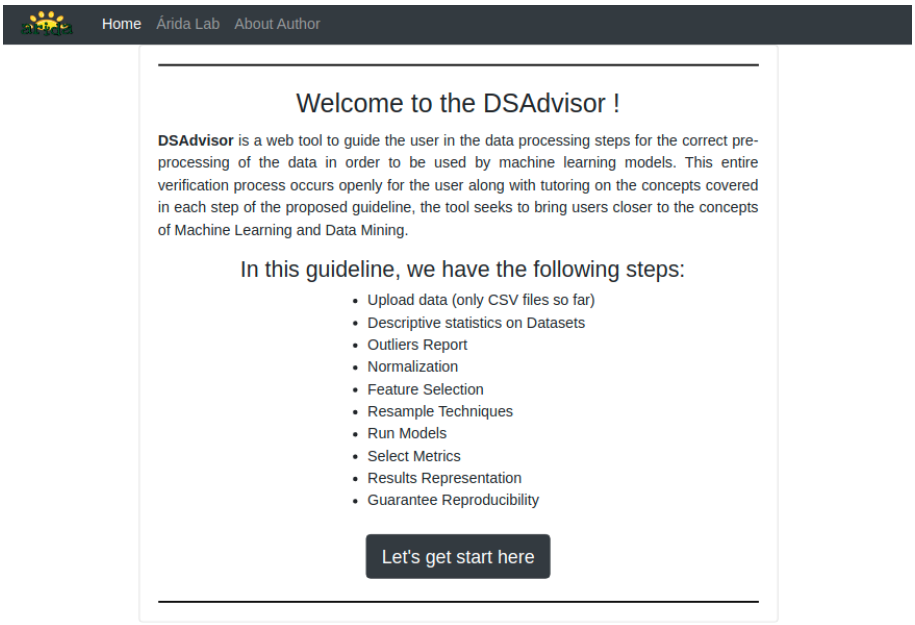


Figura 22 – Tela inicial da ferramenta "DSAdvisor. Fonte: Autor.

A *DSAdvisor* busca encorajar usuários não especialistas a construir modelos de aprendizado de máquina para solucionar tarefas preditivas (regressão ou classificação). A seguir, descreveremos em detalhes a implementação da *DSAdvisor*. Por questões didáticas, iremos seguir as mesmas fases, etapas e atividades presentes no guia prático proposto no capítulo anterior.

4.1 Fase 1 - Análise Exploratória

Como mencionado anteriormente, a primeira fase do guia proposto, denominada “Análise Exploratória”, tem como principal objetivo explorar os dados que serão utilizados na construção de um ou mais modelos preditivos. Portanto, nesta fase, busca-se entender, descrever e resumir os dados que serão utilizados.

4.1.1 Carregar os dados

Inicialmente, o usuário precisa carregar os dados a serem explorados. Para isso, ele deve indicar um arquivo no formato csv, além de uma breve descrição deste conjunto de dados, como ilustrado na Figura 23. Em seguida (após o *upload* do arquivo), a *DSAdvisor* exibe uma amostra dos dados contendo 10 linhas, como mostra a Figura 24. Por questões didáticas, usaremos um mesmo conjunto de dados durante todo esse capítulo. Esse conjunto de dados foi criado com a finalidade de possibilitar a exemplificação de todas as funcionalidades da ferramenta *DSAdvisor* e pode ser obtido livremente em nosso repositório no *GitLab*[1].

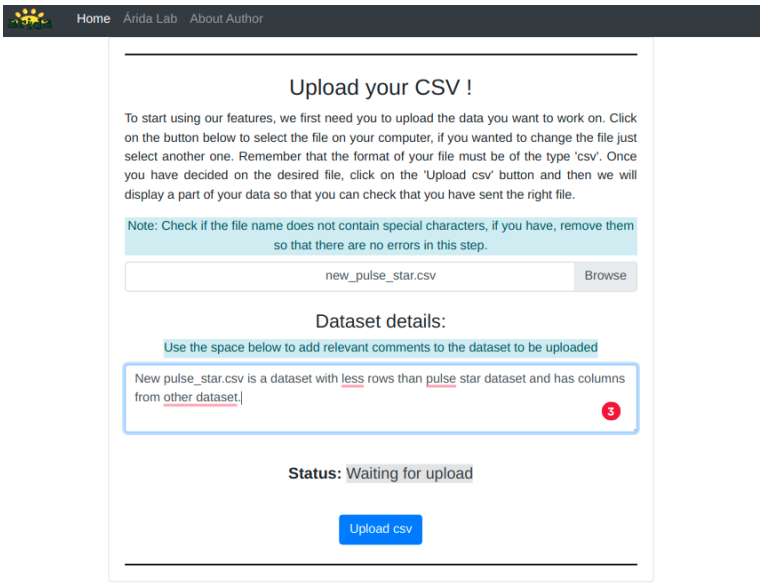


Figura 23 – Tela de upload de arquivo csv. Fonte: Autor.

[1] O data set utilizado neste capítulo pode ser obtido por meio do link https://gitlab.com/jmmonteiro/dsadvisor

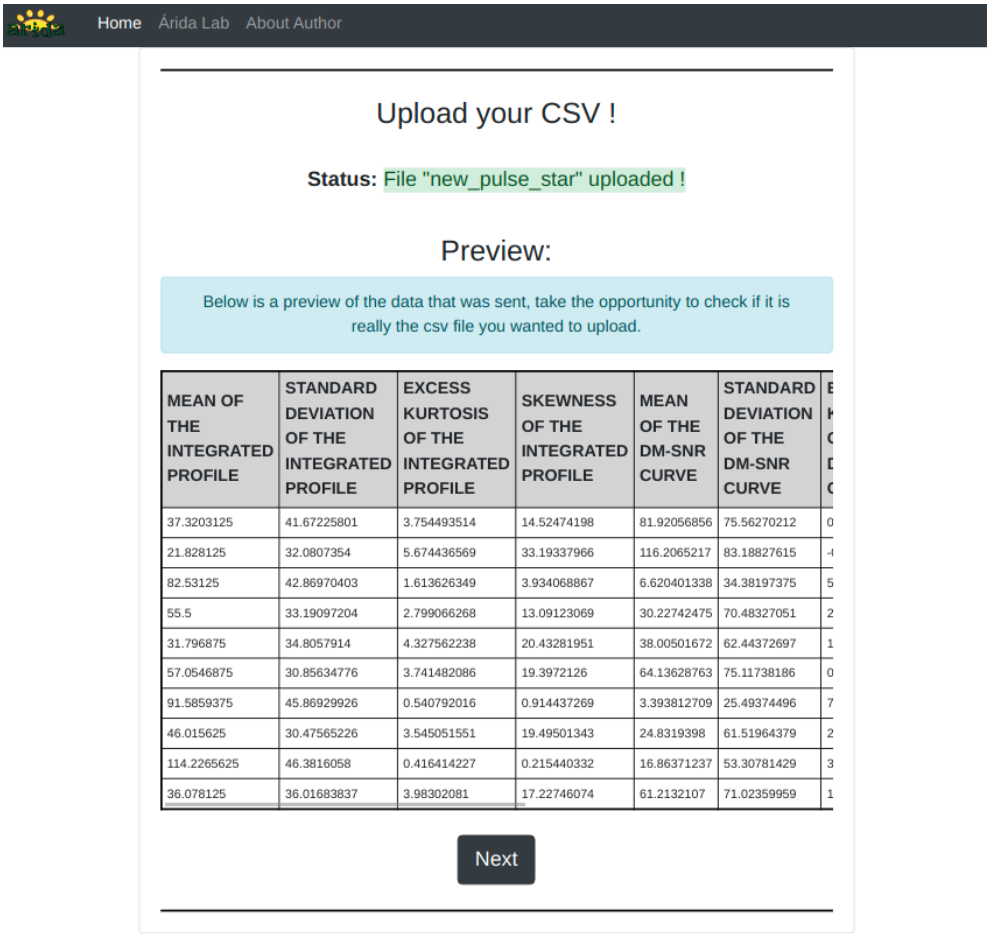


Figura 24 – Tela de confirmação de envio de arquivo. Fonte: Autor.

4.1.2 Checar o tipo de cada variável

Para cada variável (coluna) do conjunto de dados fornecido anteriormente, a ferramenta *DSAdvisor* irá identificar e mostrar o seu tipo computacional, seguindo o Algoritmo 1. A página *"Resume Variables"* apresenta uma breve explicação sobre os tipos de variáveis categóricas e numéricas (discretas ou contínuas) com a finalidade de orientar o usuário, como ilustra a Figura 25.

Algoritmo 1: Checagem do tipo de variáveis. Fonte: Autor.

```
início
    for cada variável V no dataset do
        Checar o tipo da variável V
        if V é do tipo "string" then
            Rotular V como sendo do tipo categórica
        end
        if V é do tipo "float" then
            Rotular V como sendo do tipo numérica contínua
        end
        if V é do tipo "int" then
            Rotular V como sendo do tipo numérica discreta
        end
    end
fim
```

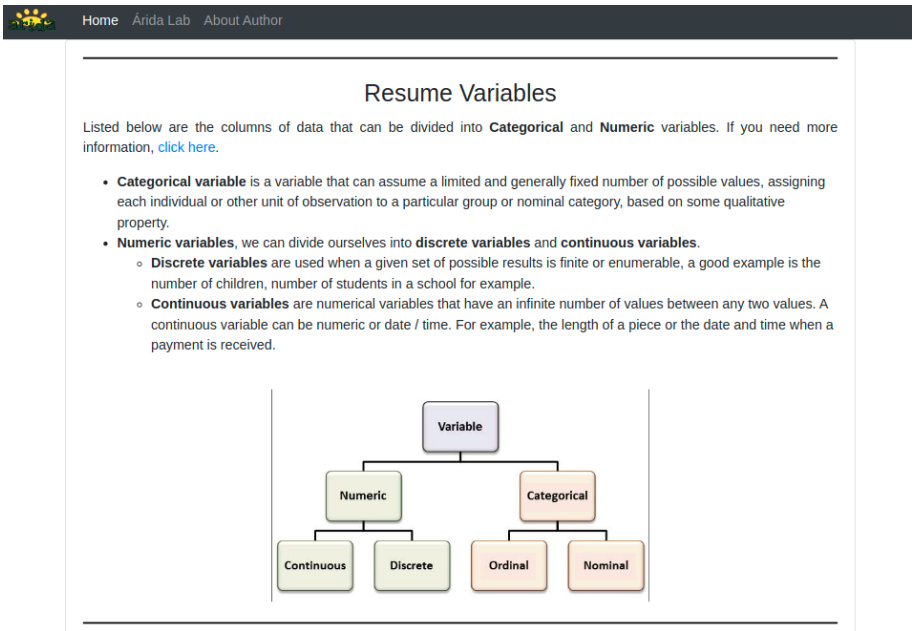


Figura 25 – Tela de resumo de variáveis. Fonte: Autor.

4.1.3 Filtrar variáveis

Nesta etapa, todas as variáveis presentes no conjunto de dados serão exibidas. Em seguida, caso deseje, o usuário pode selecionar algumas variáveis para serem removidas do conjunto de dados. Desta forma, o usuário consegue definir um subconjunto contendo as variáveis com as quais deseja continuar trabalhando. O Algoritmo 2 descreve os passos realizados pela *DSAdvisor* com a finalidade de auxiliar o usuário a filtrar variáveis. A Figura 26 ilustra a tela da *DSAdvisor* onde o usuário pode indicar as variáveis que serão removidas do conjunto de dados. Após a realização dessa tarefa, a *DSAdvisor* gera-se um novo conjunto de dados contendo apenas as variáveis que o usuário deseja utilizar, descartando-se as demais. Esse novo conjunto de dados será utilizado pela *DSAdvisor* nas tarefas seguintes.

Algoritmo 2: Filtragem de variáveis. Fonte: Autor.

início
Exibir variáveis agrupadas por seus respectivos tipos
O usuário seleciona quais variáveis remover
As variáveis selecionadas são então removidas
fim

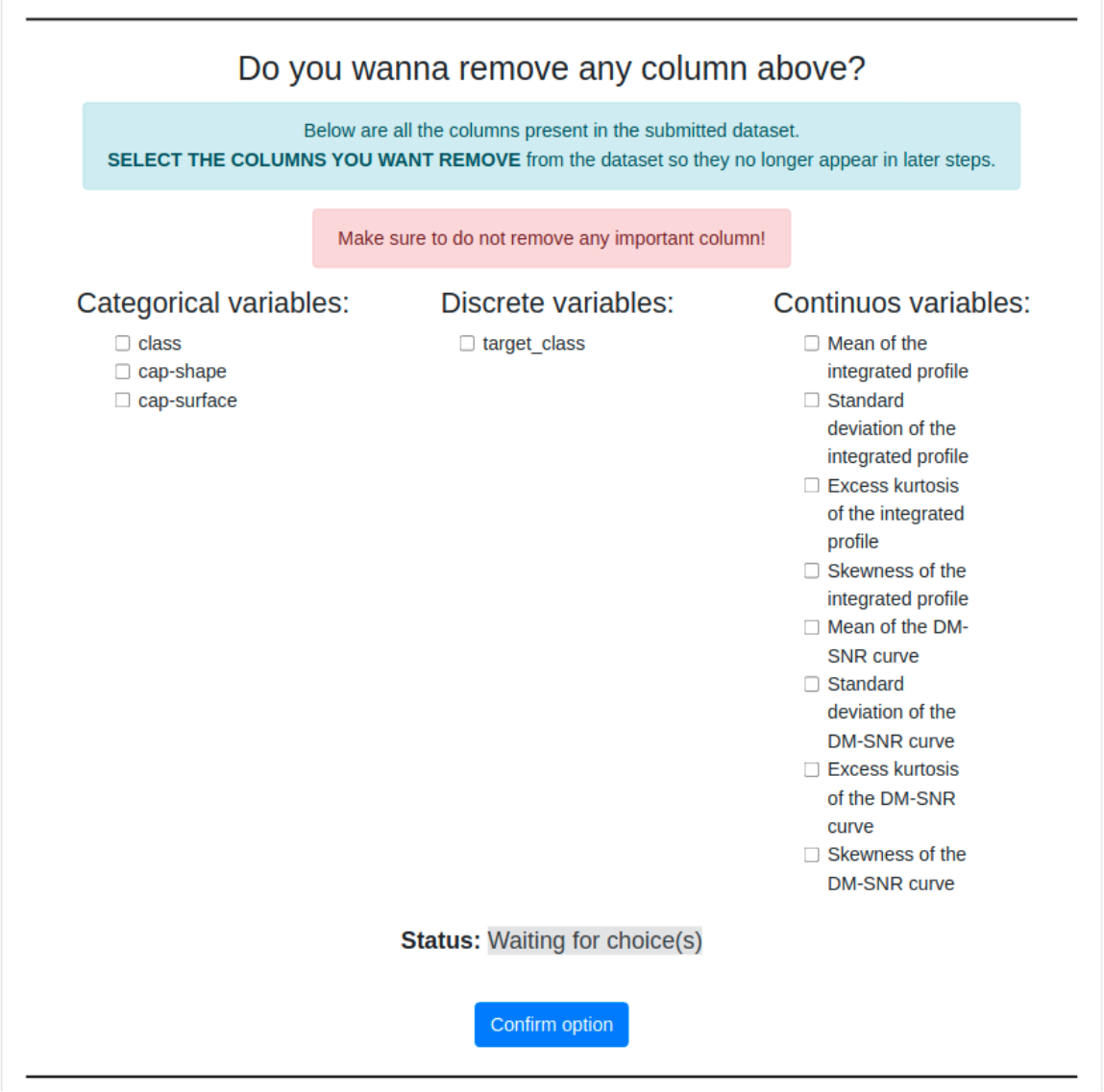


Figura 26 – Tela de remoção de variáveis. Fonte: Autor.

4.1.4 Definir os códigos para os valores faltantes

Diferentes códigos podem ser utilizados para representar a ausência de dados. Muitos conjuntos de dados disponíveis publicamente utilizam códigos distintos para representar valores faltantes. A ferramenta *DSAdvisor* permite que o usuário selecione até três representações diferentes para indicar valores faltantes, são elas: *"Nan and None"*, "Empty String" e "Outro". Nesta última opção, o usuário deve indicar o código (ou caractere) que será utilizado para representar valores ausentes. A Figura 27 mostra a tela da ferramenta *DSAdvisor* na qual o usuário indica os códigos que serão utilizados para representar valores faltantes. O Algoritmo 3 descreve o funcionamento desta tarefa.

Algoritmo 3: Listar códigos para representação das variáveis faltantes. Fonte: Autor.

início

Exibir as representações de valores faltantes ao usuário:

Nan/None-EmptyString-Other code(Outros códigos)

O usuário seleciona quais códigos utilizar

fim

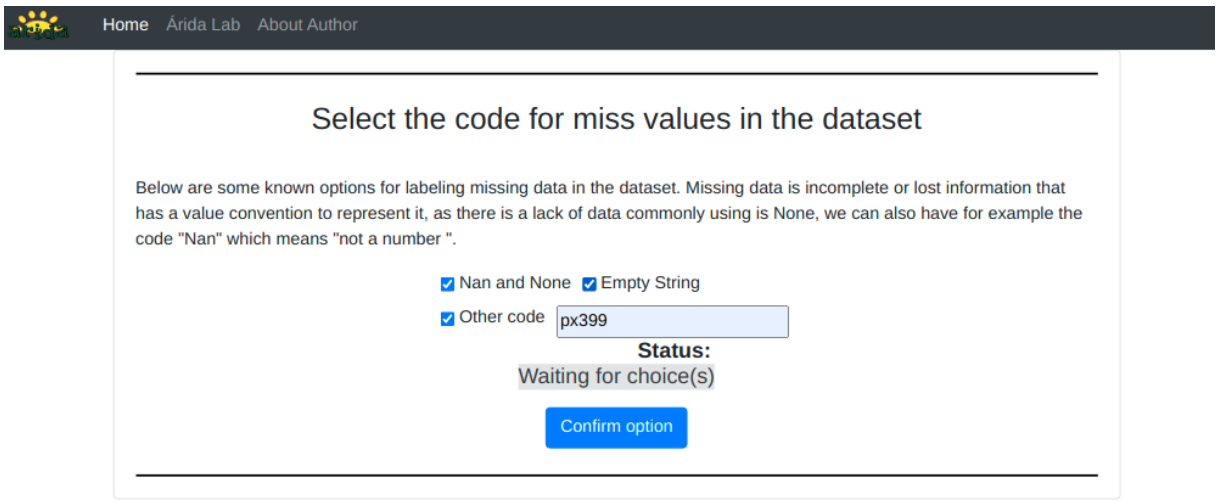


Figura 27 – Tela para escolha de códigos para valores faltantes. Fonte: Autor.

4.1.5 Checar valores faltantes

Após o usuário indicar os códigos que serão utilizados para representar dados faltantes, a ferramenta *DSAdvisor* irá percorrer o conjunto de dados e computar a porcentagem e a quantidade de linhas com valores ausentes para cada um dos códigos anteriormente definidos. O Algoritmo 4 ilustra a execução desta tarefa.

Algoritmo 4: Contar os valores faltantes pelos códigos selecionados. Fonte: Autor.

```
início
    for cada variável V no dataset do
        vector-Nan/None[V]←0 /*Contador para valores Nan/None*/
        vector-EmptyString[V]←0 /*Contador para valores EmptyString*/
        vector-Othercodes[V]←0 /*Contador para valores Outros códigos*/
    end
    for cada variável V no dataset do
        for cada linha L no dataset do
            if V[L] == Code1 then
                vector-Nan/None[V]←vector-Nan/None[V] + 1
            end
            if V[L] == Code2 then
                vector-EmptyString[V]←vector-EmptyString[V] + 1
            end
            if V[L] == Code3 then
                vector-Othercodes[V]←vector-Othercodes[V] + 1
            end
        end
    end
fim
```

4.1.6 Mostrar relatório de valores faltantes

Para cada código definido para representar dados faltantes, a ferramenta *DSAdvisor* exibe uma tabela contendo a porcentagem e a quantidade de instâncias (linhas) do conjunto de dados utilizado contendo valores faltante entretanto a ferramenta não realiza nenhuma imputação ou remoção de dados, cabendo ao usuário fazer separadamente da forma que preferir. A Figura 28 ilustra a tela da *DSAdvisor* que exibe o relatório da checagem de dados faltantes.

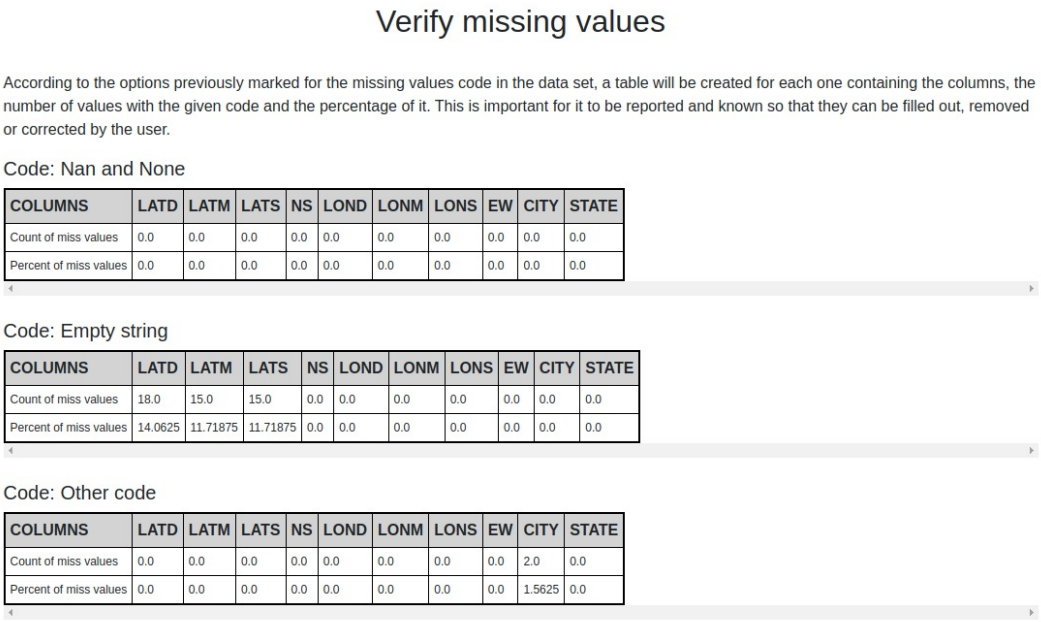
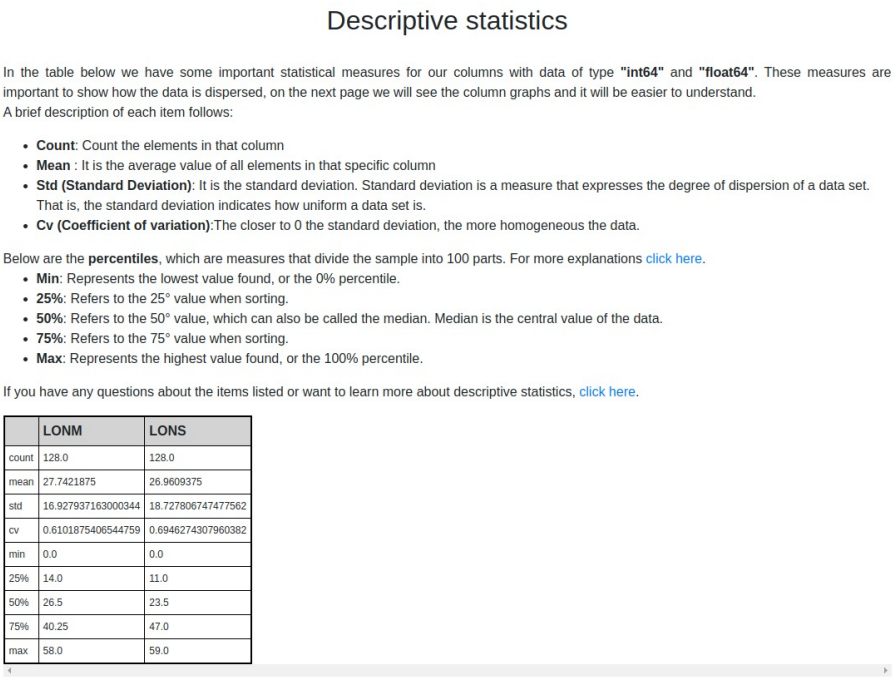


Figura 28 – Tela de resumo de verificações de dados faltantes. Fonte: Autor.

4.1.7 Exibir estatísticas descritivas e os tipos computacionais das variáveis

Nessa etapa a ferramenta *DSAdvisor* irá classificar cada variável do conjunto de dados em dois tipos principais: variáveis numéricas e variáveis categóricas. Para as variáveis numéricas serão exibidas as seguintes medidas estatísticas: número de linhas, média, desvio padrão, coeficiente de variação, o menor valor, o mair valor e os percentis (25%, 50%, 75%) conforme ilustra a Figura 29. Já para as variáveis categóricas a *DSAdvisor* irá mostrar o número de linhas, o número de valores distintos, o elemento mais frequente e a frequência do valor mais comum, conforme ilustra a Figura 30. Adicionalmente, a *DSAdvisor* exibe também uma tabela contendo o tipo computacional de cada variável, conforme mostra a Figura 31. O Algoritmo 5 descreve como a ferramenta *DSAdvisor* implementa essa etapa do guia proposto.



Descriptive statistics

In the table below we have some important statistical measures for our columns with data of type **"int64"** and **"float64"**. These measures are important to show how the data is dispersed, on the next page we will see the column graphs and it will be easier to understand.
A brief description of each item follows:

- **Count**: Count the elements in that column
- **Mean** : It is the average value of all elements in that specific column
- **Std (Standard Deviation)**: It is the standard deviation. Standard deviation is a measure that expresses the degree of dispersion of a data set. That is, the standard deviation indicates how uniform a data set is.
- **Cv (Coefficient of variation)**:The closer to 0 the standard deviation, the more homogeneous the data.

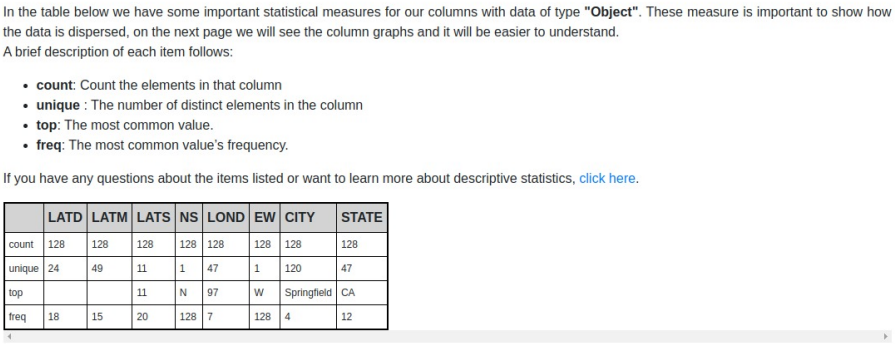
Below are the **percentiles**, which are measures that divide the sample into 100 parts. For more explanations click here.

- **Min**: Represents the lowest value found, or the 0% percentile.
- **25%**: Refers to the 25° value when sorting.
- **50%**: Refers to the 50° value, which can also be called the median. Median is the central value of the data.
- **75%**: Refers to the 75° value when sorting.
- **Max**: Represents the highest value found, or the 100% percentile.

If you have any questions about the items listed or want to learn more about descriptive statistics, click here.

	LONM	LONS
count	128.0	128.0
mean	27.7421875	26.9609375
std	16.927937163000344	18.727806747477562
cv	0.6101875406544759	0.6946274307960382
min	0.0	0.0
25%	14.0	11.0
50%	26.5	23.5
75%	40.25	47.0
max	58.0	59.0

Figura 29 – Tela de estatísticas descritivas para as variáveis numéricas. Fonte: Autor.



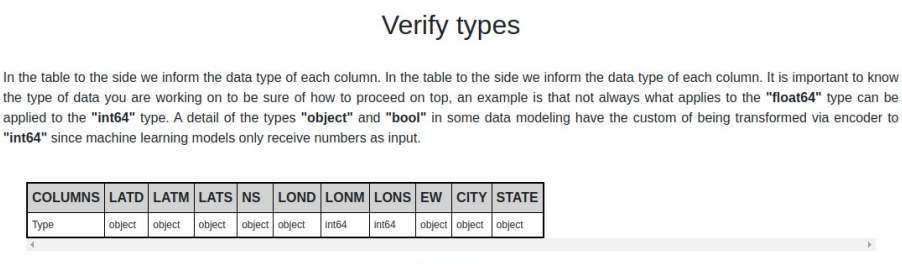
In the table below we have some important statistical measures for our columns with data of type **"Object"**. These measure is important to show how the data is dispersed, on the next page we will see the column graphs and it will be easier to understand.
A brief description of each item follows:

- **count**: Count the elements in that column
- **unique** : The number of distinct elements in the column
- **top**: The most common value.
- **freq**: The most common value's frequency.

If you have any questions about the items listed or want to learn more about descriptive statistics, click here.

	LATD	LATM	LATS	NS	LOND	EW	CITY	STATE
count	128	128	128	128	128	128	128	128
unique	24	49	11	1	47	1	120	47
top			11	N	97	W	Springfield	CA
freq	18	15	20	128	7	128	4	12

Figura 30 – Tela de estatísticas descritivas para as variáveis categóricas. Fonte: Autor.



Verify types

In the table to the side we inform the data type of each column. In the table to the side we inform the data type of each column. It is important to know the type of data you are working on to be sure of how to proceed on top, an example is that not always what applies to the **"float64"** type can be applied to the **"int64"** type. A detail of the types **"object"** and **"bool"** in some data modeling have the custom of being transformed via encoder to **"int64"** since machine learning models only receive numbers as input.

COLUMNS	LATD	LATM	LATS	NS	LOND	LONM	LONS	EW	CITY	STATE
Type	object	object	object	object	object	int64	int64	object	object	object

Figura 31 – Tela de listagem de tipos computacionais. Fonte: Autor.

Algoritmo 5: Exibir estatísticas das variáveis, resumo de dados faltantes e tipos computacionais. Fonte: Autor.

```
início
    for cada variável V no dataset do
        if V é do tipo numérico then
            Exibir medidas estatísticas descritivas para variáveis numéricas.(cont, mean,
             std, cv, min, 15%, 50%, 75%, max)
        end
        if V é do tipo categórica then
            Exibe medidas estatísticas descritivas para variáveis categóricas. (Cont,
             unique, top, freq)
        end
    end
    for cada código de valor ausente MVC do
        Mostra a quantidade de cada MVC Exibir a porcentagem de cada MVC
    end
    Exibir total de dados ausentes para todos os códigos MVC
    Mostra o tipo computacional de cada variável V
fim
```

4.1.8 Exibir histogramas para cada variável numérica discreta e a proporção de valores para cada variável categórica

Para cada variável categórica, a ferramenta *DSAdvisor* mostra um gráfico de pizza, se o número de categorias das variáveis for menor que quatro, ou um gráfico de barras, caso contrário, conforme ilustra a Figura 32. Adicionalmente, para cada variável numérica discreta, a *DSAdvisor* constrói e exibe um histograma assim, conforme mostra a Figura 33. O Algoritmo 6 descreve como a ferramenta *DSAdvisor* implementa essa etapa do guia proposto.

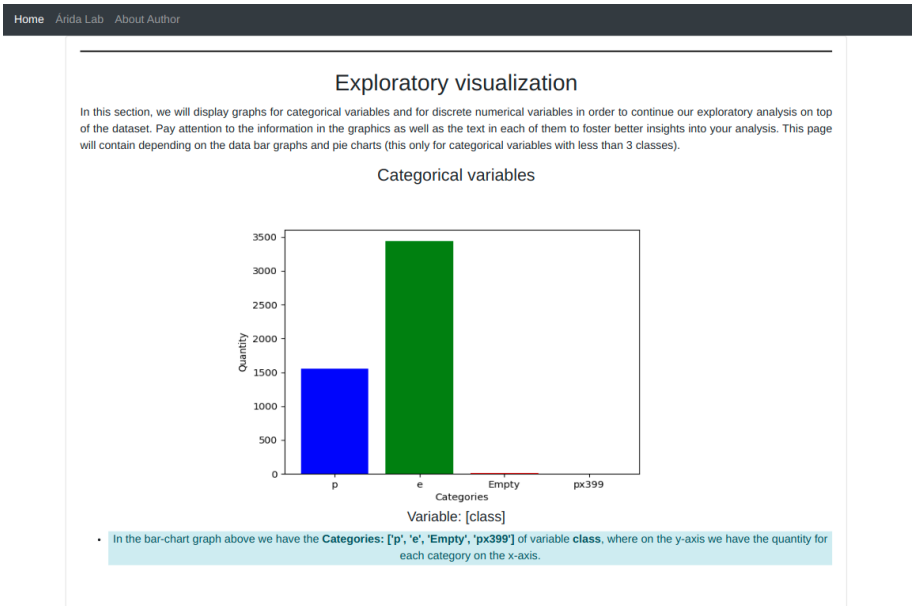


Figura 32 – Tela de exibição de variáveis categóricas. Fonte: Autor.

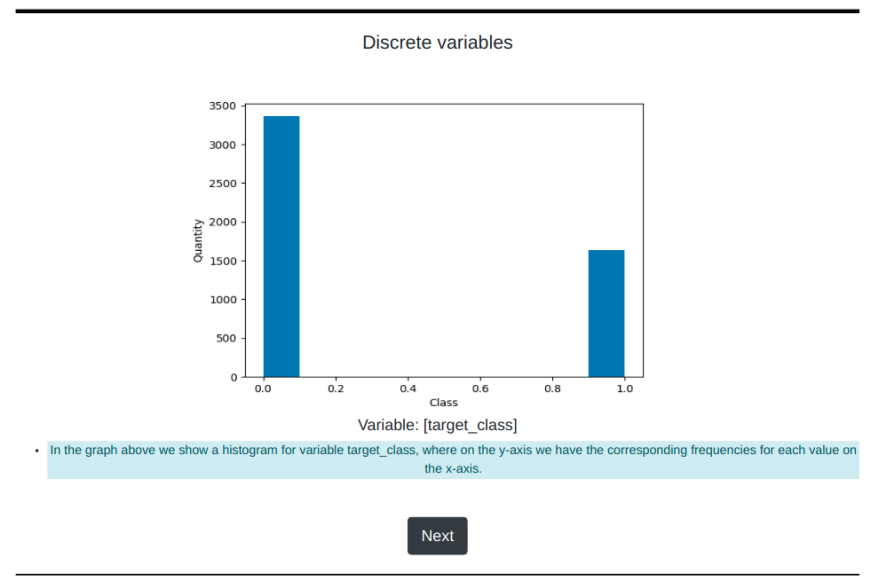


Figura 33 – Tela de exibição de variáveis numéricas discretas. Fonte: Autor.

Algoritmo 6: Exibir variáveis categóricas e numéricas discretas. Fonte: Autor.

```
início
    for cada variável V no dataset do
        if V é do tipo numérico discreta then
            Exibir histograma para cada variável discreta V. /*Exibir distribuição de
             dados*/
        end
        if V é do tipo categórica then
            if V possui mais de 3 categorias then
                Exibir gráfico de barra para cada variável categórica de V.
            end
            if V possui menos de 4 categorias then
                Exibir gráfico de pizza para cada variável categórica V. /*ver a proporção
                 de cada categoria da variável V*/
            end
        end
    end
fim
```

4.1.9 Exibir a distribuição do "bestfit" para cada variável contínua

Inicialmente, o usuário deve selecionar um conjunto de distribuições de probabilidade. A Figura 34 mostra as distribuições disponíveis na ferramenta *DSAdvisor*.

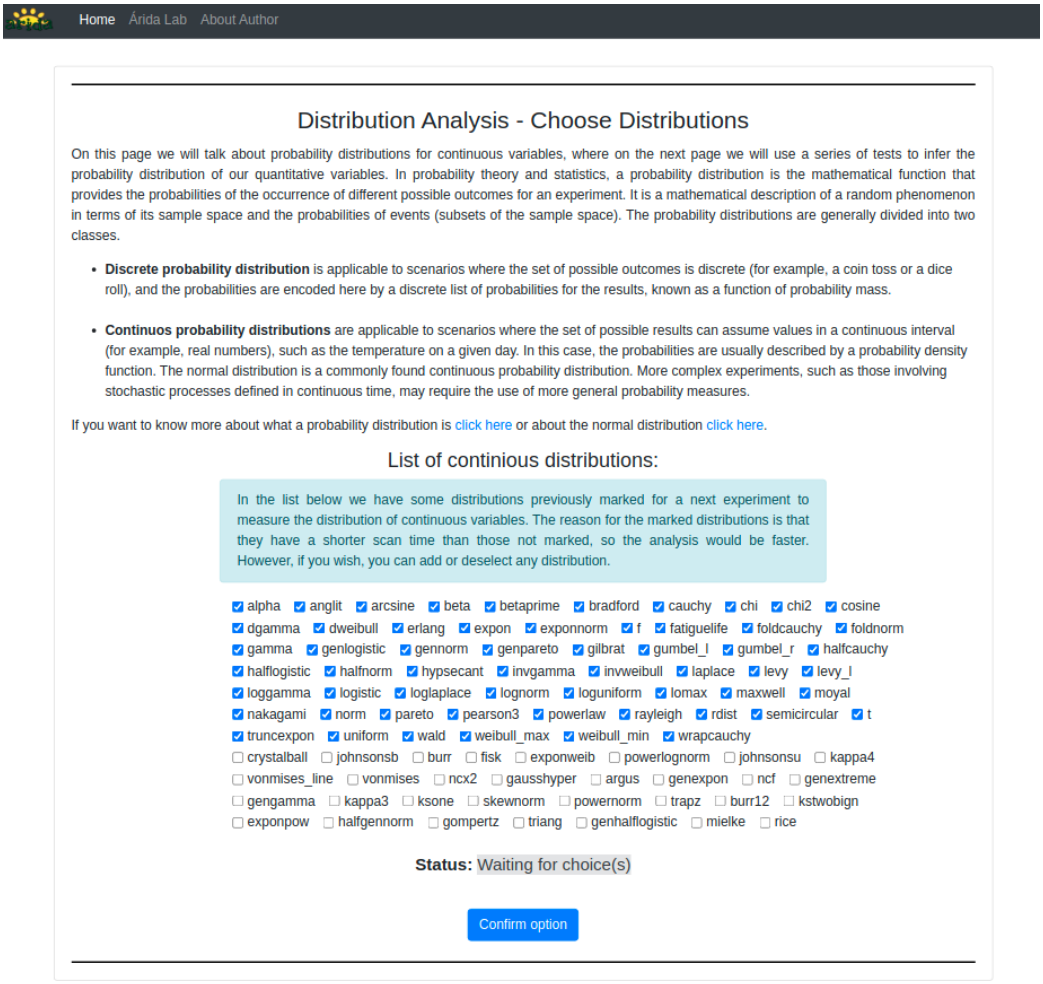


Figura 34 – Distribuições estatísticas disponíveis na DSAdvisor. Fonte: Autor.

Em seguida, a *DSAdvisor* aplica, para cada variável numérica contínua, os seguintes testes: K-quadrado de D'Agostino (D'AGOSTINO, 1970), Lilliefors (LILLIEFORS, 1967), Shapiro-Wilk(SHAPIRO; WILK, 1965) e Kolmogorov-Smirnov (SMIRNOV, 1948). O teste de Kolmogorov-Smirnov é utilizado para avaliar qual das distribuições selecionadas pelo usuário é a que mais se aproxima da distribuição da variável. Os demais testes são utilizados para avaliar a normalidade dos dados, ou seja, se a uma determinada variável segue uma distribuição normal ou não, como ilustra a Figura 35.

Para cada variável numérica contínua, a *DSAdvisor* irá exibir o resultado de cada um dos testes de normalidade, além do seu histograma, a curva referente à distribuição normal e a curva da distribuição que obteve a maior aproximação segundo o teste de Kolmogorov-Smirnov. Em seguida, para cada variável, o usuário tem a opção de definir qual das distribuições parece se adequar melhor, caso a distribuição normal ou a distribuição retornada pelo "Bestfit" não seja considerada adequada, conforme ilustra a Figura 36.

O Algoritmo 7 descreve como a *DSAdvisor* implementa a tarefa de *"bestfit"*. Vale

destacar que a ferramenta *DSAdvisor* irá avaliar somente as distribuições previamente selecionadas pelo usuário.

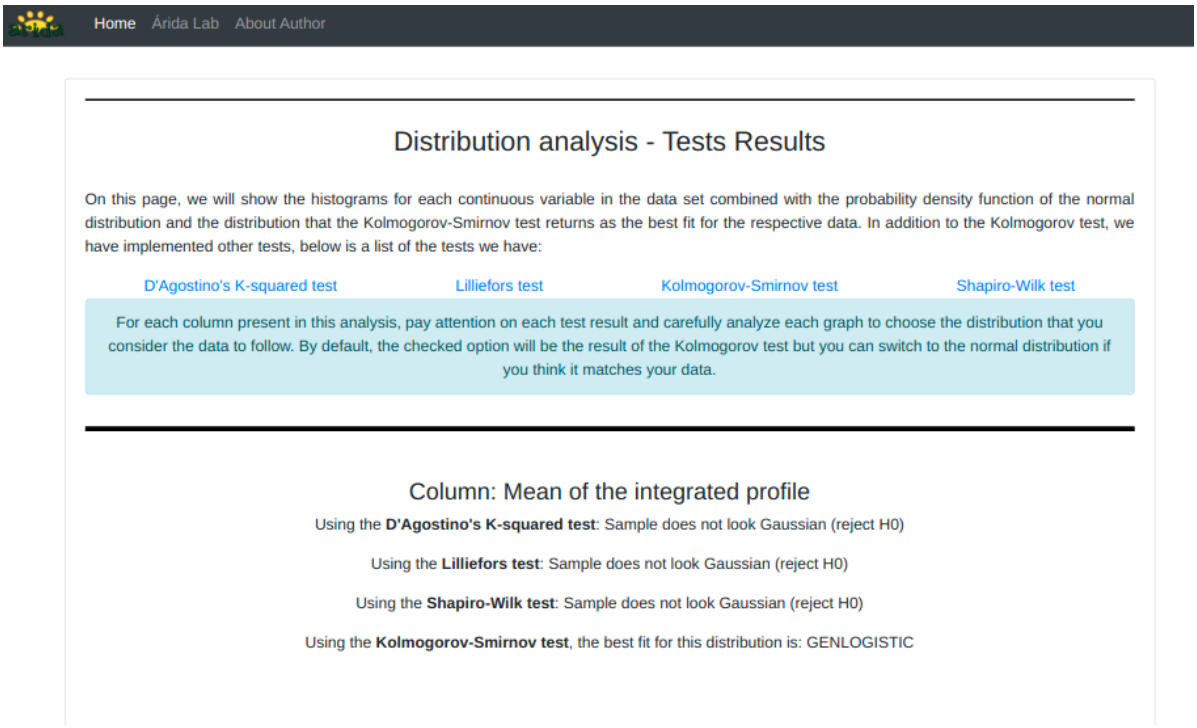


Figura 35 – Análise de normalidade dos métodos K-quadrado de D'Agostino, Lilliefors, Shapiro-Wilk e o resultado do teste de Kolmogorov-Smirnov. Fonte: Autor.

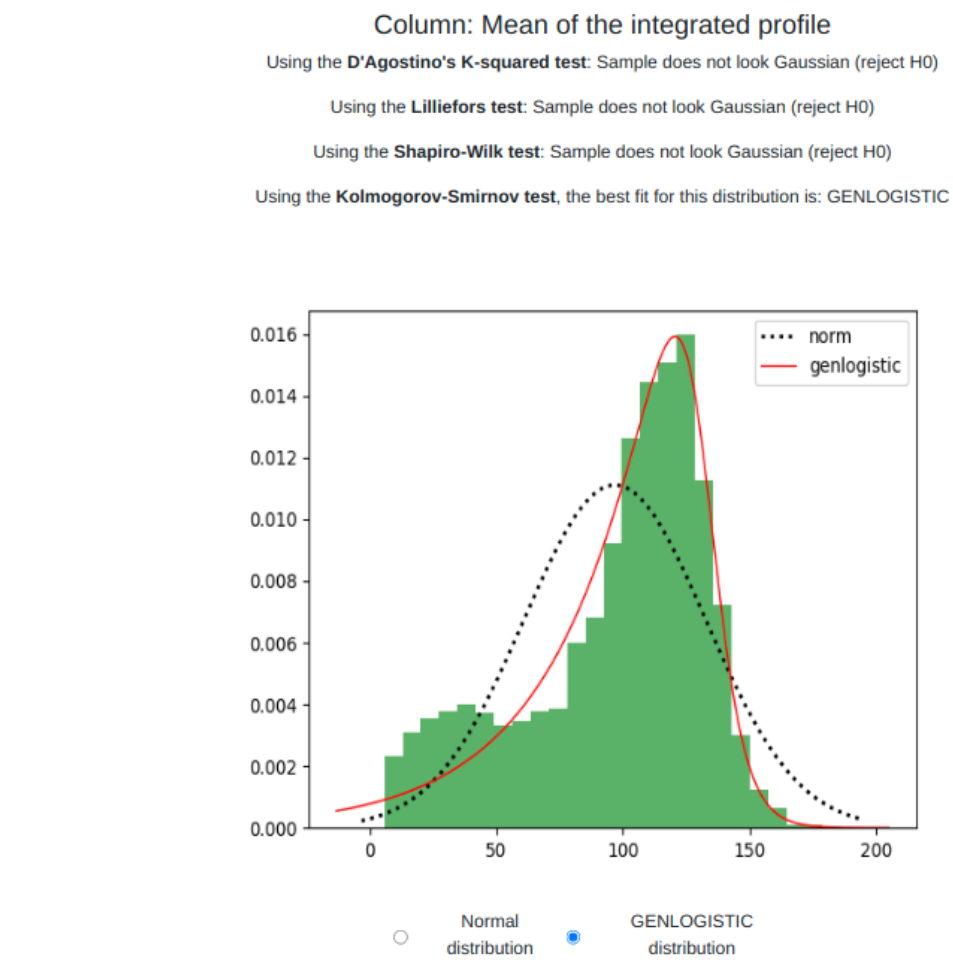


Figura 36 – Exemplo de análise da distribuição de cada variável com o método "Bestfit". Fonte: Autor.

Algoritmo 7: Mostrar distribuição de melhor ajuste aos dados("Best fit"). Fonte: Autor.

```
início
    Exibir lista de distribuições disponíveis para comparação com as variáveis
    O usuário seleciona um conjunto de distribuições(CD)
    for cada variável(V) do
        vector-Bestfitpvalue[V] ← 0
        vector-Bestfitdist ← [] /*Empty List*/
        for cada distribuição(D) pertencente a (CD) do
            Pvalue, Dist = Executa kolmogorovTest(V,D)
            if Pvalue > Bestfitpvalue[V] then
                Bestfitpvalue[V] ← Pvalue
                Bestfitdist[V] ← Dist
            end
        end
        Rodar teste de cramer-von misses(V)
        Rodar teste de d'agostino(V)
        Rodar teste de lilliefors(V)
        Rodar teste de Shapiro-wilk(V)
        Exibir resultado de cada teste
        Exibir histograma
        Mostrar melhor curva da distribuição[V] e a curva normal /*o usuário escolhe
         uma das duas distribuições para associar a cada variável*/
    end
fim
```

4.1.10 Mostrar coeficientes de correlação para cada par de variáveis

Inicialmente, a ferramenta *DSAdvisor* exibe uma série de informações sobre o conceito de correlação, a definição dos principais coeficientes de correlação, além da forma de computar e interpretar os valores desses coeficientes, conforme ilustra a Figura 37. Tudo isso buscando auxiliar o usuário na compreensão dessa importante etapa do guia proposto.

A ferramenta *DSAdvisor* exibe o coeficiente de correlação de acordo com as variáveis presentes no conjunto de dados. Para todos os pares de variáveis numéricas contínuas, a *DSAdvisor* irá computar e exibir o coeficiente de correlação de Spearman (SPEARMAN, 1961). Adicionalmente, para cada par de variáveis que possuem distribuição normal, a *DSAdvisor* calcula e mostra o coeficiente de correlação de Pearson (PEARSON, 1895).

Neste sentido, a *DSAdvisor* constrói e exibe um gráfico personalizado denominado de “gráfico de matriz de correlação”, no qual, abaixo da diagonal temos gráficos de dispersão; na diagonal, a distribuição da variável; e acima da diagonal, os valores das correlações personalizadas com o tamanho da fonte, denotando a relevância da associação entre as variáveis e a quantidade de asteriscos vermelhos, variando de 1 a 3, refletindo um valor p menor ou maior. A Figura 38 ilustra o “gráfico de matriz de correlação” gerado pela *DSAdvisor* para o conjunto de dados “Pulse Star” utilizando o coeficiente de correlação de Spearman.

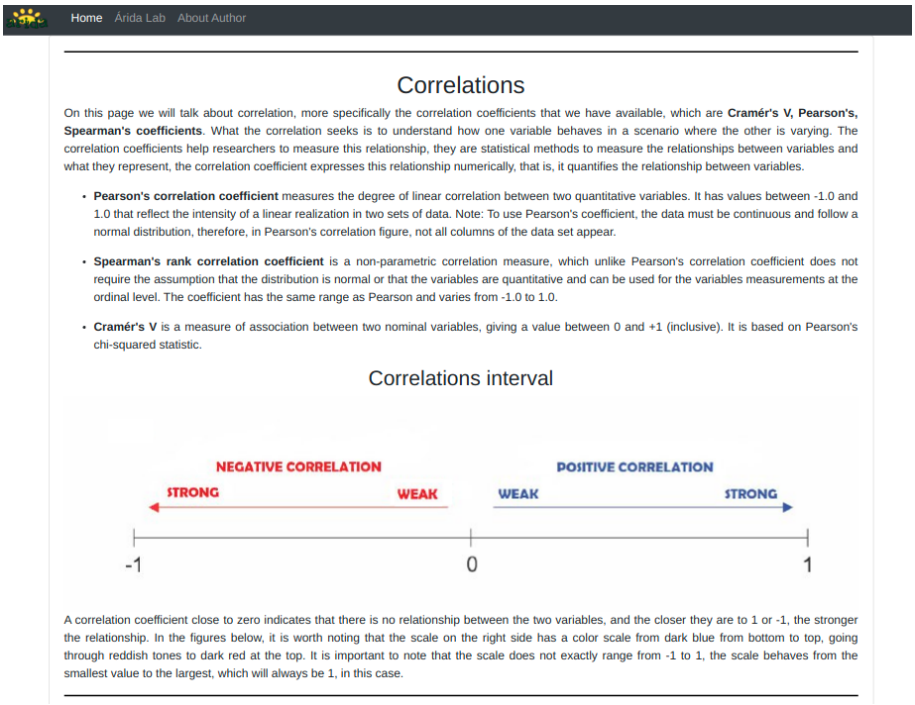


Figura 37 – Informações básicas sobre correlações. Fonte: Autor.

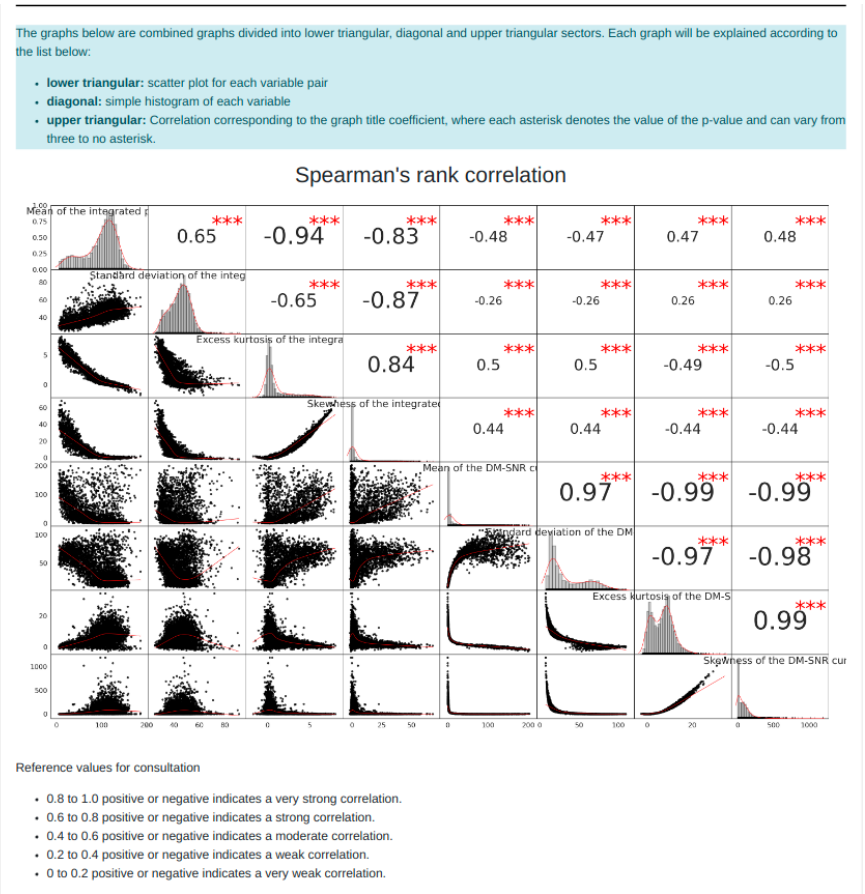


Figura 38 – Matriz de correlação de Spearman referente ao conjunto de dados "Pulse Star".
Fonte: Autor.

4.1.11 Mostrar o valor da medida V de Cramer para cada par de variáveis categóricas

Se o conjunto de dados utilizado pelo usuário contiver variáveis categóricas, a ferramenta *DSAdvisor* exibe o resultado da medida V de Cramer (CRAMÉR, 1999) para cada par de variáveis categóricas, utilizando para isso um um gráfico do tipo “mapa de calor” (*heat map*), conforme ilustra a Figura 39.

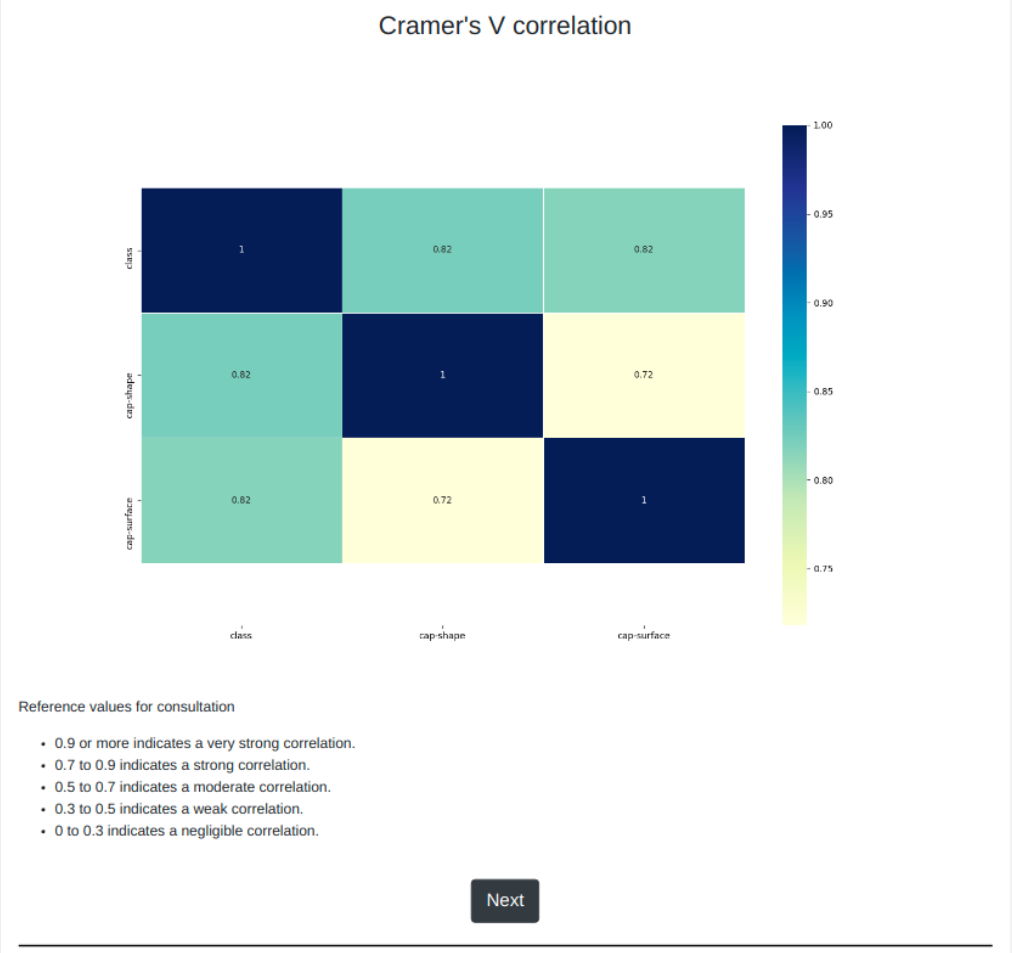


Figura 39 – Mapa de calor com o resultao do teste V de Cramer. Fonte: Autor.

4.2 Fase 2 - Pré-processamento dos dados

O principal objetivo da segunda fase do guia proposto no capítulo anterior, chamada de pré-processamento dos dados, consiste em preparar os dados para que estes possam ser utilizados na construção de modelos preditivos. Seguindo este guia, a ferramenta *DSAdvisor* inclui diferentes funcionalidades relacionadas à detecção de valores discrepantes, normalização de dados, escolha da variável independente, seleção de atributos, balanceamento de dados, seleção de recursos e divisão de conjuntos de treinamento e teste.

4.2.1 Escolher a variável dependente

Inicialmente, o usuário deve escolher dentre todas as colunas do conjunto de dados aquela que será a variável dependente, chamada de "y", enquanto as demais variáveis serão chamadas de "x". A Figura 40 ilustra a tela da ferramenta *DSAdvisor* na qual o usuário define a coluna que será utilizada como variável dependente.

4.2.2 Escolher a porcentagem de divisão para os conjuntos de treinamento e teste

Nesta etapa, o usuário deve definir a proporção dos conjuntos de treinamento e teste. A ferramenta *DSAdvisor* sugere quatro possibilidades distintas:

- 50% para treinamento - 50% para teste
- 60% para treinamento - 40% para teste
- 70% para treinamento - 30% para teste
- 80% para treinamento - 20% para teste

A Figura 40 ilustra a tela da *DSAdvisor* onde o usuário pode selecionar uma das opções para divisão dos dados nos conjuntos de treinamento e teste. Após o usuário confirmar sua escolha, a *DSAdvisor* irá executar a separação dos dados de forma aleatória e mantendo a proporção entre as classes, caso a variável dependente seja categórica (o que em geral ocorre em problemas de classificação).

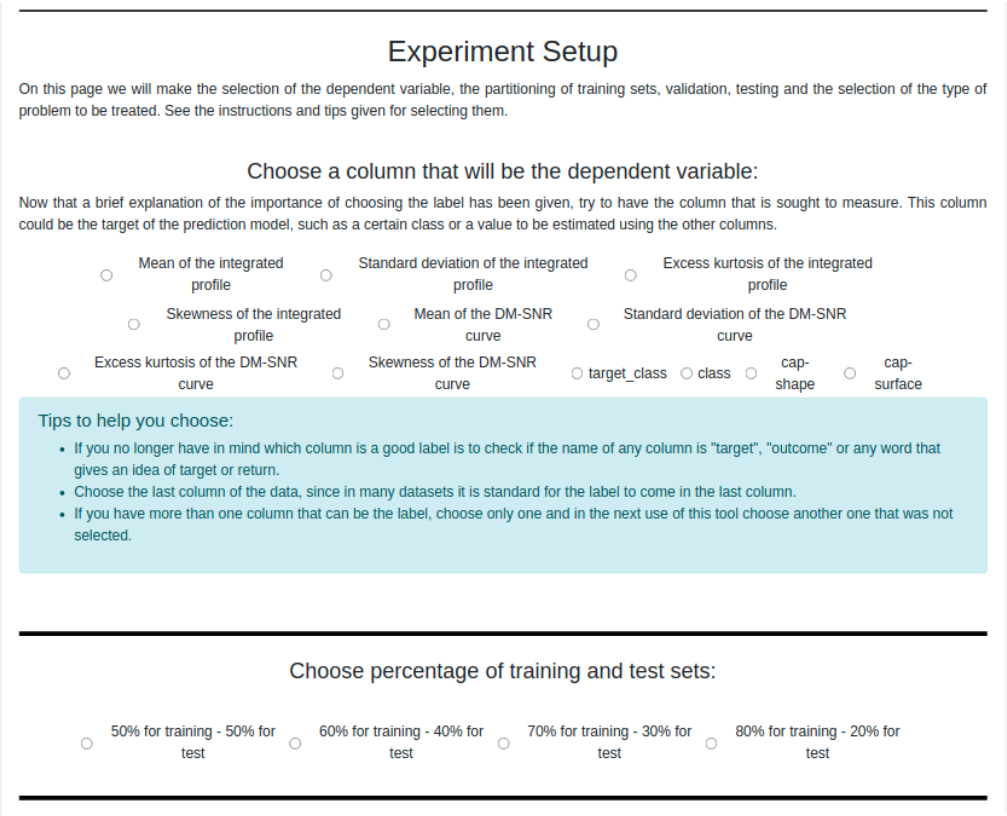


Figura 40 – Tela de configuração do experimento. Fonte: Autor.

4.2.3 Escolher o tipo de problema a ser tratado (regressão ou classificação)

Nesta etapa, o tipo de tarefa preditiva (regressão ou classificação) deve ser definido. A ferramenta *DSAdvisor* solicita que o usuário indique o tipo de problema preditivo que será

solucionado: uma classificação ou uma regressão. Para orientar essa escolha, a *DSAdvisor* exibe informações básicas sobre cada categoria de problema, como mostrado na Figura 41.

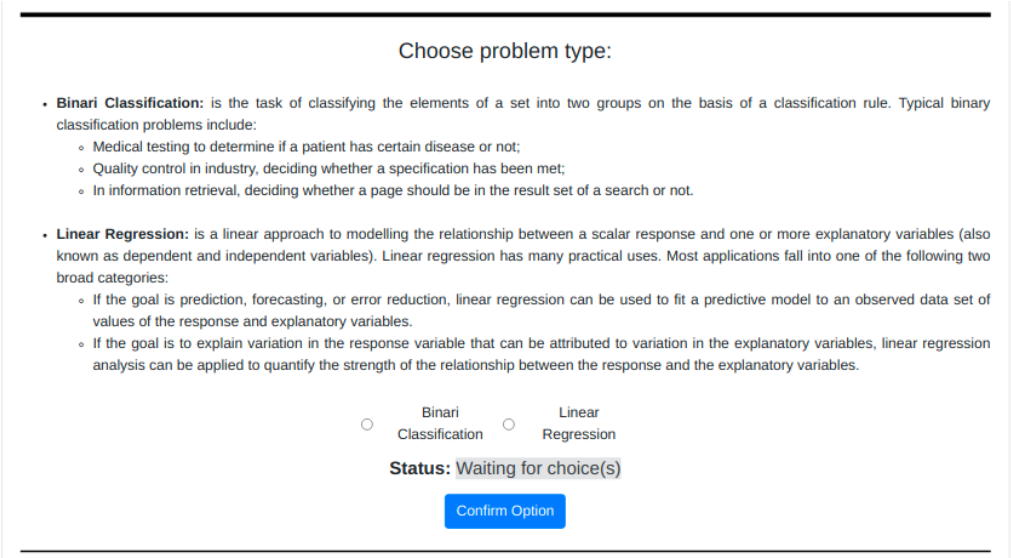


Figura 41 – Tela para escolha do tipo de problema a ser solucionado. Fonte: Autor.

4.2.4 Aplicar o encoder de rótulos para as variáveis categóricas

Conforme orienta o guia proposto no capítulo anterior, o conjunto de dados fornecido pelo usuário pode conter variáveis categóricas. Porém, sabe-se que alguns algoritmos preditivos não funcionam adequadamente com variáveis categóricas. A solução para este problema consiste em converter as variáveis categóricas em variáveis numéricas, o que é chamado de "*encoder*" (SHARMA *et al.*, 2020). Neste sentido, a ferramenta *DSAdvisor* aplica o método *Label Encoder* nas variáveis categóricas. Esta estratégia foi escolhida por ser de fácil implementação e por não aumentar a dimensionalidade do conjunto de dados, o que ocorre, por exemplo, com o método *One-hot Encoder*.

4.2.5 Escolher algoritmos preditivos

Uma vez definido o tipo de problema a ser solucionado, o usuário precisa indicar que algoritmos deseja avaliar. A ferramenta *DSAdvisor* disponibiliza 14 diferentes algoritmos, são eles:

- Algoritmos de regressão disponíveis:
  - *Linear Regression*
  - *Decision Tree*

    - *MLP-Regressor*
    - *Support Vector Regression*
    - *Gaussian Process Regressor*
- Algoritmos de classificação disponíveis:
    - *Logistic Regression*
    - *K-Nearest Neighbors*
    - *Decision Trees*
    - *Support Vector Machine*
    - *Naive Bayes*
    - *MLP-Classifier*
    - *Gaussian Process Classifier*
    - *Linear Discriminant Analysis*
    - *Quadratic Discriminant Analysis*

Contudo, com a finalidade de orientar a escolha do usuário, a ferramenta *DSAdvisor* irá exibir apenas os algoritmos relacionados ao tipo de problema previamente definido. Assim, se o usuário indicou que o problema investigado é de classificação, serão exibidos apenas os algoritmos Logistic Regression, K-Nearest Neighbors, Decision Trees, Support Vector Machine, Naive Bayes, MLP-Classifier, Gaussian Process Classifier, Linear Discriminant Analysis e Quadratic Discriminant Analysis, conforme ilustra a Figura 42.

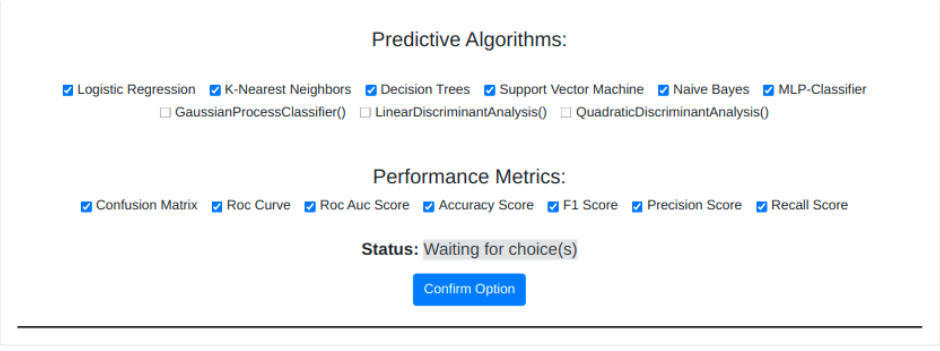


Figura 42 – Tela de seleção de algoritmos e métricas. Fonte: Autor.

4.2.6 Selecionar métricas de desempenho

Uma vez definido o tipo de problema a ser solucionado e os algoritmos que serão avaliados, o usuário precisa indicar que métricas serão utilizadas para avaliar o desempenho dos modelos preditivos. Logicamente, é importante escolher métricas que sejam adequadas ao tipo

de problema (regressão ou classificação), ao algoritmo utilizado e às características do conjunto de dados (balanceado ou desbalanceado, por exemplo). A ferramenta *DSAdvisor* disponibiliza 11 métricas distintas, são eles:

- Métricas para problemas de regressão:
  - *Mean Absolute Error*
  - *Mean Squared Error*
  - *Mean Squared Log Error*
  - *Mean Absolute Error*
  - *R2 Score*

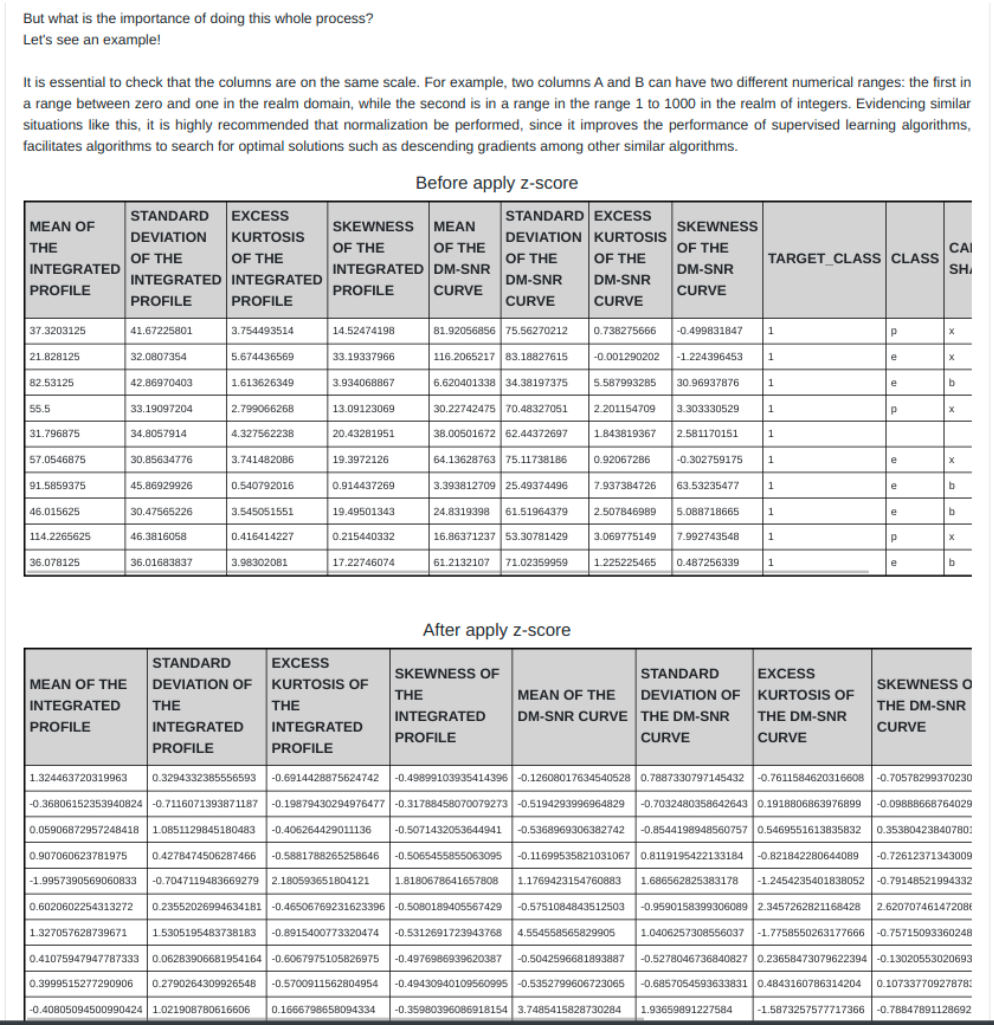
  Métricas para problemas de classificação:
- - *Confusion Matrix*
  - *Roc Curve*
  - *Roc Auc Score*
  - *Accuracy Score*
  - *F1 Score*
  - *Precision Score*
  - *Recall Score*

Contudo, com a finalidade de orientar a escolha do usuário, a ferramenta *DSAdvisor* irá exibir apenas as métricas relacionadas ao tipo de problema previamente definido. Assim, se o usuário indicou que o problema investigado é de classificação, serão exibidos apenas as métricas Confusion Matrix, Roc Curve, Roc Auc Score, Accuracy Score, F1 Score, Precision Score e Recall Score, conforme ilustra a Figura 42.

4.2.7 Escolher técnicas de normalização

Como mencionado anteriormente, a normalização e a padronização são técnicas frequentemente aplicadas na etapa de preparação dos dados, com o objetivo de colocá-los em um intervalo de valores comuns, a fim de evitar que o modelo preditivo fique enviesado para as variáveis com maior ordem de grandeza. Tanto a normalização quanto a padronização possuem o mesmo objetivo: transformar todas as variáveis na mesma ordem de grandeza. A diferença básica é que a padronização de uma determinada variável irá resultar em valores com uma média igual a 0 e um desvio padrão igual a 1. Já a normalização irá resultar em valores dentro do intervalo de 0 e 1, e caso tenha resultado negativo entre -1 e 1.

A ferramenta *DSAdvisor* fornece suporte a duas técnicas diferentes: a padronização *z-score* e a "normalização Min-Máx". Inicialmente, a *DSAdvisor* exibe uma tela contendo informações básicas, exemplos de utilização e dicas para interpretação dos resultados dessas duas técnicas, conforme ilustra a Figura 43. Em seguida, a *DSAdvisor* permite ao usuário escolher qual das duas técnicas deseja utilizar. Contudo, o método *z-score* é pré-selecionado por padrão, como mostra a Figura 44. Essa escolha deve-se ao fato da técnica *z-score* ser mais indicada para algoritmos baseados no gradiente descendente (PONTI; COSTA, 2018).



But what is the importance of doing this whole process?
Let's see an example!

It is essential to check that the columns are on the same scale. For example, two columns A and B can have two different numerical ranges: the first in a range between zero and one in the realm domain, while the second is in a range in the range 1 to 1000 in the realm of integers. Evidencing similar situations like this, it is highly recommended that normalization be performed, since it improves the performance of supervised learning algorithms, facilitates algorithms to search for optimal solutions such as descending gradients among other similar algorithms.

Before apply z-score

MEAN OF THE INTEGRATED PROFILE	STANDARD DEVIATION OF THE INTEGRATED PROFILE	EXCESS KURTOSIS OF THE INTEGRATED PROFILE	SKEWNESS OF THE INTEGRATED PROFILE	MEAN OF THE DM-SNR CURVE	STANDARD DEVIATION OF THE DM-SNR CURVE	EXCESS KURTOSIS OF THE DM-SNR CURVE	SKEWNESS OF THE DM-SNR CURVE	TARGET_CLASS	CLASS	CA SH
37.3203125	41.67225801	3.754493514	14.52474198	81.92056856	75.56270212	0.738275666	-0.499831847	1	p	x
21.828125	32.0807354	5.674436569	33.19337966	116.2065217	83.18827615	-0.001290202	-1.224396453	1	e	x
82.53125	42.86970403	1.613626349	3.934068867	6.620401338	34.38197375	5.587993285	30.96937876	1	e	b
55.5	33.19097204	2.799066268	13.09123069	30.22742475	70.48327051	2.201154709	3.303330529	1	p	x
31.796875	34.8057914	4.327562238	20.43281951	38.00501672	62.44372697	1.843819367	2.581170151	1		
57.0546875	30.85634776	3.741482086	19.3972126	64.13628763	75.11738186	0.92067286	-0.302759175	1	e	x
91.5859375	45.86929926	0.540792016	0.914437269	3.393812709	25.49374496	7.937384726	63.53235477	1	e	b
46.015625	30.47565226	3.545051551	19.49501343	24.8319398	61.51964379	2.507846989	5.088718665	1	e	b
114.2265625	46.3816058	0.416414227	0.215440332	16.86371237	53.30781429	3.069775149	7.992743548	1	p	x
36.078125	36.01683837	3.98302081	17.22746074	61.2132107	71.02359959	1.225225465	0.487256339	1	e	b

After apply z-score

MEAN OF THE INTEGRATED PROFILE	STANDARD DEVIATION OF THE INTEGRATED PROFILE	EXCESS KURTOSIS OF THE INTEGRATED PROFILE	SKEWNESS OF THE INTEGRATED PROFILE	MEAN OF THE DM-SNR CURVE	STANDARD DEVIATION OF THE DM-SNR CURVE	EXCESS KURTOSIS OF THE DM-SNR CURVE	SKEWNESS O THE DM-SNR CURVE
1.324463720319963	0.3294332385556593	-0.6914428875624742	-0.49899103935414396	-0.12608017634540528	0.7887330797145432	-0.7611584620316608	-0.70578299370230
-0.36806152353940824	-0.7116071393871187	-0.19879430294976477	-0.31788458070079273	-0.5194293996964829	-0.7032480358642643	0.1918806863976899	-0.09888668764029
0.059068729572484 18	1.0851129845180483	-0.406264429011136	-0.5071432053644941	-0.5368969306382742	-0.8544198948560757	0.5469551613835832	0.353804238407801
0.907060623781975	0.4278474506287466	-0.5881788265258646	-0.5065455855063095	-0.11699535821031067	0.8119195422133184	-0.821842280644089	-0.72612371343009
-1.9957390569060833	-0.7047119483669279	2.180593651804121	1.8180678641657808	1.1769423154760883	1.686562825383178	-1.2454235401838052	-0.79148521994332
0.6020602254313272	0.23552026994634181	-0.46506769231623396	-0.5080189405567429	-0.5751084843512503	-0.9590158399306089	2.3457262821168428	2.620707461472086
1.327057628739671	1.530519548373818 3	-0.8915400773320474	-0.5312691723943768	4.554558565829905	1.0406257308556037	-1.7758550263177666	-0.75715093360248
0.41075947947787333	0.062839066819541 64	-0.6067975105826975	-0.4976986939620387	-0.5042596681893887	-0.5278046736840827	0.2365847307962239 4	-0.13020553020693
0.3999515277290906	0.2790264309926548	-0.5700911562804954	-0.4943094010956099 5	-0.5352799606723065	-0.6857054593633831	0.4843160786314204	0.10733770927878
-0.40805094500990424	1.021908780616606	0.1666798658094334	-0.35980396086918154	3.7485415828730284	1.93659891227584	-1.5873257577717366	-0.78847891128692

Figura 43 – Tela de técnicas de normalização com um exemplo prático. Fonte: Autor.

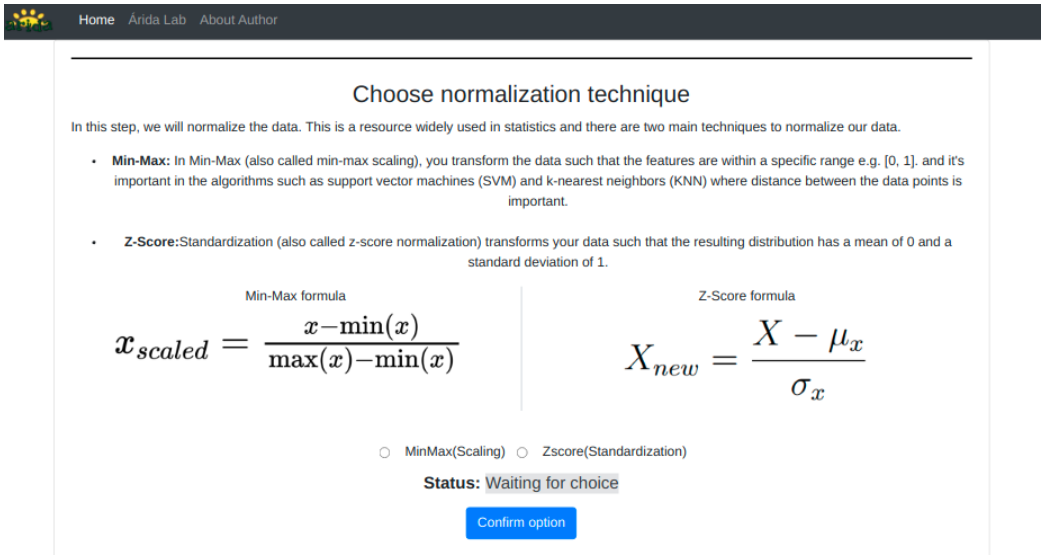


Figura 44 – Tela de técnicas de normalização com opção de escolha entre "*Minmax*"e "*Z-score*". Fonte: Autor.

4.2.8 Aplicar métodos de detecção de outliers

Como mencionado anteriormente, *Outliers* são um dos principais problemas enfrentados ao se construir um modelo preditivo, pois podem influenciar negativamente o seu desempenho, principalmente se forem decorrentes de erros na coleta dos dados. Neste caso, os *outliers* podem ser descartados. Portanto, é de fundamental importânciaidentificar a presença de *outliers* (LIU *et al.*, 2004). Neste sentido, a ferramenta *DSAdvisor* aplica, para cada variável numérica presente no conjunto de dados, o método do IQR ajustado. Em seguida, a *DSAdvisor* exibe o número de instâncias contendo *outliers*, além da porcentagem e do total de *outliers* detectados, como ilustra a Figura 45. O funcionamento da *DSAdvisor* na detecção de *outliers* é descrito no Algoritmo 8. Adicionalmente, se o usuário desejar saber exatamente quais são os valores discrepantes a fim de realizar uma análise mais detalhada, a *DSAdvisor* exibe uma tabela contendo, para cada variável numérica, os *outliers* detectados e as posições onde esses se encontram no conjunto de dados, conforme ilustra a Figura 46.

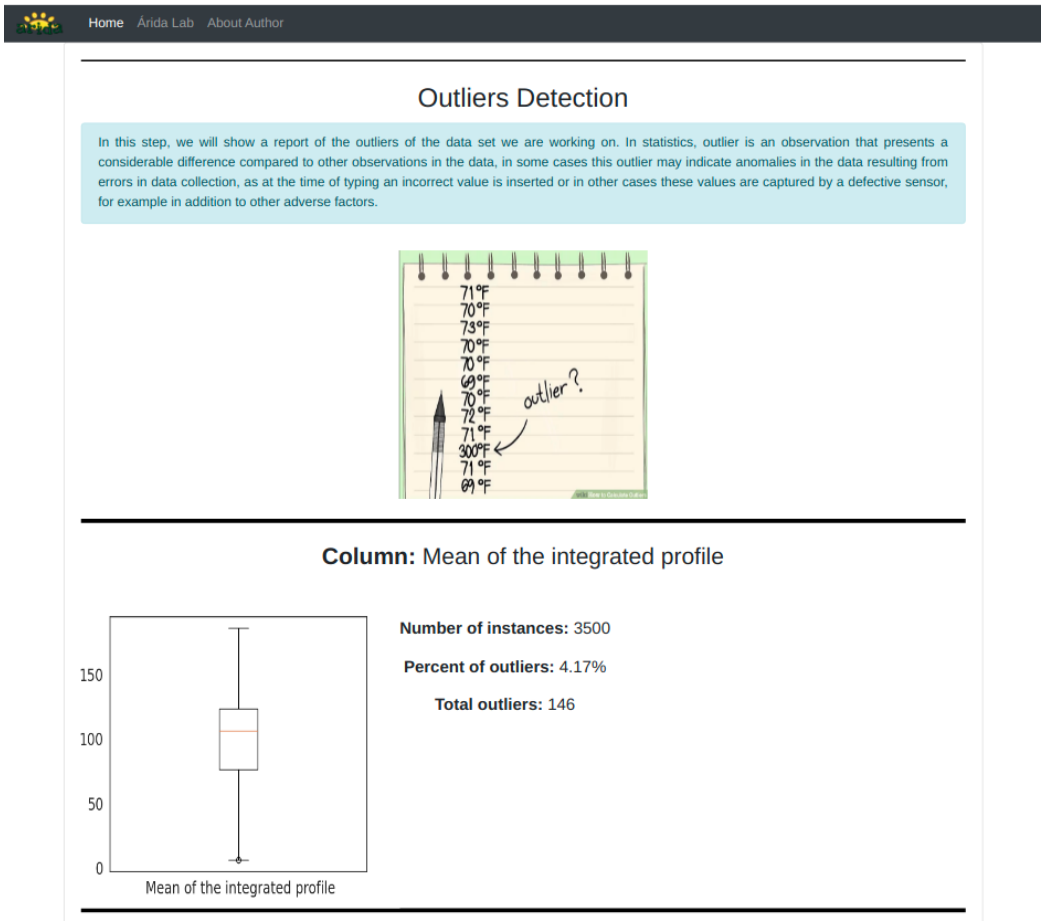
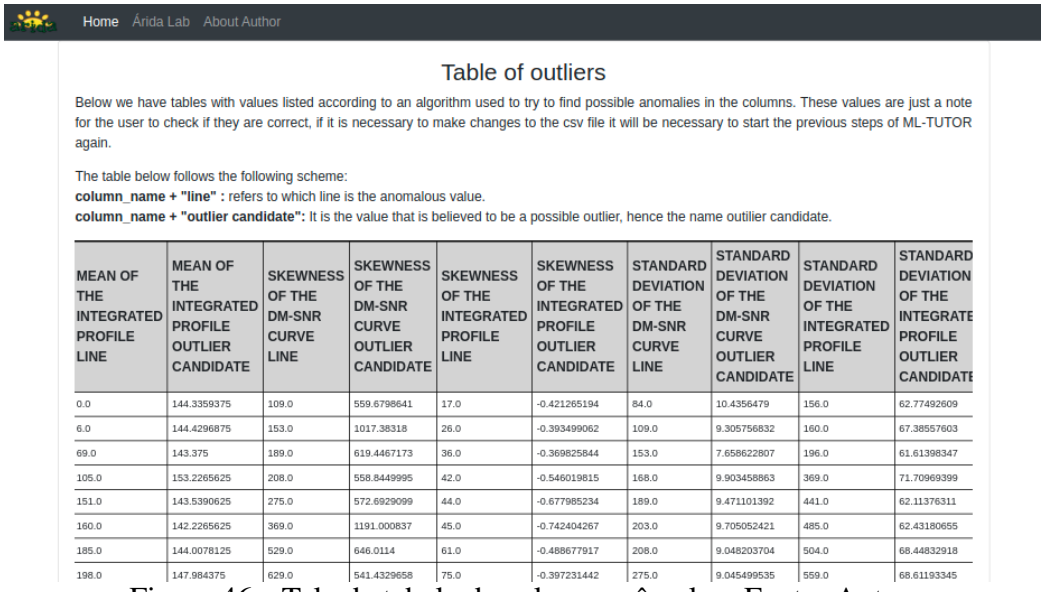


Figura 45 – Tela da funcionalidade de "outlier detection". Fonte: Autor.



MEAN OF THE INTEGRATED PROFILE LINE	MEAN OF THE INTEGRATED PROFILE OUTLIER CANDIDATE	SKEWNESS OF THE DM-SNR CURVE LINE	SKEWNESS OF THE DM-SNR CURVE OUTLIER CANDIDATE	SKEWNESS OF THE INTEGRATED PROFILE LINE	SKEWNESS OF THE INTEGRATED PROFILE OUTLIER CANDIDATE	STANDARD DEVIATION OF THE DM-SNR CURVE LINE	STANDARD DEVIATION OF THE DM-SNR CURVE OUTLIER CANDIDATE	STANDARD DEVIATION OF THE INTEGRATED PROFILE LINE	STANDARD DEVIATION OF THE INTEGRATE PROFILE OUTLIER CANDIDATE
0.0	144.3359375	109.0	559.6798641	17.0	-0.421265194	84.0	10.4356479	156.0	62.77492609
6.0	144.4296875	153.0	1017.38318	26.0	-0.393499062	109.0	9.305756832	160.0	67.38557603
69.0	143.375	189.0	619.4467173	36.0	-0.369825844	153.0	7.658622807	196.0	61.61398347
105.0	153.2265625	208.0	558.8449995	42.0	-0.546019815	168.0	9.903458863	369.0	71.70969399
151.0	143.5390625	275.0	572.6929099	44.0	-0.677985234	189.0	9.471101392	441.0	62.11376311
160.0	142.2265625	369.0	1191.000837	45.0	-0.742404267	203.0	9.705052421	485.0	62.43180655
185.0	144.0078125	529.0	646.0114	61.0	-0.488677917	208.0	9.048203704	504.0	68.44832918
198.0	147.984375	629.0	541.4329658	75.0	-0.397231442	275.0	9.045499535	559.0	68.61193345

Figura 46 – Tela da tabela de valores anômalos. Fonte: Autor.

Algoritmo 8: Aplicar detecção de outliers. Fonte: Autor.

início

for *cada variável V numérica* **do**
NumOutliers[V] $\leftarrow$ 0
PercentOutilers[V] $\leftarrow$ 0
end
for *cada variável V numérica* **do**
NumOutliers[V], PercentOutilers[V] $\leftarrow$ AdjustableIQR(V)
end
Exibir NumOutliers[V], PercentOutilers[V]

fim

4.2.9 Aplicar métodos de seleção de atributos

Como destacado anteriormente, um conjunto de dados pode conter atributos (variáveis independentes) redundantes. Um atributo redundante é aquele que pode ter seu valores inferidos por meio de outros atributos, ou seja, que possui uma forte correlação com outros atributos. Assim, sua presença no conjunto de dados acaba não contribuindo para a melhoria do desempenho do modelo preditivo (KUHN *et al.*, 2013). Logo, esse atributo pode ser removido do conjunto de dados. Portanto, a seleção de atributos pode ser definida como o processo de obtenção de um subconjunto dos dados originais sem atributos redundantes.

A ferramenta *DSAdvisor* implementa uma heurística de seleção de atributos baseada em 5 métodos distintos (9):

- *Chi Squared*
- *Information Gain*
- *Mutual Info*
- *F-Value*
- *Gain Ratio*

Os métodos Chi Squared e Information Gain foram implementados diretamente na ferramenta *DSAdvisor*. Já para os métodos Mutual Info, F-Value e Gain Ratio utilizamos a implementação do *Sklearn*.

O 9 descreve a heurística de seleção de atributos implementada pela *DSAdvisor*. Inicialmente, computa-se, para cada variável do conjunto de dados, o resultado de cada um dos cinco métodos de seleção de atributos: Chi Squared, Information Gain, Mutual Info, F-Value e Gain Ratio. Para cada variável armazena-se a soma dos resultados desses cinco métodos,

utilizando-se para isso o vetor “VecSumTotal”. Em seguida, iremos copiar o maior valor presente no vetor “VecSumTotal” para avariável “TopScoreSum”. A heurística utiliza como “threshold” o valor de “TopScoreSum” dividido por dois. Esse “threshold” será utilizado para definir as variáveis que serão selecionadas, ou seja, que continuarão presentes no conjunto de dados. Neste sentido, as variáveis cuja soma dos resultados dos cinco métodos for menor que o valor do “threshold” serão removidas do conjunto de dados. Por fim, a ferramenta *DSAdvisor* mostra os resultados obtidos, bem como exibe um “check-box” para cada uma das variáveis, sendo que aquelas selecionadas pela heurística estarão previamente marcadas, enquanto as demais estão desmarcadas, como ilustra a Figura 47. Contudo, o usuário pode alterar a sugestão da *DSAdvisor*.

Algoritmo 9: Heurística de seleção de atributos. Fonte: Autor.

```
início
    Threshold ← 0
    TopScoreSum ← 0
    for cada variável V do
        VecChiSquared[V] ← 0 /* Chi Squared para V */
        VecInforGain[V] ← 0 /* Information Gain para V */
        VecMutualInfor[V] ← 0 /* Mutual Info para V */
        VecFValue[V] ← 0 /*F-Value para V*/
        VecGainRatio[V] ← 0 /*Gain Ratio para V*/
        VecSumTotal[V] ← 0 /*Soma de cada método para V*/
    end
    for cada variável V do
        VecChiSquared[V] ← ChiSquared(V)
        VecInforGain[V] ← Information Gain(V)
        VecMutualInfor[V] ← Mutual Info(V)
        VecFValue[V] ← F Value(V)
        VecGainRatio[V] ← Gain Ratio(V)
        VecSumTotal[V] ← VecChiSquared[V] + VecInforGain[V] + VecMutualInfor[V]
         + VecFValue[V] + VecGainRatio[V]
    end
    TopScoreSum ← HighScore(VecSumTotal)
    Threshold ← TopScoreSum/2
    for cada variável V do
        if VecSumTotal[V] ≥ Threshold then
            Recomendar V para a próxima fase
        end
        else
            Não recomendar V para a próxima fase
        end
    end
fim
```

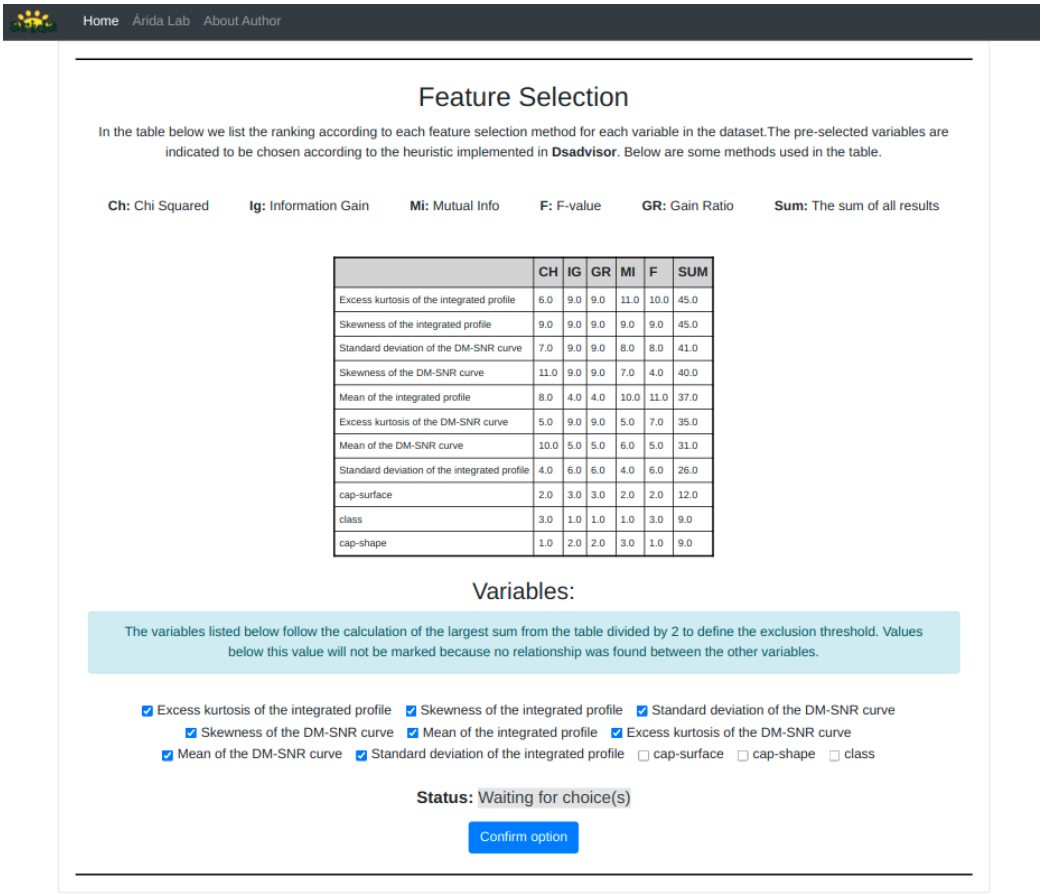


Figura 47 – Tela de seleção de variáveis com o resultado da Heurística de Seleção de Atributos aplicados ao conjunto de dados "*Pulse Star*". Fonte: Autor.

4.2.10 Mostrar a frequência para cada classe da variável dependente

Para problemas de classificação é importante verificar como se configuram as classes da variável dependente. Neste sentido, a ferramenta *DSAdvisor* exibe para cada classe presente na variável dependente a porcentagem de suas proporções, conforme ilustrado nas figuras 48 e 49. A depender das proporções das diferentes classes presentes na variável dependente, a *DSAdvisor* pode sugerir a técnica de balanceamento mais indicada.

4.2.11 Escolher técnica de balanceamento

Como mencionamos anteriormente, em problemas de classificação, dizemos que um conjunto de dados é desbalanceado quando as quantidades de instâncias em cada uma das classes presentes na variável dependente estão desequilibradas. Vale destacar que a maioria dos algoritmos de classificação funcionam melhor quando os números de instâncias em cada classe são semelhantes ou próximos (LONGADGE; DONGRE, 2013).

Neste sentido, inicialmente, a ferramenta *DSAdvisor* verifica se o tipo de problema indicado pelo usuário foi de "regressão". Nesse caso, a *DSAdvisor* passa para a atividade de geração de modelos, pois em atividades de regressão só em casos muito específicos se utiliza métodos de balanceamento de dados. Caso contrário, a *DSAdvisor* exibe a quantidade de instâncias e a porcentagem de cada classe presente na variável dependente, como ilustra a Figura 48. Em seguida, a *DSAdvisor* apresenta três opções ao usuário:

– *Oversampling*;
– *Undersampling*;
– Sem balanceamento;

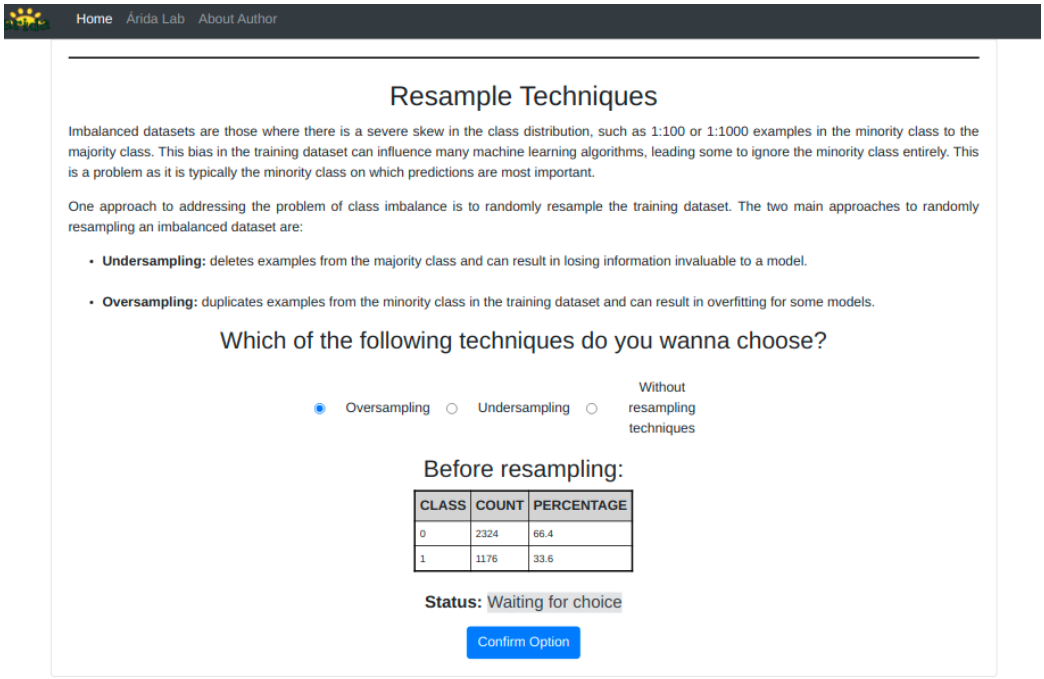


Figura 48 – Proporção das classes presentes na variável dependente antes da aplicação da técnica de balanceamento. Fonte: Autor.

A técnica de balanceamento escolhida será aplicada na fase de “geração de modelos”. Contudo, a *DSAdvisor* ilustra como ficará o balanceamento dos dados após a aplicação da técnica selecionada, como mostra a Figura 49.

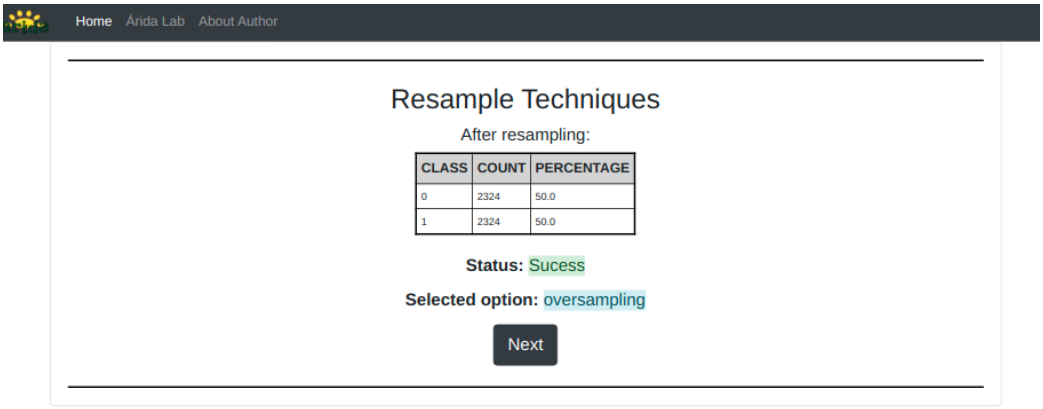


Figura 49 – Proporção das classes presentes na variável dependente após a aplicação da técnica de balanceamento. Fonte: Autor.

4.3 Fase 3 - Construção de Modelos Preditivos

Nessa última fase, a ferramenta *DSAdvisor* busca auxiliar o usuário na construção de modelos preditivos, além de orientar a análise dos resultados obtidos e assegurar a reprodutibilidade dos experimentos realizados.

4.3.1 Geração de Modelos

Nessa primeira etapa, a ferramenta *DSAdvisor* irá utilizar o conjunto de dados gerado na Fase 2, bem como as configurações realizadas anteriormente pelo usuário, tais como a escolha dos algoritmos, métricas, variável dependente e porcentagem dos conjuntos de treinamento e de teste, com a finalidade de construir modelos preditivos.

4.3.1.1 Particionamento dos conjuntos de treino e teste

Inicialmente, a ferramenta *DSAdvisor* aplica a estratégia de particionamento de dados selecionada previamente. Neste sentido, o conjunto de dados é dividido nos conjuntos de treino e teste, de acordo com a porcentagem selecionada pelo usuário. Para isso, a *DSAdvisor* utiliza o método "*train-test-split*"da biblioteca *Sklearn*, o qual recebe como entrada a porcentagem de divisão e o conjunto de dados (x, y), e retorna os conjuntos de treinamento (x-train, y-train) e teste (x-test, y-test).

4.3.1.2 Aplicar técnicas de balanceamento de dados

Nesta etapa, caso o usuário tenha verificado que o conjunto de dados utilizado está desbalanceado e selecionado uma técnica de balanceamento, esta será aplicada pela ferramenta *DSAdvisor*. As técnicas de balanceamento de dados suportadas pela *DSAdvisor* foram implementadas por meio da biblioteca "*imblearn.pipeline*". Essa biblioteca foi escolhida por permitir a construção de modelos preditivos em conjunto com a aplicação de técnicas de balanceamento e normalização do conjunto de treino.

4.3.1.3 Ajuste de hiper parâmetros

Como destacado anteriormente, todo algoritmo de predição tem um conjunto de hiperparâmetros, os quais controlam o processo de aprendizagem e determinam o seu desempenho (HUTTER *et al.*, 2019). Logo, surge o desafio de encontrar valores para os hiperparâmetros do algoritmo utilizado que proporcionem modelos preditivos com melhores desempenhos, o que é chamado de "otimização"de hiperparâmetros.

Neste sentido, a ferramenta *DSAdvisor* aplica um método de "otimização" de hiperparâmetros chamado *RandomSearchCv*, o qual é uma junção do *RandomSearch* com o método *k-fold cross validatin* configurado para 5 conjuntos randomizados. Caso o usuário tenha selecionado a classificação com o problema preditivo a ser solucionado, a *DSAdvisor* utiliza a acurácia como função de otimização. Caso, contrário, a *DSAdvisor* utiliza o coeficiente de determinação (r2) como função de otimização. Após a identificação dos melhores valores para os hiperparâmetros um novo modelo é treinado utilizando os valores encontrados e todo o conjunto de treino. Por fim, o modelo preditivo obtido é avaliado utilizando-se os dados de teste.

4.3.2 Avaliar os modelos preditivos

Nessa etapa, a ferramenta *DSAdvisor* tem por finalidade auxiliar o usuário a avaliar o desempenho dos modelos preditivos gerados. Para comparar o desempenho de modelos diferentes, a *DSAdvisor* utiliza métricas já consolidadas.

Para problemas de classificação a *DSAdvisor* computa e exibe as seguintes métricas:

- Matriz de Confusão(*Confusion Matrix*)
- Curva Roc(*Roc Curve*)
- Acurácia(*Accuracy Score*)

– Precisão(*Precision Score*)
– Revocação(*Recall Score*)

Já para problemas de regressão a *DSAdvisor* computa e exibe as seguintes métricas:

– Erro Quadrático Médio (MSE)

$$R^2 = 1 - \frac{\sum_{n=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{n=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.1}$$

– R-Quadrado
– Raiz do erro quadrático médio (RMSE)
– Erro Absoluto Médio (MAE)
– Erro Percentual Absoluto Médio (MAPE)

4.3.3 Apresentar os resultados obtidos

Nesta etapa, a ferramenta *DSAdvisor* busca organizar e apresentar os resultados obtidos, ou seja, os algoritmos, as características, as técnicas de pré-processamento aplicadas e os hiperparâmetros dos modelos que apresentaram os melhores desempenhos. O objetivo principal desta etapa é auxiliar o usuário a encontrar o modelo mais adequado para o problema investigado. Neste sentido, a *DSAdvisor* gera uma página contendo todos os resultados obtidos, incluindo para cada um dos algoritmos previamente selecionados pelo usuário, o resultado das métricas correspondentes ao problema preditivo enfrentado (classificação ou regressão), conforme ilustra a Figura 50.

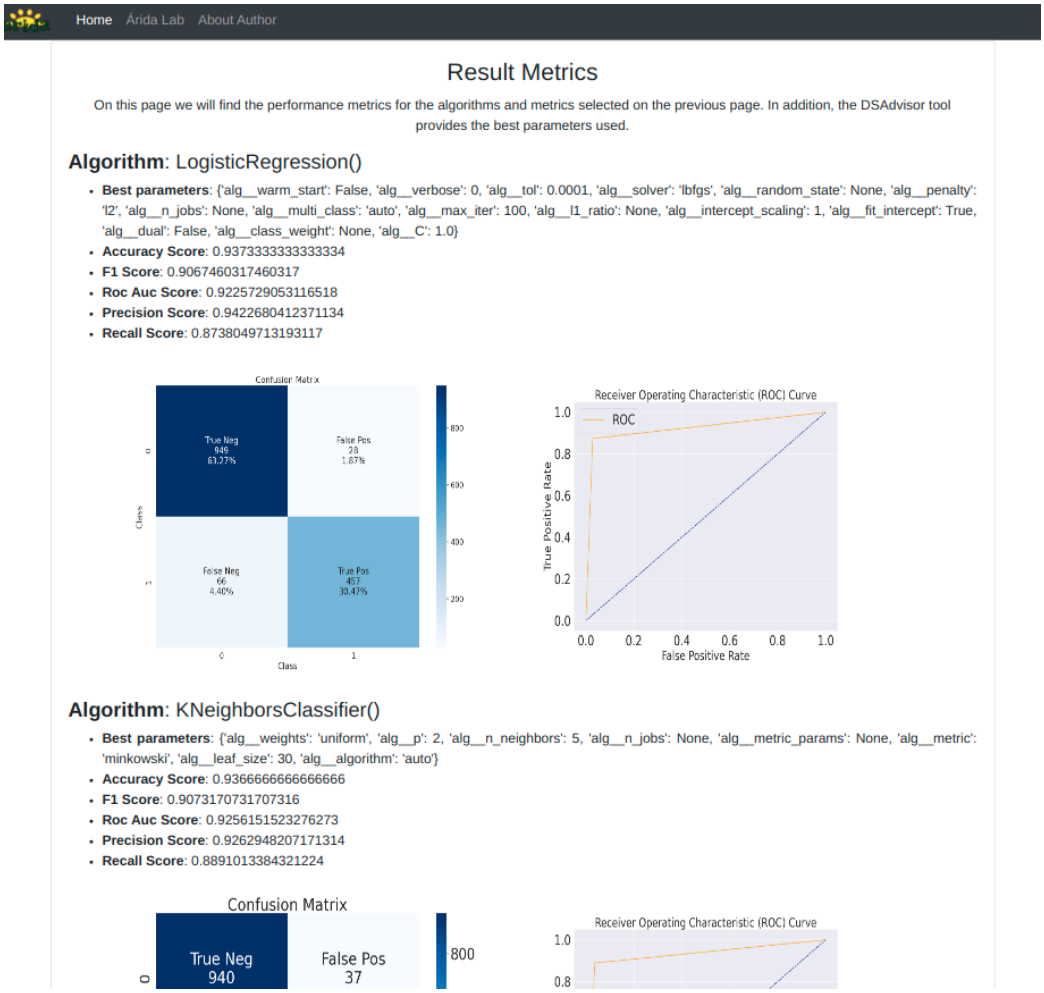


Figura 50 – Tela de avaliação dos modelos preditivos. Fonte: Autor.

4.3.4 Assegurar a reprodutibilidade

Por fim, a ferramenta *DSAdvisor* busca auxiliar o usuário a garantir a reprodutibilidade dos seus experimentos, com a finalidade de fornecer credibilidade ao estudo realizado e possibilitar que este seja executado por outros usuários. Neste sentido, a *DSAdvisor* permite que o usuário gere três arquivos distintos (conforme mostra a Figura 51):

– Os dados do conjunto de treinamento (x-train, y-train);
– Os dados do conjunto de teste (x-test, y-test) e
– As configurações realizadas pelo usuário (chamado *log*).

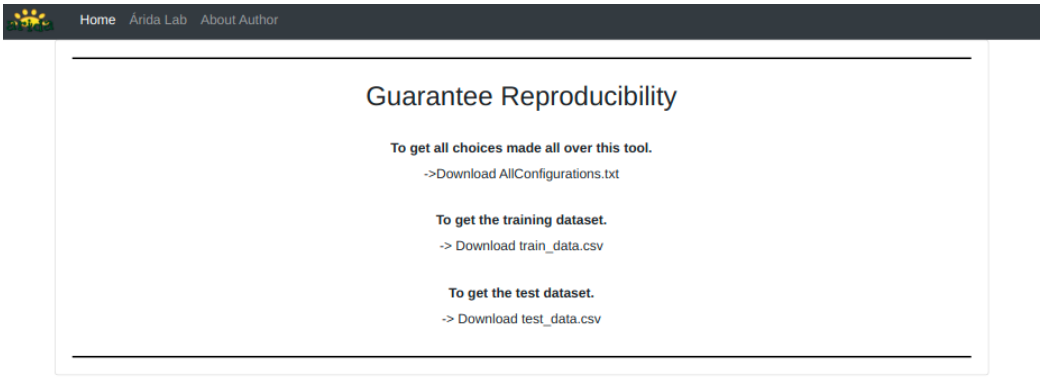


Figura 51 – Tela de reprodutibilidade com os arquivos para download. Fonte: Autor.

O arquivo de configurações (*log*) é um relatório que contém todas as opções selecionadas pelo usuário, tais como: tipo do problema, variável dependente, técnica de balanceamento de dados utilizada, técnica de normalização aplicada, conjunto de atributos selecionados, algoritmos executados, valores encontrados para os hiperparâmetros, além das tabelas e gráficos gerados pela *DSAdvisor*, conforme ilustra a Figura 52. Com este arquivo de configurações, espera-se que os experimentos executados possam ser reproduzidos por outros usuários.

```
    ALL THE CHOICES MADE:
--------------START-------------
'Dataset details':'New pulse_star.csv is a dataset with less rows than pulse star dataset and has columns from other
dataset.'
'file_name':'new_pulse_star'
'removed variables':'[]'
'missing values codes':'['nan', 'empty', 'other_code']'
'special missing value code':'px399'
'selected_ditribution_to_best_fit_method':'['alpha', 'anglit', 'arcsine', 'beta', 'betaprime', 'bradford', 'cauchy', 'chi',
'chi2', 'cosine', 'dgamma', 'dweibull', 'erlang', 'expon', 'exponnorm', 'f', 'fatiguelife', 'foldcauchy', 'foldnorm',
'gamma', 'genlogistic', 'gennorm', 'genpareto', 'gilbrat', 'gumbel_l', 'gumbel_r', 'halfcauchy', 'halflogistic',
'halfnorm', 'hypsecant', 'invgamma', 'invweibull', 'laplace', 'levy', 'levy_l', 'loggamma', 'logistic', 'loglaplace',
'lognorm', 'loguniform', 'lomax', 'maxwell', 'moyal', 'nakagami', 'norm', 'pareto', 'pearson3', 'powerlaw', 'rayleigh',
'rdist', 'semicircular', 't', 'truncexpon', 'uniform', 'wald', 'weibull_max', 'weibull_min', 'wrapcauchy']'
'users_distribution_selected':'['GENLOGISTIC', 'LOGGAMMA', 'FOLDCAUCHY', 'ALPHA', 'BETAPRIME', 'FATIGUELIFE', 'DGAMMA',
'NAKAGAMI']'
'dependent_variable':'target_class'
'test_size_percent':'0.3'
'problem_type':'Classification'
'normalization':'MinMaxScaler()'
'feature_selection_variables':'[' Excess kurtosis of the integrated profile', ' Skewness of the integrated profile', '
Standard deviation of the DM-SNR curve', ' Skewness of the DM-SNR curve', ' Mean of the integrated profile', ' Excess
kurtosis of the DM-SNR curve', ' Mean of the DM-SNR curve', ' Standard deviation of the integrated profile']'
'resample_technique_choiced':'SMOTE(random_state=42)'
'predictive_alg_list':'['LogisticRegression()', 'KNeighborsClassifier()', 'tree.DecisionTreeClassifier()', 'SVC()',
'GaussianNB()', 'MLPClassifier()']'
'metrics_list':'['confusion_matrix', 'roc_curve', 'roc_auc_score', 'accuracy_score', 'f1_score', 'precision_score',
'recall_score']'
--------------END--------------
```

Figura 52 – Exemplo do arquivo de configurações (*log*). Fonte: Autor.

5 AVALIAÇÃO DE USABILIDADE

Este capítulo descreve a avaliação da usabilidade da ferramenta *DSAdvisor* e a satisfação do usuário por meio de dois diferentes testes: *Net Promoter Score* (NPS) e *System Usability Scale* (SUS).

5.1 Testes de Usabilidade

A palavra "usabilidade" indica a finalidade de facilitar o uso de um produto ou serviço. Neste sentido, usabilidade é o termo utilizado para descrever a qualidade de uso de um produto ou serviço, em relação a diferentes aspectos, tais como: utilidade, eficiência, eficácia, satisfação, facilidade de aprendizado e acessibilidade (BEVAN *et al.*, 2016). A utilidade aborda a capacidade do usuário de utilizar um determinado produto ou serviço para atingir algum objetivo específico. Eficiência é mensurada através da rapidez e da precisão em o usuário atinge seus objetivos. A eficácia é mensurada através da capacidade do usuário de completar uma determinada tarefa. Facilidade de aprendizado é o conhecimento acumulado usado pelo usuário para manusear um determinado produto ou serviço. A satisfação se concentra na percepção do usuário sobre o produto. Acessibilidade relaciona-se com a capacidade do usuário de ter acesso ao produto sempre que necessário (CHARLTON; O'BRIEN, 2019).

Quando a usabilidade é avaliada durante o processo de desenvolvimento de um produto ou serviço, vários problemas podem ser identificados e corrigidos, como, por exemplo, pode-se reduzir o tempo de acesso à uma determinada informação, alterando a sua posição na interface a fim de que esta seja encontrada mais facilmente, evitando assim a frustração do usuário de não encontra-la. Atualmente, busca-se identificar os problemas de usabilidade tão logo eles possam ser detectados. Uma vez identificado, o problema pode ser solucionado ou, pelo menos, seus efeitos podem ser minimizados. Existe uma diversidade de métodos de avaliação que podem ser utilizados em diferentes etapas do desenvolvimento de um produto ou serviço (NIELSEN *et al.*, 2012). Os testes de usabilidade são técnicas de projeto centradas nos usuários e usadas para avaliar um produto ou serviço em situações do cotidiano. Eles permitem obter as percepções (*feedback*) dos próprios usuários que trabalham ou executam atividades com o objeto de análise. Adicionalmente, durante a realização dos testes de usabilidade, é bastante frequente que os usuários surpreendam os avaliadores ao efetuarem ações inesperadas enquanto estão avaliando o produto (CHARLTON; O'BRIEN, 2019).

Para realizar a avaliação da usabilidade de um determinado produto ou serviço, é aconselhável utilizar testes já consagrados e instrumentos de avaliações bem sedimentados. Os testes de usabilidade mais comuns são: Heurísticas de Nielsen (NIELSEN, 1995), *System Usability Scale* (SUS) (LEWIS, 2018), *Net Promoter Scores* (NPS) (MANDAL, 2014), *Software Usability Measurement Inventory* (SUMI) (KIRAKOWSKI; CORBETT, 1993), *Website Analysis and Measurement Inventory Questionnaire* (Wammi) (CLARIDGE; KIRAKOWSKI, 2011) e *User Experience Questionarie* (UEQ) (SCHREPP, 2015).

5.1.1 Net Promoter Score (NPS)

O *Net Promoter Scores* (NPS) é uma técnica que tem por finalidade medir a satisfação de um cliente ou usuário para que as empresas ou prestadoras de serviço possam avaliar e melhorar seu desempenho, seus produtos e serviços (RAS *et al.*, 2017). A ideia principal do NPS é que os clientes ou usuários sejam abordados quanto à probabilidade de recomendarem produtos/serviços para seus colegas (KORNETA, 2014).

No NPS, os clientes são rotulados como Promotores (*Promoters*), Passivos (*Passives*) ou Detratores(*Detractors*). Os promotores são classificados como clientes fiéis que sempre fornecerão recomendações de produtos/serviços para terceiros. Já os usuários com perfil passivo estão satisfeitos com os produtos/serviços da empresa, mas tem potencial para aceitar outros produtos/serviços ofertados por concorrentes. Por fim, os detratores são clientes insatisfeitos e desleais, que incentivam outras pessoas a não utilizarem os produtos/serviços da empresa. Os clientes que pontuam de 9 a 10 são chamados de promotores, já os que atribuem valores de 7 a 8 são chamados passivos, e de 1 a 6 são chamados de detratores.

5.1.1.1 Cálculo do NPS

O NPS é calculado da seguinte forma: a porcentagem do número de Promotores menos a porcentagem do total de Detratores. Importante ressaltar que os usuários passivos são desconsiderados no cálculo do NPS. Desta forma, o resultado do NPS pode variar de -100 a 100. Valores acima de 70 são considerados “excelentes”. Já os valores entre 50 e 70 são considerados “muito bons”. Por fim, valores entre 0 e 50 são considerados “bons” (LEE, 2018). A Figura 53 ilustra como calcular o NPS.

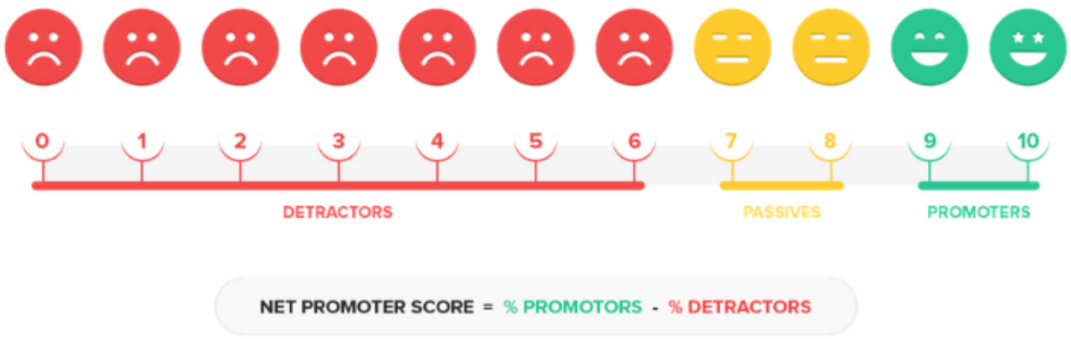


Figura 53 – Demonstração de cálculo do NPS. Fonte: link.

5.1.2 System Usability Scale (SUS)

O *System Usability Scale* (SUS) é um questionário padronizado amplamente utilizado para a avaliação da usabilidade percebida. Lewis e Sauro (2009) relatam que o método SUS foi utilizado em 43% das avaliações de usabilidade industrial. As citações do *Google Scholar* (examinadas em 25/08/2022) mostraram 13539 citações para o artigo que introduziu o SUS, mais precisamente, (BROOKE, 1996). Em seu formato mais comum, o SUS possui 10 itens de cinco pontos cada um, incluindo questões de cunho positivo e negativo, de maneira alternada, conforme ilustra a Figura 54.

5.1.2.1 Cálculo do SUS

Antes de discutirmos como realizar o cálculo do SUS, é importante lembrar que:

- O questionário possui 10 itens (ou questões);
- Cada item recebe valores entre 1 e 5;
- O questionário alterna itens positivos (ímpares) e negativos (pares).

Para calcular o valor do SUS, inicialmente, precisamos converter as pontuações brutas dos 10 itens em pontuações ajustadas (também conhecidas como “contribuições de pontuação”), as quais variam de 1 (avaliação mais baixa) a 5 (melhor avaliação). Então executamos os seguintes passos:

- Para os itens ímpares (Q1,Q3,Q5,Q7,Q9), subtraia 5 do total da pontuação obtida dos itens ímpares.
- Já para os itens pares (Q2,Q4,Q6,Q8,Q19), subtraia a pontuação obtida dos itens pares de 25.

System Usability Scale Questionnaire	Strongly Disagree				Strongly Agree
1. I think that I would like to use this product frequently.	1	2	3	4	5
2. I found the product unnecessarily complex.	1	2	3	4	5
3. I thought the product was easy to use.	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this product.	1	2	3	4	5
5. I found the various functions in the product were well integrated.	1	2	3	4	5
6. I thought there was too much inconsistency in this product.	1	2	3	4	5
7. I imagine that most people would learn to use this product very quickly.	1	2	3	4	5
8. I found the product very awkward to use.	1	2	3	4	5
9. I felt very confident using the product.	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this product.	1	2	3	4	5

Figura 54 – Questionário padrão do SUS. Fonte: link.

– Por fim, calcule a soma dos valores dos itens acima e multiplique por 2,5 para obter a pontuação padrão do SUS.

A equação a seguir mostra uma maneira mais concisa de calcular uma pontuação SUS a partir dos valores brutos dos 10 itens:

$$SUS = 2.5 * [20 + [(Q1 + Q3 + Q5 + Q7 + Q9) - (Q2 + Q4 + Q6 + Q8 + Q10)]] \tag{5.1}$$

Vale destacar ainda que o sistema de pontuação do SUS exige que todos os 10 itens sejam respondidos. Portanto, se um entrevistado deixar um item em branco, ele deve receber uma pontuação bruta de 3 (o centro da escala de cinco pontos). Além disso, o resultado do SUS pode ser interpretado da seguinte maneira (BANGOR *et al.*, 2009):

– Valores menores que 25 são considerados "Muito ruins";
– Valores entre 25 e 38 são considerados "Pobres";
– Valores entre 38 e 52 são considerados "OK";
– Valores entre 52 e 72 são considerados "Bons";

- Valores entre 72 e 85 são considerados "Excelentes";
- Valores acima de 85 são considerados "Acima das expectativas";

5.2 Configurações da avaliação de usabilidade

A avaliação da usabilidade da ferramenta *DSAdvisor* realizada nesta dissertação incluiu dos tipos (perfis) distintos de usuários: i) pessoas experientes na área da ciências de dados e ii) iniciantes na área de ciência de dados (pessoas com conhecimento superficial ou sem nenhum conhecimento na referida área). Os testes de usabilidade foram realizados de forma remota. Portanto, os participantes realizavam as operações na ferramenta *DSAdvisor* remotamente, uma vez que esta é uma aplicação Web. A ferramenta *DSAdvisor* foi disponibilizada por meio de uma instância EC2 da *Amazon*. Assim, o participante recebia um *link* e por meio deste acessava a *DSAdvisor*, sem a necessidade de instalar qualquer tipo de *software*.

Por fim, além de executarem todas as funcionalidades da *DSAdvisor*, os participantes preencheram, antes de acessar a ferramenta, um formulário demográfico para mapeamento dos seus respectivos perfis, um para marcar a entrevista e no fim um outro formulário após completarem as tarefas, incluindo as questões dos métodos NPS e SUS, além de algumas perguntas adicionais sobre aspectos considerados relevantes. Os formulários seguem na seção de anexos **??**.

5.2.1 População

A população da avaliação consistiu de 20 pessoas de diversas áreas profissionais, sendo que 10 possuíam um perfil de (i) pessoas experientes na área da ciências de dados e 10 de (ii) iniciantes na área de ciência de dados. A faixa etária dos participantes variou entre 20 e 40 anos como mostra a Figura 76.

5.2.2 Entrevistas para avaliação de usabilidade

Para a realização das avaliações de usabilidade foram previamente agendadas 20 entrevistas, uma com cada participante. Os horários das entrevistas foram definidos pelos participantes no formulário de perfil demográfico. Esse formulário coletou dados como idade, gênero, experiência do usuário (através da atribuição de uma nota de zero a dez), experiência na área de ciência de dados, se já utilizou alguma ferramenta ou linguagem de programação com

foco em ciência de dados, e, por fim, o entrevistado informava a data e hora disponíveis para a entrevista.

Na entrevista, o usuário acessava a ferramenta *DSAdvisor* por meio de um *link* de acesso a uma instância EC2 da *Amazon*, porém hoje não se encontra mais disponível e para acesso da ferramenta basta acessar o *link* do git[1]. Junto com este *link*[2] foi enviado um arquivo ".csv"[3] contendo o conjunto de dados que deveria ser utilizado para realizar a tarefa proposta de utilizar a *DSAdvisor*em tarefas preditivas na avaliação, permitindo ao participante explorar todas as funcionalidades da *DSAdvisor*.

Após tentar realizar as tarefas propostas utilizando o conjunto de dados fornecido e a ferramenta *DSAdvisor*, cada participante respondeu o segundo questionário, contendo as questões do NPS e do SUS. Adicionalmente, o questionário incluiu questões sobre as seguintes funcionalidades:

- "Bestfit";
- "Outlier Detection";
- Funcionamento por meio da Web
- Armazenamento de todas as decisões tomadas pelo usuário em arquivo de texto.

[1] https://gitlab.com/jmmonteiro/dsadvisor
[2] https://forms.gle/1pB39E52Y3668SkPA
[3] shorturl.at/iN056

6 RESULTADOS

Neste capítulo, iremos discutir os resultados obtidos na avaliação de usabilidade da ferramenta *DSAdvisor*, mais precisamente nos testes *Net Promoter Score* (NPS) e *System Usability Scale* (SUS). Adicionalmente, iremos abordar as respostas das questões relacionadas às principais funcionalidades da *DSAdvisor*.

6.1 Resultados dos testes de usabilidades

Nesta seção, iremos apresentar e discutir os resultados dos testes *Net Promoter Score* (NPS) e *System Usability Scale* (SUS). Por questões didáticas, iremos apresentar os resultados em três cenários distintos:

- pessoas experientes na área da ciências de dado (chamados de usuários especialistas);
- iniciantes na área de ciência de dados, ou seja, pessoas com conhecimento superficial ou sem nenhum conhecimento na referida área (chamados usuários não especialistas) e
- todos os participantes.

Com esta organização, será possível analisar a usabilidade da ferramenta para diferentes públicos alvo (ou perfis).

6.1.1 Resultados do NPS

Como mencionado no capítulo anterior, a fórmula de cálculo do NPS é dada pela subtração da porcentagem dos Promotores menos a porcentagem dos Detratores. Valores de NPS acima de 70 são considerados "excelentes". Já os valores entre 50 e 70 são considerados "muito bons". Por fim, valores entre 0 e 50 são considerados "bons"(LEE, 2018). A seguir, iremos detalhar os resultados obtidos para cada um dos três cenários definidos anteriormente.

6.1.1.1 NPS para usuários não especialistas

A Figura 55 mostra a relação de usuários não especialistas com suas respectivas notas. Note que, dos dez participantes, seis atribuíram notas entre 9 e 10 (sendo classificados promotores), dois atribuíram nota 8 (sendo classificados como passivos ou neutros) e dois participantes atribuíram notas menores ou iguais a 6 (sendo classificados como detratores). Porém, as notas atribuídas pelos detratores, respectivamente de 6 e 5, estão próximas do limite

inferior da classificação dos usuários passivos, ou seja, 7. O teste do NPS teve com resultado o valor de 40%. Esse resultado é classificado como "bom".

Votos	Classificação
9	PROMOTORES
10	PROMOTORES
10	PROMOTORES
9	PROMOTORES
8	NEUTROS
6	DETRATORES
10	PROMOTORES
8	NEUTROS
5	DETRATORES
10	PROMOTORES
Resultados 40%	

Figura 55 – NPS aplicado aos usuários não especialistas. Fonte: Autor.

6.1.1.2 NPS para usuários especialistas

A Figura 56 mostra a relação de usuários especialistas com suas respectivas notas. O teste do NPS teve com resultado o valor de 70%. Esse resultado é classificado como “muito bom”. Esse melhor resultado decorre do fato de que para usuários especialistas muitas das funcionalidades fornecidas pela *DSAdvisor* já eram conhecidas, seja de forma teórica, seja por meio da utilização de outras ferramentas, o que melhora a percepção do usuário e leva a notas maiores.

Votos	Classificação
9	PROMOTORES
2	DETRATORES
9	PROMOTORES
8	NEUTRO
9	PROMOTORES
9	PROMOTORES
9	PROMOTORES
10	PROMOTORES
9	PROMOTORES
9	PROMOTORES
Resultado:	70,00%

Figura 56 – NPS aplicado aos usuários especialistas. Fonte: Autor.

6.1.1.3 NPS com todos os usuários

A Figura 57 mostra a relação dos usuários (incluindo especialistas e não especialistas) com suas respectivas notas. O teste do NPS teve com resultado o valor de 55%. Esse resultado é classificado como “muito bom”.

Votos	Classificação
9	PROMOTORES
9	PROMOTORES
10	PROMOTORES
10	PROMOTORES
10	PROMOTORES
9	PROMOTORES
9	PROMOTORES
9	PROMOTORES
9	PROMOTORES
6	DETRATORES
10	PROMOTORES
9	PROMOTORES
9	PROMOTORES
8	NEUTROS
2	DETRATORES
8	NEUTROS
5	DETRATORES
9	PROMOTORES
10	PROMOTORES
8	NEUTROS
Resultado: 55,00%	

Figura 57 – NPS aplicado a ambos os perfis de usuários. Fonte: Autor.

6.1.2 Resultados do SUS

Como mencionado no capítulo anterior, para calcular o valor do SUS, primeiramente, convertemos as pontuações brutas dos 10 itens em pontuações ajustadas, o que é realizado da seguinte forma. Para os itens ímpares, subtraímos 1 da pontuação bruta e, para os itens pares, subtraímos a pontuação bruta de 5. Em seguida, calculamos a soma das pontuações ajustadas e multiplicamos por 2,5, obtendo assim a pontuação do SUS. A seguir, iremos detalhar os resultados obtidos no teste SUS para cada um dos três cenários definidos anteriormente.

6.1.2.1 *SUS para usuários não especialistas*

A Figura 58 mostra a relação de usuários não especialistas com suas respectivas notas. O resultado obtido pelo método SUS foi de 66,75, o que se configura como “Ok”, pela classificação mencionada anteriormente (valores entre 52 e 72). Porém, vale observar que o resultado obtido está próximo do limite inferior dos valores considerados “bons” (valores entre 72 e 85).

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SUS Raw Score	SUS Final Score
5	2	4	3	5	2	4	2	4	3	30	75
4	2	4	3	5	2	5	2	4	1	32	80
4	2	4	4	5	3	4	2	4	3	27	67,5
4	2	4	5	5	4	4	2	3	2	25	62,5
4	2	3	4	4	3	4	3	3	5	21	52,5
3	1	4	4	3	3	4	2	2	3	23	57,5
5	1	3	4	4	1	4	1	4	4	29	72,5
3	3	4	2	4	2	3	2	4	1	28	70
3	2	4	4	4	2	2	2	3	2	24	60
4	2	4	5	5	1	4	1	4	4	28	70
										AVG=	66,75

Figura 58 – SUS aplicado aos usuários não especialistas. Fonte: Autor.

6.1.2.2 *SUS para usuários especialistas*

A Figura 59 ilustra a relação de usuários especialistas com suas respectivas notas. O resultado obtido pelo método SUS foi de 70,25, o que se configura como “Ok”, pela classificação mencionada anteriormente (valores entre 52 e 72). Porém, vale observar que o resultado obtido está próximo do limite inferior dos valores considerados “bons” (valores entre 72 e 85). Adicionalmente, podemos notar que o resultado obtido pelo método SUS entre os usuários especialistas (70,25) foi maior que o resultado obtido entre os usuários não especialistas (66,75).

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SUS Raw Score	SUS Final Score
5	3	3	4	5	2	3	2	4	4	25	62,5
2	5	2	5	4	2	1	4	1	4	10	25
3	4	5	4	1	1	4	1	4	4	23	57,5
3	2	4	1	4	2	4	2	3	1	30	75
4	2	4	4	4	2	4	2	4	2	28	70
5	2	4	3	4	1	5	2	4	1	33	82,5
5	1	4	2	4	2	4	2	4	2	32	80
4	1	5	2	5	2	4	1	5	1	36	90
4	4	5	1	5	1	5	1	5	1	36	90
4	3	4	2	4	1	4	1	3	4	28	70
										AVG=	70,25

Figura 59 – SUS aplicado aos usuários especialistas. Fonte: Autor.

6.1.2.3 SUS com ambos os usuários

A Figura 60 mostra a relação de todos os usuários com suas respectivas notas. O resultado obtido pelo método SUS foi de 68,5, o que se configura como "Ok", pela classificação mencionada anteriormente (valores entre 52 e 72). Porém, vale observar que o resultado obtido está próximo do limite inferior dos valores considerados "bons"(valores entre 72 e 85). Esse resultado indica que, público geral, a ferramenta *DSAdvisor* poderia ser útil.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	SUS Raw Score	SUS Final Score
5	2	4	3	5	2	4	2	4	3	30	75
5	1	4	2	4	2	4	2	4	2	32	80
4	1	5	2	5	2	4	1	5	1	36	90
4	2	4	3	5	2	5	2	4	1	32	80
4	2	4	4	5	3	4	2	4	3	27	67,5
4	2	4	4	4	2	4	2	4	2	28	70
4	2	4	5	5	4	4	2	3	2	25	62,5
5	2	4	3	4	1	5	2	4	1	33	82,5
4	3	4	2	4	1	4	1	3	4	28	70
3	1	4	4	3	3	4	2	2	3	23	57,5
5	1	3	4	4	1	4	1	4	4	29	72,5
4	4	5	1	5	1	5	1	5	1	36	90
3	4	5	4	1	1	4	1	4	4	23	57,5
4	2	3	4	4	3	4	3	3	5	21	52,5
2	5	2	5	4	2	1	4	1	4	10	25
3	3	4	2	4	2	3	2	4	1	28	70
3	2	4	4	4	2	2	2	3	2	24	60
5	3	3	4	5	2	3	2	4	4	25	62,5
4	2	4	5	5	1	4	1	4	4	28	70
3	2	4	1	4	2	4	2	3	1	30	75
										AVG=	68,5

Figura 60 – SUS aplicado a ambos os perfis. Fonte: Autor.

6.2 Avaliação de funcionalidades específicas

Após a aplicação dos testes NPS e SUS, os participantes responderam algumas questões sobre as principais funcionalidades da ferramenta:

- "BestFit'"
- "Outlier Detection"
- Funcionamento por meio da Web
- Armazenamento de todas as decisões tomadas pelo usuário em arquivo de texto.

Para essas questões, os participantes atribuíram notas entre 1 e 5. A seguir, iremos detalhar os resultados obtidos para cada uma dessas funcionalidades.

6.2.1 Avaliação da funcionalidade de Bestfit

A funcionalidade de "Bestfit"é uma heurística para aproximar a distribuição do conjunto de dados inseridos. Ela serve para que o usuário possa verificar a distribuição do dado, e se aquele dado segue uma distribuição que o próprio usuário julgue que ela siga.

6.2.1.1 Avaliação da funcionalidade Bestfit com usuários não especialistas

A Figura 61 exibe a relação de usuários não especialistas com suas respectivas notas. Podemos observar que 70% dos participantes atribuíram notas entre 4 e 5. Porém, 30% dos participantes atribuíram nota 3, o que pode indicar que eles não compreenderam a importância dessa funcionalidade, o que seria esperado em se tratando de usuários sem conhecimento em Ciência de Dados.

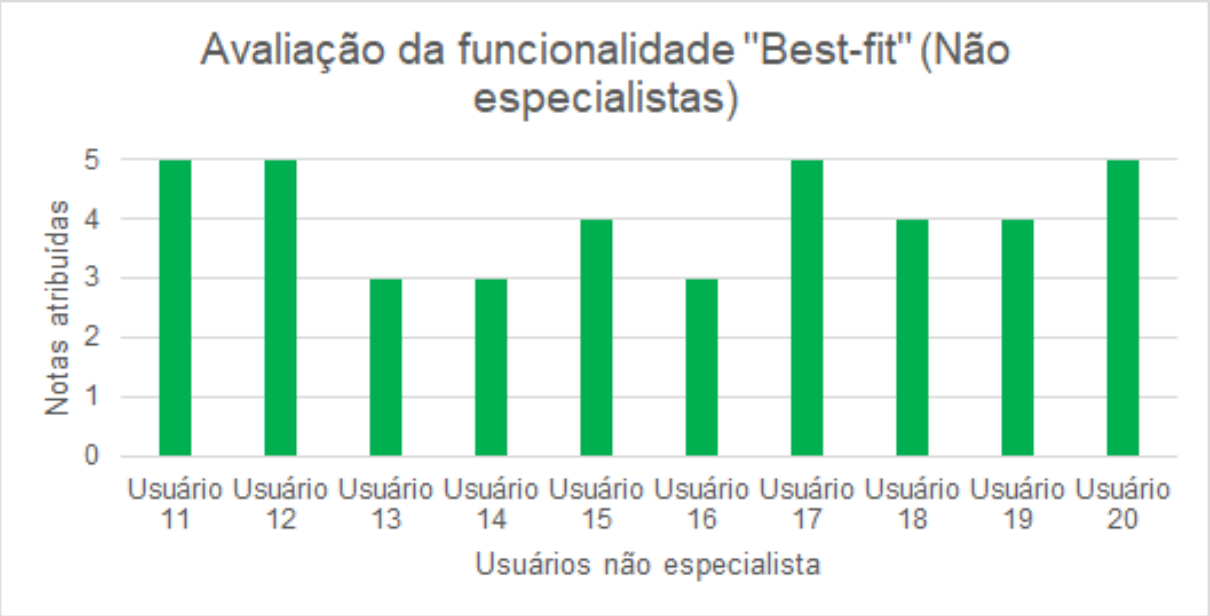


Figura 61 – Avaliação da funcionalidade "Bestfit"com usuários não especialistas. Fonte: Autor.

6.2.1.2 Avaliação da funcionalidade "Bestfit"com usuários especialistas

A Figura 62 exibe a relação de usuários especialistas com suas respectivas notas. Podemos observar que 80% dos participantes atribuíram nota 5 e 20% nota 4. Esse resultado indica que os usuários especialistas consideraram essa funcionalidade bastante relevante.

Figura 62 – Avaliação da funcionalidade "Bestfit"com usuários especialistas. Fonte: Autor.

6.2.1.3 Avaliação da funcionalidade "Bestfit"com todos os usuários

A Figura 63 exibe a relação de usuários com suas respectivas notas. Podemos observar que 60% dos participantes atribuíram nota 5. Vale destacar também que os usuários especialistas atribuíram notas mais altas que os usuários não especialistas. Isso indica que, de fato, existe uma exigência de conhecimentos prévios acerca de distribuições de probabilidades e sua implicação nas tarefas relacionadas ao pre-processamento dos dados.

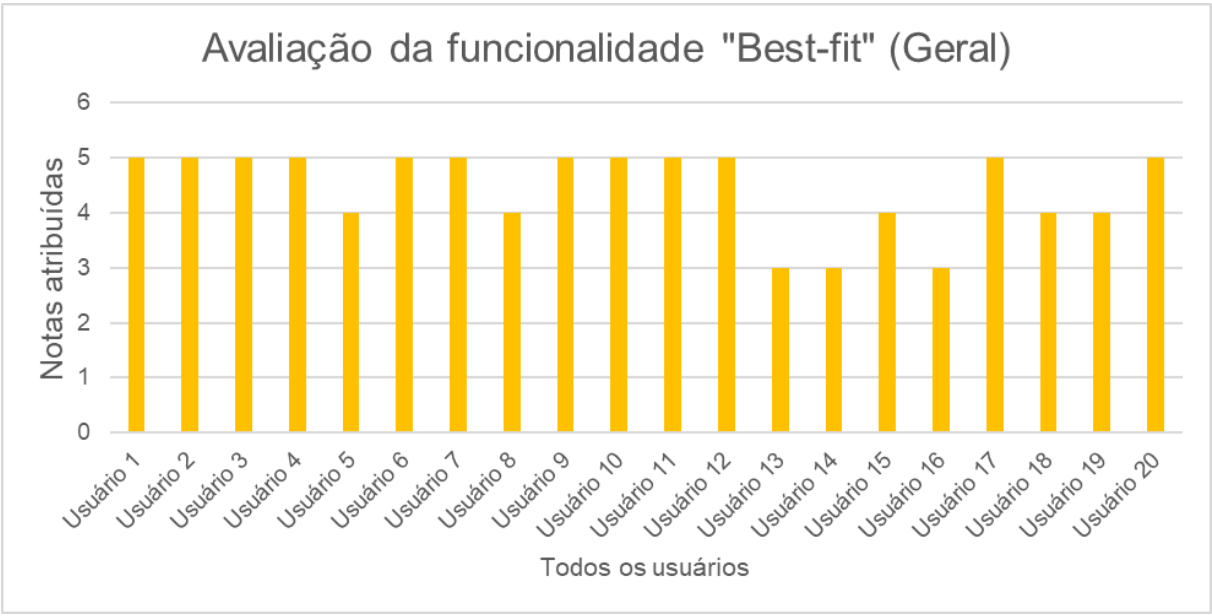


Figura 63 – Avaliação da funcionalidade "Bestfit"com ambos usuários. Fonte: Autor.

6.2.2 Avaliação da funcionalidade de Outlier Detection

Outra funcionalidade importante presente na ferramenta *DSAdvisor* é a detecção de valores discrepantes, chamada de "*Outlier Detection*". Essa funcionalidade tem por finalidade percorrer o conjunto de dados fornecido pelo usuário e detectar valores discrepantes ou anômalos. Adicionalmente, a ferramenta indica a posição onde esses valores foram encontrados no conjunto de dados para que usuário possa investigar se a origem destes valores se deve a algum erro na captura dos dados ou não. Por fim, a ferramenta auxilia na remoção dos valores discrepantes, caso seja necessário.

6.2.2.1 Avaliação da funcionalidade "Outlier Detection"com usuários não especialistas

A Figura 64 exibe a relação de usuários não especialistas com suas respectivas notas. Podemos observar que 80% dos participantes atribuíram notas entre 4 e 5. Porém, 20% dos participantes atribuíram nota 3, o que pode indicar que eles não compreenderam a importância dessa funcionalidade, o que seria esperado em se tratando de usuários sem conhecimento em Ciência de Dados.

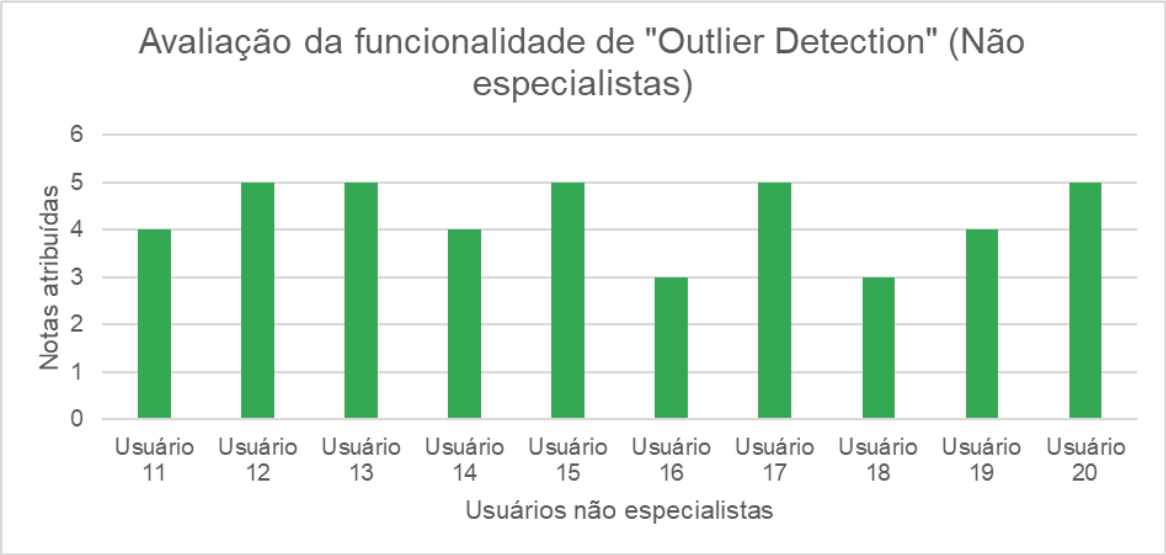


Figura 64 – Avaliação da funcionalidade Outlier Detection com usuários não especialistas.
Fonte: Autor.

6.2.2.2 Avaliação da funcionalidade "Outlier Detection"com usuários especialistas

A Figura 65 exibe a relação de usuários especialistas com suas respectivas notas. Podemos observar que 80% dos participantes atribuíram notas entre 5 e 20% dos participantes atribuíram nota 4. Esse resultado indica que os usuários especialistas consideraram essa funcionalidade bastante relevante.

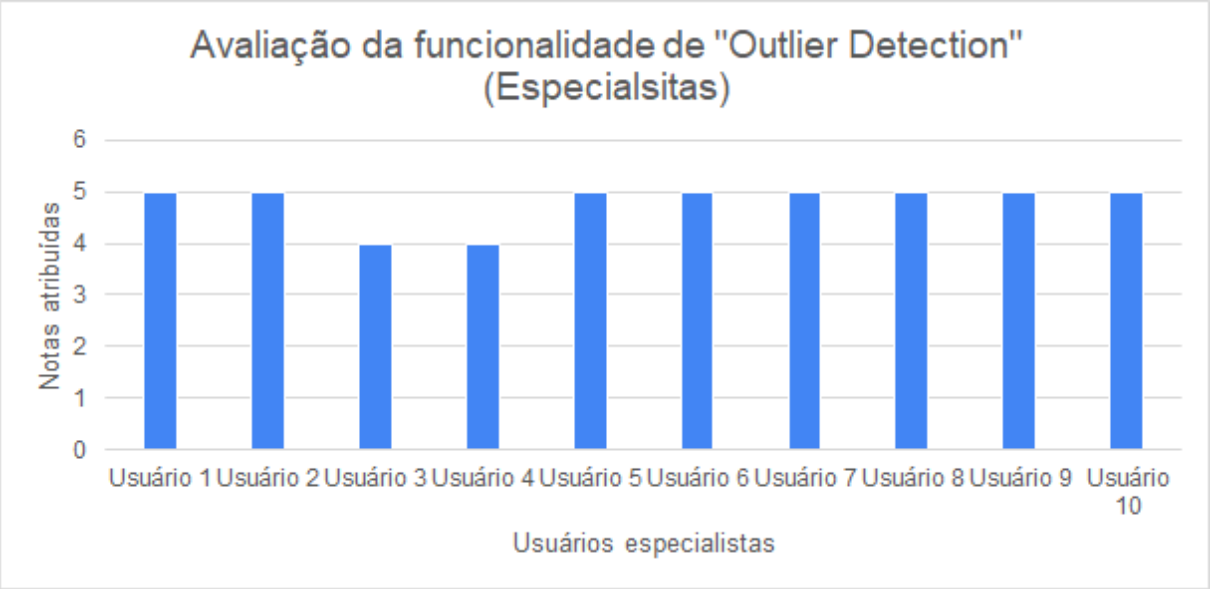


Figura 65 – Avaliação da funcionalidade Outlier Detection com usuários especialistas. Fonte: Autor.

6.2.2.3 Avaliação da funcionalidade "Outlier Detection"com todos os usuários

A Figura 66 exibe a relação de usuários com suas respectivas notas. Podemos observar que 90% dos participantes atribuíram notas entre 4 e 5. Vale destacar também que os usuários especialistas atribuíram notas mais altas que os usuários não especialistas. Isso indica que, de fato, existe uma exigência de conhecimentos prévios acerca do conceito, da importância e dos métodos de detecção de valores discrepantes.

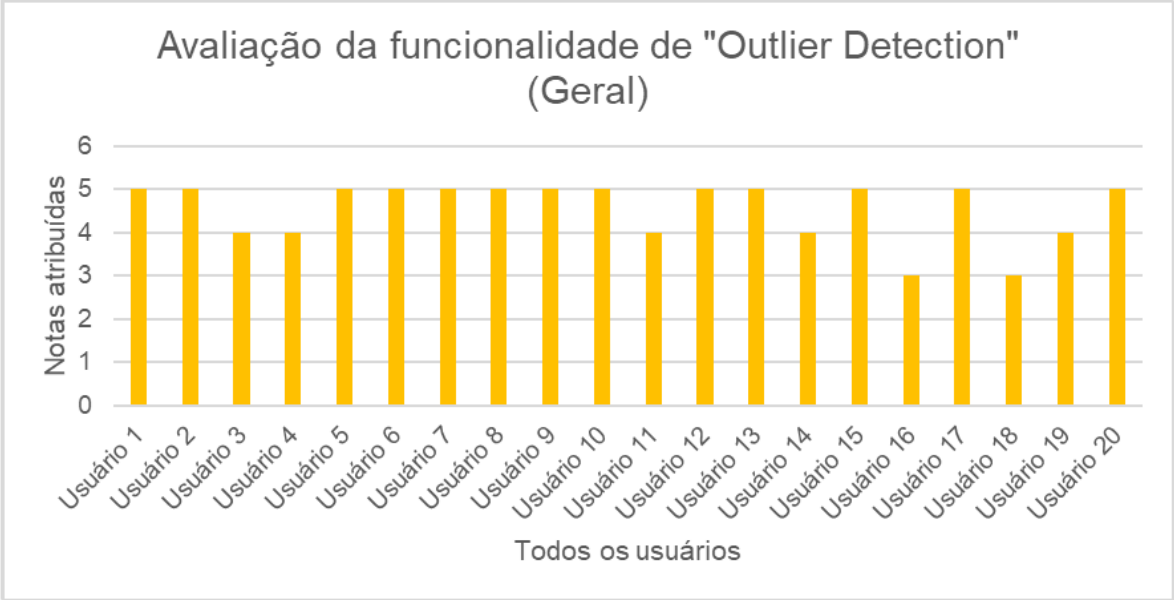


Figura 66 – Avaliação da funcionalidade "Outlier Detection"com ambos os tipos de usuários. Fonte: Autor.

6.2.3 Avaliação da funcionalidade de garantir a reprodutibilidade ("Ensure Reproducibility")

Por fim, a última funcionalidade avaliada pelos usuários foi a garantia da reprodutibilidade, chamada de "*Ensure Reproducibility*". Essa funcionalidade permite que o usuário armazene em um arquivo .txt todas as decisões por ele tomadas durante a utilização da *DSAdvisor*.

As Figuras 67, 68 e 69 exibem um resumo dos resultados obtidos para essa funcionalidade. Observe que nos três cenários foram obtidos resultados semelhantes.

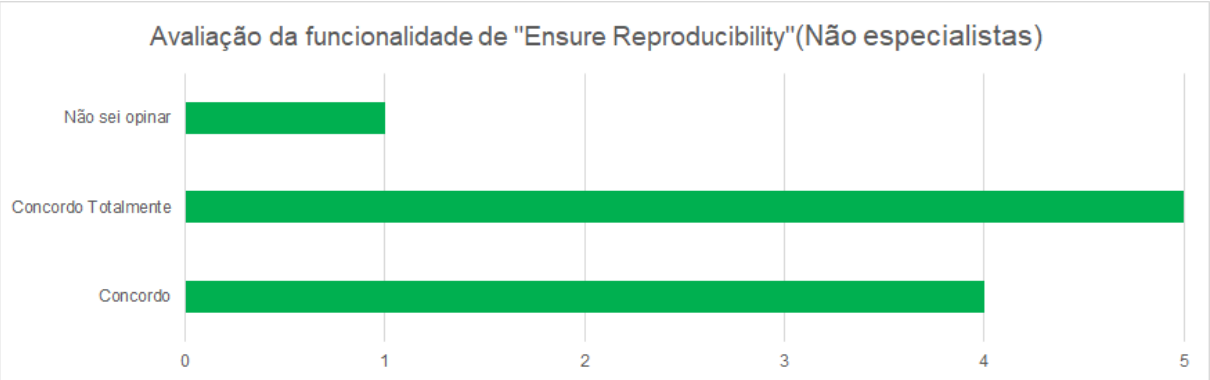


Figura 67 – Avaliação da funcionalidade “Ensure Reproducibility” com usuários não especialistas. Fonte: Autor.

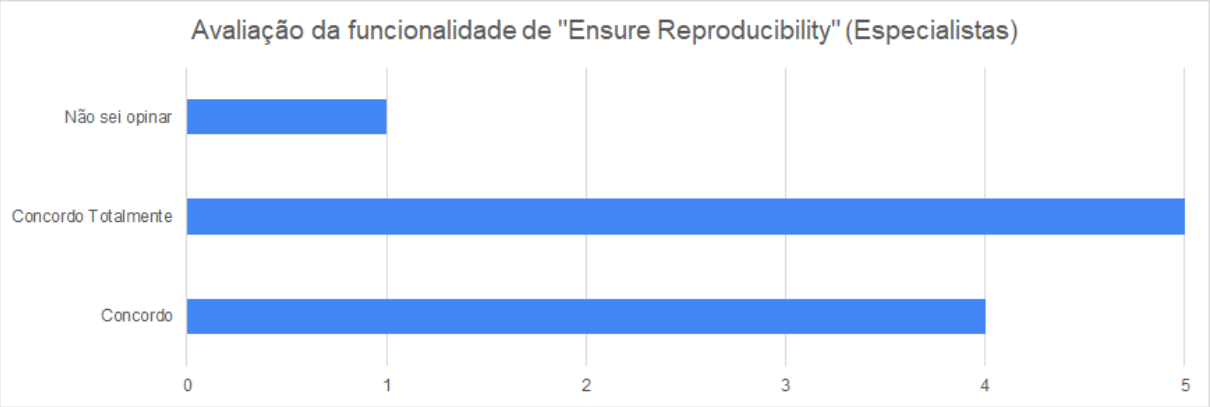


Figura 68 – Avaliação da funcionalidade “Ensure Reproducibility” com usuários especialistas. Fonte: Autor.

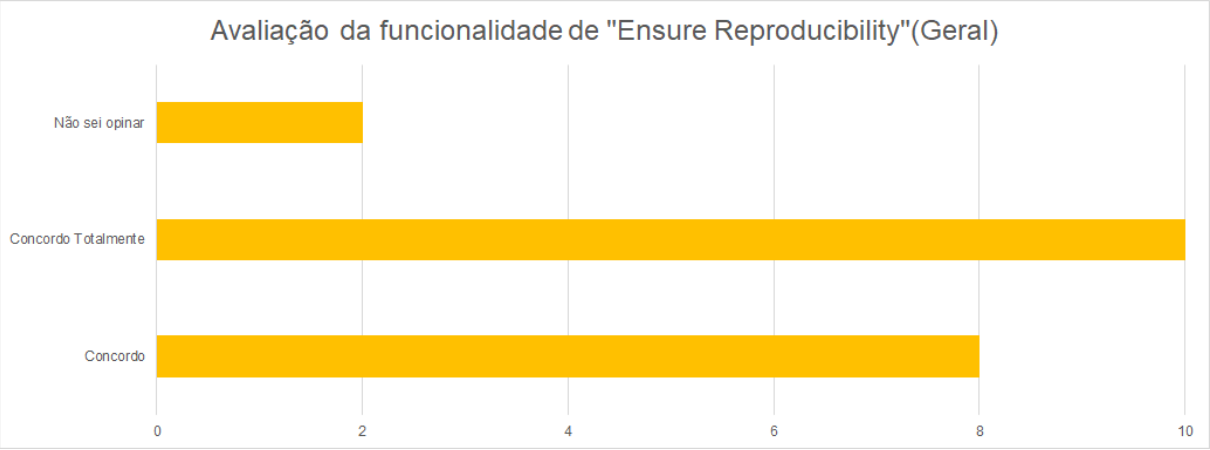


Figura 69 – Avaliação da funcionalidade “Ensure reproducibility” com todos os usuários. Fonte: Autor.

6.3 Avaliação geral

Por fim, os participantes realizaram uma avaliação qualitativa, destacando os pontos negativos e positivos da ferramenta *DSAdvisor*. As respostas dos participantes estão descritas na íntegra no Anexo A. Os comentários elaborados pelos participantes permitiram obter *feedbacks* importantes.

Como pontos positivos temos que a maioria dos usuários, tanto especialistas como leigos, relataram facilidade no manuseio da ferramenta. Um comentário relevante realizado por um participante diz respeito à diversidade de opções disponíveis ara escolha em algumas funcionalidades da ferramenta *DSAdvisor*, como, por exemplo, a quantidade de distribuições de probabilidade que podem ser selecionadas.

Já os pontos negativos encontrados pelos usuários leigos foram de instabilidades

na aplicação que ocasionavam atrasos no carregamento das páginas; trechos da ferramenta se demonstraram complexos para o entendimento de usuários leigos, havendo assim a necessidade de reavaliação de como explicar certas etapas da aplicação ou até mesmo alguns casos de que não estava claro a escolha que o usuário precisava fazer em determinada página da ferramenta, por exemplo.

Para os usuários especialista um dos pontos negativos encontrados foi de que a aplicação detém conteúdo textual excessivo para algumas páginas da aplicação. Em vista desse cenário apontado, os próprios entrevistados recomendaram que houve-se uma reformulação nas páginas da aplicação visando reduzir o conteúdo textual por outras alternativas. Outro ponto negativo citado é a escassez de *links* de conteúdo recomendado de algumas páginas. Para alguns usuários materiais de apoio poderiam ajudar a enriquecer a experiência ao se utilizar a ferramenta.

Por fim, no último ponto das entrevista foi de avaliação geral, onde os usuários em sua maioria afirmaram que a ferramenta *DSAdvisor* atende ao propósito de auxiliar nas atividades preditivas, bem como facilita a experiência do usuário final através de uma interface simples, direta e funcional.

Ademais, alguns participantes destacaram que a *DSAdvisor* poderia ser utilizada nas aulas prática das disciplinas que envolvam mineração dos dados. Por outro lado, alguns usuários leigos relataram uma frustração devido a presença de conceitos específicos da estatística, por eles desconhecidos.

7 CONCLUSÕES E TRABALHOS FUTUROS

Nesta dissertação, propomos um guia prático que tem por finalidade auxiliar profissionais de diferentes áreas do conhecimento nas diversas atividades envolvidas na solução de problemas de predição, mais especificamente, regressão e classificação. O guia proposto nesta dissertação é organizado em três fases, sendo elas: análise exploratória, pré-processamento dos dados e criação de modelos preditivos. Cada fase é composta por uma sequência de tarefas. Para cada tarefa, o guia descreve sua finalidade, justificativa teórica, orientações práticas e artefatos utilizados como entrada ou produzidos como saída. A primeira fase do guia proposto, denominada análise exploratória, tem por finalidade explorar os dados que serão utilizados na construção de um ou mais modelos preditivos. O principal objetivo desta fase consiste em entender, descrever e resumir os dados que serão utilizados. A segunda-fase, chamada pré-processamento dos dados, tem por objetivo preparar os dados para que estes possam ser utilizados na construção de modelos preditivos. Esta fase inclui atividades relacionadas à detecção de valores discrepantes, normalização de dados, escolha da variável independente, seleção de atributos, balanceamento de dados, seleção de atributos (variáveis) e divisão dos conjuntos de treinamento e teste. A terceira e última fase do guia proposto, denominada criação de modelos preditivos, tem por finalidade gerar modelos preditivos, analisar seus resultados e assegurar que os experimentos realizados possam ser reproduzidos.

Adicionalmente, seguindo o guia prático proposto, desenvolvemos uma ferramenta, denominada *DSAdvisor*, busca auxiliar os usuários na execução das diversas atividades envolvidas em um problema de predição. A *DSAdvisor* visa encorajar usuários leigos a construir modelos de aprendizado de máquina para executar tarefas preditivas, extraindo conhecimento de seus próprios repositórios de dados. A *DSAdvisor* atua como um consultor para usuários não especialistas, descrevendo para cada atividade a sua finalidade, os conceitos teóricos nela envolvidos e um conjunto de dicas para a sua execução. Além disso, a *DSAdvisor* segue um fluxo bem definido, faz a gestão de todos os artefatos utilizados (com entrada ou saída) e de todas as configurações realizadas pelo usuário, facilitando assim a reprodutibilidade dos experimentos realizados. A *DSAdvisor* foi desenvolvida em *Python* (ROSSUM, 1995) utilizando bibliotecas como *Flask* (GRINBERG, 2018), *scikit-learn*, *seaborn*, *matplotlib*, *seaborn*, *scipy*, *numpy*, *pandas*, dentre outras. Para mais, a *DSAdvisor* está disponível em um repositório *online* e pode ser usada livremente [1].

[1] A código da ferramenta *DSAdvisor* pode ser obtido livremente por meio do *link*

Por fim, avaliamos a ferramenta *DSAdvisor* utilizando dois diferentes testes de usabilidade e satisfação do usuário: *Net Promoter Score* (NPS) e *System Usability Scale* (SUS). O *Net Promoter Scores* (NPS) é uma técnica que tem por finalidade medir a satisfação de um cliente ou usuário para que as empresas ou prestadoras de serviço possam avaliar e melhorar seu desempenho, seus produtos e serviços (RAS *et al.*, 2017). A ideia principal do NPS é que os clientes ou usuários sejam abordados quanto à probabilidade de recomendarem produtos/serviços para seus colegas (KORNETA, 2014). Já o *System Usability Scale* (SUS) é um questionário padronizado amplamente utilizado para a avaliação da usabilidade percebida. O método SUS foi utilizado em 43% das avaliações de usabilidade industrial (LEWIS; SAURO, 2009). As citações do *Google Scholar* (examinadas em 25/08/2022) mostraram 13539 citações para o artigo que introduziu o método SUS, mais precisamente, (BROOKE, 1996).

A avaliação da usabilidade da ferramenta *DSAdvisor* incluiu dos tipos (perfis) distintos de usuários: i) pessoas experientes na área da ciências de dados e ii) iniciantes na área de ciência de dados (pessoas com conhecimento superficial ou sem nenhum conhecimento na referida área). Os testes de usabilidade foram realizados de forma remota, uma vez que a *DSAdvisor* é uma aplicação Web, por meio de uma instância EC2 da *Amazon*.

Assim, o participante recebia um *link* e por meio deste acessava a *DSAdvisor*, sem a necessidade de instalar qualquer tipo de *software*. Por fim, além de executarem todas as funcionalidades da *DSAdvisor*, os participantes preencheram, antes de acessar a ferramenta, um formulário demográfico para mapeamento dos seus respectivos perfis, um para marcar a entrevista e no fim um outro formulário após completarem as tarefas, incluindo as questões dos métodos NPS e SUS, além de algumas perguntas adicionais sobre aspectos considerados relevantes.

A população da avaliação consistiu de 20 pessoas de diversas áreas profissionais, sendo que 10 possuíam um perfil de (i) pessoas experientes na área da ciências de dados e 10 de (ii) iniciantes na área de ciência de dados. A faixa etária dos participantes variou entre 20 e 40 anos. O teste NPS realizado com os usuários iniciantes (não especialistas) resultou no valor de 40%, o que é classificado como "bom". Já para os usuários especialistas o teste NPS resultou no valor de 70%, o que é classificado como "muito bom". Considerando os dois perfis em conjunto, o resultado do teste NPS foi de 55%, o que é classificado como "muito bom". O resultado obtido pelo método SUS para usuários não especialistas foi de 66,75, o que pode ser classificado como

https://gitlab.com/jmmonteiro/dsadvisor

"Bom"(valores entre 52 e 72).

O resultado obtido pelo método SUS para usuários especialistas foi de 70,25, o que também pode ser classificado como "Bom", já o resultado obtido entre os usuários não especialistas foi de 66,75. Considerando os dois perfis em conjunto, o resultado do teste SUS foi de 68,5, o que se configura como "Bom". Esse resultado indica que, público geral, a ferramenta *DSAdvisor* poderia ser útil.

7.1 Ameaças à Validade

Mesmo com os devidos cuidados para a realização da avaliação da usabilidade da ferramenta *DSAdvisor*, essa análise ainda pode ser afetada por diferentes fatores que podem invalidar as conclusões obtidas.

Validade Interna. Visando aumentar a validade interna, foram usados os formulários do NPS e do SUS com suas traduções literais com poucas alterações e substituindo apenas a referência do foco da pesquisa para a ferramenta *DSAdvisor*. Outro fator relevante é que a população do experimento constituída por 20 usuários divididos em dois grupos de dez pessoas, sendo um grupo formado apenas por usuários inexperientes e o outro formado por usuários mais experientes. Em (ALROOBAEA; MAYHEW, 2014) foi aferido que com 5 usuários é possível obter cerca de 85% dos problemas de usabilidade de uma ferramenta. Logo, com 10 participantes em cada grupo é possível obter uma avaliação ainda mais consistente.

Validade de Construção. A validade de construção está preocupada com a relação entre a teoria e a observação. Neste contexto, a principal preocupação do estudo é a corretude do guia proposto para as tarefas preditivas, o qual também foi utilizado para balizar o funcionamento da ferramenta *DSAdvisor*. O guia prático proposto se configura como uma extensão de um guia já existente criado por Melo (2020), mas expandindo para problemas de regressão e classificação. A *DSAdvisor* é uma ferramenta em desenvolvimento que possui uma interface direcionada e direta para a tomada de decisões do usuário, onde em todas as páginas existe um conteúdo personalizado para indicar a atividade corrente, bem como sua ligação com as atividades posteriores.

Validade de Conclusão. A validade de conclusão diz respeito à extensão com que as conclusões sobre a presença de uma relação estatisticamente significativa entre os tratamentos e os resultados são válidos. Para mitigar as ameaças à validade da conclusão e aumentar a confiabilidade desse estudo realizamos os cálculos do NPS e do SUS com as suas respectivas tabelas de classificação considerando três cenários distintos: usuários não especialistas, usuários

especialistas e todos os usuários em conjunto.

7.2 Resultados Alcançados

O guia prático para tarefas preditivas foi publicado na conferência *International Conference on Enterprise Information Systems* (ICEIS) de 2021. Este guia pode ser utilizado com o suporte de uma ferramenta ou biblioteca já existente. Adicionalmente, a ferramenta *DSAdvisor*, que segue o fluxo do guia prático proposto, foi publicada apresentada na sessão de ferramentas e demonstrações do Simpósio Brasileiro de Banco de Dados de 2021.

7.3 Trabalhos Futuros

Em razão da *DSAdvisor* ainda se encontrar em uma versão de desenvolvimento é possível encontrar alguns erros durante a execução de certas atividades, tais como: falhas no carregamento do conteúdo de algumas páginas, por exemplo, um determinado gráfico que não é apresentado, imagens com qualidade abaixo do esperado, bem como problemas de acesso concorrente, quando muitos usuários utilizam a ferramenta simultaneamente.

Adicionalmente, uma melhoria sugerida por alguns usuários foi a construção de uma nova interface que possibilite melhor aproveitamento do guia prático por meio de conteúdos em diferentes mídias, tais como vídeos e gifs animados. Essa vertente também viabilizará para a ferramenta ser utilizada em meios educacionais auxiliando em práticas sobre CD.

Com a criação da nova interface para a *DSAdvisor* é necessário uma nova avaliação de usabilidade com os mesmo entrevistados ou comparando a versão anterior com a nova para avaliar as questões apontadas nesse trabalho foram melhor avaliadas ou não. Ademais é importante avaliações comparativas entre as outras ferramentas citadas neste trabalho para descobrir qual a ferramenta preferível pelos usuários em práticas de CD.

Outra melhoria consiste em fornecer a possibilidade do usuário escolher seu próprio fluxo de atividades, com base em sua experiência e necessidade particular de se , por exemplo, usar algoritmos mais sofisticados de *deep learning* onde é necessário adaptar tanto o guia proposto como a ferramenta *DSAdvisor* para dar suporte a análise e processamento dos dados além de também adicionar os algoritmos de *deep learning* que deseja-se utilizar.

Por fim, usar o arquivo gerado pela *DSAdvisor* com as configurações realizadas pelo usuário para reproduzir os experimentos realizados de forma automática foi mencionada como

uma funcionalidade que seria interessante.

REFERÊNCIAS

ALCALÁ, R.; ALCALÁ-FDEZ, J.; CASILLAS, J.; CORDÓN, O.; HERRERA, F. Hybrid learning models to get the interpretability–accuracy trade-off in fuzzy modeling. **Soft Computing**, Springer, v. 10, n. 9, p. 717–734, 2006.

ALROOBAEA, R.; MAYHEW, P. J. How many participants are really enough for usability studies? In: IEEE. **2014 Science and Information Conference**. [*S. l.*], 2014. p. 48–56.

BANGOR, A.; KORTUM, P.; MILLER, J. Determining what individual sus scores mean: Adding an adjective rating scale. **Journal of usability studies**, Citeseer, v. 4, n. 3, p. 114–123, 2009.

BASHEER, I.; HAJMEER, M. Artificial neural networks: fundamentals, computing, design, and application. **Journal of Microbiological Methods**, v. 43, n. 1, p. 3 – 31, 2000. ISSN 0167-7012. Neural Computting in Micrbiology.

BATISTA, G.; PRATI, R.; MONARD, M.-C. A study of the behavior of several methods for balancing machine learning training data. **SIGKDD Explorations**, v. 6, p. 20–29, 06 2004.

BERNADÓ-MANSILLA, E.; HO, T. K. Domain of competence of xcs classifier system in complexity measurement space. **IEEE Transactions on Evolutionary Computation**, IEEE, v. 9, n. 1, p. 82–104, 2005.

BERTHOLD, M. R.; CEBRON, N.; DILL, F.; GABRIEL, T. R.; KÖTTER, T.; MEINL, T.; OHL, P.; THIEL, K.; WISWEDEL, B. Knime-the konstanz information miner: version 2.0 and beyond. **AcM SIGKDD explorations Newsletter**, ACM New York, NY, USA, v. 11, n. 1, p. 26–31, 2009.

BEVAN, N.; CARTER, J.; EARTHY, J.; GEIS, T.; HARKER, S. New iso standards for usability, usability reports and usability measures. In: SPRINGER. **International conference on human-computer interaction**. [*S. l.*], 2016. p. 268–278.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. **Pattern recognition**, Elsevier, v. 30, n. 7, p. 1145–1159, 1997.

BROOKE, J. Sus: a "quick and dirty'usability. **Usability evaluation in industry**, v. 189, n. 3, 1996.

BRYS, G.; HUBERT, M.; STRUYF, A. A robust measure of skewness. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 13, n. 4, p. 996–1017, 2004.

BYTES, E. **Ayrton Senna - Fala sobre o primeiro campeonato e como melhorar como pessoa**. 2020. Disponível em: https://www.youtube.com/watch?v=umhJS2UxyIA.

CAI, J.; LUO, J.; WANG, S.; YANG, S. Feature selection in machine learning: A new perspective. **Neurocomputing**, Elsevier, v. 300, p. 70–79, 2018.

CANO, J. R.; HERRERA, F.; LOZANO, M. Using evolutionary algorithms as instance selection for data reduction in kdd: An experimental study. **IEEE transactions on evolutionary computation**, IEEE, v. 7, n. 6, p. 561–575, 2003.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEA-RER, C.; WIRTH, R. **CRISP-DM 1.0 Step-by-step data mining guide/CRISP-DM consortium. 2000**. [*S. l.*]: Forschungsbericht, 2019.

CHARLTON, S. G.; O'BRIEN, T. G. **Handbook of human factors testing and evaluation**. [*S. l.*]: CRC Press, 2019.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

CHERTCHOM, P. A comparison study between data mining tools over regression methods: Recommendation for smes. In: IEEE. **2018 5th International Conference on Business and Industrial Research (ICBIR)**. [*S. l.*], 2018. p. 46–50.

CLARIDGE, N.; KIRAKOWSKI, J. Wammi: website analysis and measurement inventory questionnaire. **Retrived May**, v. 20, n. 2013, p. 57–66, 2011.

CRAMÉR, H. **Mathematical methods of statistics**. [*S. l.*]: Princeton university press, 1999. v. 43.

D'AGOSTINO, R. B. Transformation to normality of the null distribution of g1. **Biometrika**, JSTOR, p. 679–681, 1970.

DEMŠAR, J.; CURK, T.; ERJAVEC, A.; GORUP, Č.; HOČEVAR, T.; MILUTINOVIČ, M.; MOŽINA, M.; POLAJNAR, M.; TOPLAK, M.; STARIČ, A. *et al.* Orange: data mining toolbox in python. **the Journal of machine Learning research**, JMLR. org, v. 14, n. 1, p. 2349–2353, 2013.

DEMŠAR, J.; ZUPAN, B. Orange: Data mining fruitful and fun. **Informacijska Družba- IS**, v. 6, 2012.

ELHASSAN, T.; ALJURF, M. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. 2016.

FARIAS, A. M. L. de. Estatística descritiva. **Universidade Federal Fluminense. Instituto de Matemática**, 2006.

FERNÁNDEZ, A.; GARCÍA, S.; GALAR, M.; PRATI, R. C.; KRAWCZYK, B.; HERRERA, F. **Learning from imbalanced data sets**. [*S. l.*]: Springer, 2018. v. 10.

GÉRON, A. Hands-on machine learning with scikit-learn and tensorflow: Concepts. **Tools, and Techniques to build intelligent systems**, O'Reilly Media, 2017.

GRAHAM, J. W. *et al.* Missing data analysis: Making it work in the real world. **Annual review of psychology**, Palo Alto, v. 60, n. 1, p. 549–576, 2009.

GRINBERG, M. **Flask web development**. [*S. l.*]: "O'Reilly Media, Inc.", 2018.

GUAN, D.; YUAN, W.; LEE, Y.-K.; LEE, S. Nearest neighbor editing aided by unlabeled data. **Information Sciences**, Elsevier, v. 179, n. 13, p. 2273–2282, 2009.

GUEDES, T. A.; MARTINS, A. B. T.; ACORSI, C. R. L.; JANEIRO, V. Estatística descritiva. **Projeto de ensino aprender fazendo estatística**, Universidade Estadual de Maringá Maringá, p. 1–49, 2005.

GULATI, P. Hybrid resampling technique to tackle the imbalanced classification problem. 2020.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM New York, NY, USA, v. 11, n. 1, p. 10–18, 2009.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**. 3rd edition. ed. [*S. l.*]: Morgan Kaufman, 2012.

HASIM, N.; HARIS, N. A. A study of open-source data mining tools for forecasting. In: **Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication**. [*S. l.*: *s. n.*], 2015. p. 1–4.

HEIJDEN, G. J. Van der; DONDERS, A. R. T.; STIJNEN, T.; MOONS, K. G. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. **Journal of clinical epidemiology**, Elsevier, v. 59, n. 10, p. 1102–1109, 2006.

HELLIWELL, J. F.; HUANG, H.; WANG, S.; NORTON, M. Social environments for world happiness. **World happiness report 2020**, JSTOR, v. 1, p. 13–45, 2020.

HIRAKATA, V. N.; MANCUSO, A. C. B.; CASTRO, S. M. d. J. Teste de hipóteses. **Teste de hipóteses: perguntas que você sempre quis fazer, mas nunca teve coragem. Vol. 39, n. 2, 2019, p. 181-185**, 2019.

HOFMANN, M.; KLINKENBERG, R. **RapidMiner**. [*S. l.*]: CRC Press, 2016.

HUBERT, M.; VANDERVIEREN, E. An adjusted boxplot for skewed distributions. **Computational statistics & data analysis**, Elsevier, v. 52, n. 12, p. 5186–5201, 2008.

HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. **Automated machine learning: methods, systems, challenges**. [*S. l.*]: Springer Nature, 2019.

JAIN, Y. K.; BHANDARE, S. K. Min max normalization based data perturbation method for privacy protection. **International Journal of Computer & Communication Technology**, v. 2, n. 8, p. 45–50, 2011.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [*S. l.*]: Springer, 2013. v. 112.

KANG, H. The prevention and handling of the missing data. **Korean journal of anesthesiology**, The Korean Society of Anesthesiologists, v. 64, n. 5, p. 402–406, 2013.

KIRAKOWSKI, J.; CORBETT, M. Sumi: The software usability measurement inventory. **British journal of educational technology**, Wiley Online Library, v. 24, n. 3, p. 210–212, 1993.

KORNETA, P. What makes customers willing to recommend a retailer-the study on roots of positive net promoter score index. **Central European Review of Economics & Finance**, v. 5, n. 2, p. 61–74, 2014.

KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. *et al.* Handling imbalanced datasets: A review. **GESTS International Transactions on Computer Science and Engineering**, v. 30, n. 1, p. 25–36, 2006.

KUHN, M.; JOHNSON, K. *et al.* **Applied predictive modeling**. [*S. l.*]: Springer, 2013. v. 26.

LAND, S.; FISCHER, S. Rapid miner 5. **Rapid-I GmbH**, 2012.

LEE, S. Net promoter score: Using nps to measure it customer support satisfaction. In: **Proceedings of the 2018 ACM SIGUCCS Annual Conference**. [*S. l.*: *s. n.*], 2018. p. 63–64.

LEWIS, J. R. The system usability scale: past, present, and future. **International Journal of Human–Computer Interaction**, Taylor & Francis, v. 34, n. 7, p. 577–590, 2018.

LEWIS, J. R.; SAURO, J. The factor structure of the system usability scale. In: SPRINGER. **International conference on human centered design**. [*S. l.*], 2009. p. 94–103.

LI, H.; LI, J.; CHANG, P.-C.; SUN, J. Parametric prediction on default risk of chinese listed tourism companies by using random oversampling, isomap, and locally linear embeddings on imbalanced samples. **International Journal of Hospitality Management**, Elsevier, v. 35, p. 141–151, 2013.

LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. Grid search, random search, genetic algorithm: a big comparison for nas. **arXiv preprint arXiv:1912.06059**, 2019.

LILLIEFORS, H. W. On the kolmogorov-smirnov test for normality with mean and variance unknown. **Journal of the American statistical Association**, Taylor & Francis, v. 62, n. 318, p. 399–402, 1967.

LIU, H.; SHAH, S.; JIANG, W. On-line outlier detection and data cleaning. **Computers & Chemical Engineering**, v. 28, n. 9, p. 1635 – 1647, 2004. ISSN 0098-1354. Disponível em: http://www.sciencedirect.com/science/article/pii/S0098135404000249.

LIU, X.-Y.; WU, J.; ZHOU, Z.-H. Exploratory undersampling for class-imbalance learning. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, IEEE, v. 39, n. 2, p. 539–550, 2008.

LONGADGE, R.; DONGRE, S. Class imbalance problem in data mining review. **arXiv preprint arXiv:1305.1707**, 2013.

MANDAL, P. C. Net promoter score: a conceptual analysis. **International Journal of Management Concepts and Philosophy**, Inderscience Publishers, v. 8, n. 4, p. 209–219, 2014.

MARTÍNEZ-ESTUDILLO, A.; MARTÍNEZ-ESTUDILLO, F.; HERVÁS-MARTÍNEZ, C.; GARCÍA-PEDRAJAS, N. Evolutionary product unit based neural networks for regression. **Neural Networks**, Elsevier, v. 19, n. 4, p. 477–486, 2006.

MELO, C. S. Supporting change-prone class prediction. 2020.

MELO, C. S.; CRUZ, M. M. L. da; MARTINS, A. D. F.; MATOS, T.; FILHO, J. M. da S. M.; MACHADO, J. de C. A practical guide to support change-proneness prediction. In: **ICEIS (2)**. [*S. l.*: *s. n.*], 2019. p. 269–276.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. [*S. l.*]: Saraiva Educação SA, 2017.

NIELSEN, J. How to conduct a heuristic evaluation. **Nielsen Norman Group**, v. 1, p. 1–8, 1995.

NIELSEN, J. *et al.* Usability 101: Introduction to usability. 2012.

OLORISADE, B. K.; BRERETON, P.; ANDRAS, P. Reproducibility in machine learning-based studies: An example of text mining. 2017.

OZDEMIR, S. **Principles of data science**. [*S. l.*]: Packt Publishing Ltd, 2016.

PEARSON, K. **Notes on Regression and Inheritance in the Case of Two Parents Proceedings of the Royal Society of London, 58, 240-242**. [*S. l.*]: ed, 1895.

PIATETSKY-SHAPIRO, G. Kdnuggets news on sigkdd service award. 2005.

PONTI, M. A.; COSTA, G. B. P. D. Como funciona o deep learning. **arXiv preprint arXiv:1806.07908**, 2018.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. **Big data**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 1, n. 1, p. 51–59, 2013.

RAMAMOHAN, Y.; VASANTHARAO, K.; CHAKRAVARTI, C. K.; RATNAM, A. *et al.* A study of data mining tools in knowledge discovery process. **International Journal of Soft Computing and Engineering (IJSCE) ISSN**, Citeseer, v. 2, n. 3, p. 2231–2307, 2012.

RAS, Z. W.; TARNOWSKA, K. A.; KUANG, J.; DANIEL, L.; FOWLER, D. User friendly nps-based recommender system for driving business revenue. In: SPRINGER. **International Joint Conference on Rough Sets**. [*S. l.*], 2017. p. 34–48.

ROSSUM, G. van. **Python tutorial**. Amsterdam, 1995.

SCHREPP, M. User experience questionnaire handbook. **All you need to know to apply the UEQ successfully in your project**, 2015.

SELIYA, N.; KHOSHGOFTAAR, T. M.; HULSE, J. V. A study on the relationships of classifier performance metrics. In: **2009 21st IEEE International Conference on Tools with Artificial Intelligence**. [*S. l.*: *s. n.*], 2009. p. 59–66.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

SHARMA, N.; BHANDARI, H. V.; YADAV, N. S.; SHROFF, H. Optimization of ids using filter-based feature selection and machine learning algorithms. **Int. J. Innov. Technol. Explor. Eng**, v. 10, n. 2, p. 96–102, 2020.

SKIENA, S. S. **The data science design manual**. [*S. l.*]: Springer, 2017.

SMIRNOV, N. Table for estimating the goodness of fit of empirical distributions. **The annals of mathematical statistics**, JSTOR, v. 19, n. 2, p. 279–281, 1948.

SPEARMAN, C. The proof and measurement of association between two things. Appleton-Century-Crofts, 1961.

TANG, L.; SONG, J.; BELIN, T. R.; UNÜTZER, J. A comparison of imputation methods in a longitudinal randomized clinical trial. **Statistics in medicine**, Wiley Online Library, v. 24, n. 14, p. 2111–2128, 2005.

VEERABHADRAPPA; RANGARAJAN, L. Bi-level dimensionality reduction methods using feature selection and feature extraction. **International Journal of Computer Applications**, v. 4, 07 2010.

VENKATESH, B.; ANURADHA, J. A review of feature selection and its methods. **Cybernetics and Information Technologies**, Sciendo, v. 19, n. 1, p. 3–26, 2019.

WANG, A. Y.-T.; MURDOCK, R. J.; KAUWE, S. K.; OLIYNYK, A. O.; GURLO, A.; BRGOCH, J.; PERSSON, K. A.; SPARKS, T. D. Machine learning for materials scientists: An introductory guide towards best practices. **Chemistry of Materials**, ACS Publications, 2020.

YANG, Q.; SUH, J.; CHEN, N.-C.; RAMOS, G. Grounding interactive machine learning tool design in how non-experts actually build models. In: **Proceedings of the 2018 Designing Interactive Systems Conference**. [*S. l.*: *s. n.*], 2018. p. 573–584.

YAP, B. W.; RANI, K. A.; RAHMAN, H. A. A.; FONG, S.; KHAIRUDIN, Z.; ABDULLAH, N. N. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: SPRINGER. **Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)**. [*S. l.*], 2014. p. 13–22.

ANEXO A – AVALIAÇÃO QUALITATIVA

	Quais os pontos POSITIVOS sobre a funcionalidade de realizar tarefas preditivas.
Usuário 11	Fácil escolha dos algoritmos
Usuário 12	O sistema dispõe de uma facilidade muito grande para realizar tarefas preditivas.
Usuário 13	A diversidade de opções disponibilizada
Usuário 14	Automatizar o processo e executar várias análises de uma única vez.
Usuário 15	Estimar possíveis resultados pode ajudar a preparar como melhor conduzir uma tarefa.
Usuário 16	Facilitar a análise dos dados
Usuário 17	Acredito que as tarefas preditivas auxiliam na preparação e análise dos dados. Acredito também que, analisando especificamente o DSAdvisor, as fases de validação proporcionam um melhor feedback ao profissional que utiliza a ferramenta.
Usuário 18	sistema bem esquematizado e intuitivo.
Usuário 19	Agilizar a tomada de decisões
Usuário 20	Facilita analisar base de dados, podendo predizer informações da base de forma dinâmica e interativa.

Figura 70 – Pontos positivos destacados pelos usuários não especialistas. Fonte: Autor.

	Quais os pontos POSITIVOS sobre a funcionalidade de realizar tarefas preditivas.
Usuário 1	- Consegue fornecer tanto uma visão geral simples quanto amostragens mais rebuscadas de um mesmo dataset - Possibilidade de comparação entre métodos preditivos - Visualização gráfica passo a passo - Dicas e explicações
Usuário 2	A geração de análises apenas clicando em caixas com o mouse.
Usuário 3	Experiência bem simples e gera-se as análises exploratórias rapidamente
Usuário 4	Boa exploração dos dados, muitas opções de algoritmos
Usuário 5	O sistema apresenta confirmações ao usuário no decorrer do processos e resultados preliminares, o que auxilia na compreensão das etapas.
Usuário 6	Possibilidade de escolher modelos, avaliar distribuições e verificar diversas características do conjunto de dados.
Usuário 7	A automatização de algoritmos de machine learning, melhorando assim o tempo inicial.
Usuário 8	Tecnicamente completo, segue bem metodologias utilizadas por cientistas de dados, intuitivo
Usuário 9	Fácil de usar e com muitas funcionalidades
Usuário 10	Você consegue gerar com facilidade um benchmark dos modelos utilizados.

Figura 71 – Pontos positivos destacados pelos usuários especialistas. Fonte: Autor.

	Quais os pontos NEGATIVOS sobre a funcionalidade de realizar tarefas preditivas.
Usuário 11	nehum
Usuário 12	Nenhum
Usuário 13	Lentidão
Usuário 14	Algumas vezes não sabia o que escolher ou para que serve alguma seleção.
Usuário 15	Determinadas tarefas podem possuir variáveis demasiadamente complexas para a realização de tarefas preditivas.
Usuário 16	Não encontrei pontos negativos
Usuário 17	Sinceramente, eu não consigo pensar em nenhuma agora.
Usuário 18	nenhum
Usuário 19	Dependência e preguiça
Usuário 20	Nenhum

Figura 72 – Pontos negativos destacados pelos usuários não especialistas. Fonte: Autor.

	Quais os pontos NEGATIVOS sobre a funcionalidade de realizar tarefas preditivas.
Usuário 1	Algumas dicas adicionais de como utilizar algumas métricas seriam bem vindas.
Usuário 2	Muito texto na tela, interface ruim, muitos dados sendo exibidos sem necessidade.
Usuário 3	Acredito que precisa de precisa de uma explicação mais clara sobre os testes estatísticos e modelos para pessoas que não sabem do assunto. Talvez perguntar no inicio se a pessoa tem conhecimento. Caso sim, explica-se os conceitos com exemplos, caso não, não aparece.
Usuário 4	Muitas telas até a última etapa. Embora tenha muitas opções, talvez fosse bom ter como alterar as decisões da ferramenta.
Usuário 5	Somente a Interface
Usuário 6	Apenas pequenos detalhes de clareza da seleção das opções
Usuário 7	Nada a comentar
Usuário 8	Poderia ser um pouco mais organizado, talvez se fosse possível visualizar o fluxo por completo em cada fase poderia ser interessante, poderia ter mais métodos de seleção de atributos e redução de dimensionalidade
Usuário 9	Alguns problemas com imagens
Usuário 10	Acho que as descrições estão um pouco imprecisas, talvez seria interessante que fossem mais claras em relação aos objetivos.

Figura 73 – Pontos negativos destacados pelos usuários especialistas. Fonte: Autor.

	Qual a sua avaliação geral sobre a ferramenta DSAdvisor
Usuário 11	Ferramenta intuitiva e de fácil uso
Usuário 12	É uma ótima ferramenta para quem trabalha com ciência de dados, pois possui muitas funcionalidades importantes.
Usuário 13	Uma ótima ferramenta para ciência de dados, simples de usar.
Usuário 14	Muito boa
Usuário 15	De forma geral, considero uma boa ferramenta especialmente pela acessibilidade e praticidade inerentes de uma ferramenta web.
Usuário 16	A proposta da ferramenta é bem interessante. Só senti falta de algumas funcionalidades, com a de voltar para análise anterior e a de baixar gráficos e outras saídas.
Usuário 17	Acredito que iniciativas que propõem e reúnem metodologias, boas práticas e etc contribuem diretamente na rotina de um profissional de TI proporcionando direcionamento, agilidade em atividades. Senti isso com a ferramenta proposta e, caso eu atuasse na área de Ciência de Dados, eu teria curiosidade em testá-la em minha rotina de trabalho.
Usuário 18	nota 8
Usuário 19	Não serve tanto assim para o público leigo, visto que estatística não é tão fácil de entender quanto parece
Usuário 20	Ferramenta instrutiva, que ajudar a pessoas iniciantes a manipular base de dados, e facilitando o entendimento dos conceitos.

Figura 74 – Avaliação geral pelos usuários não especialistas. Fonte: Autor.

	Qual a sua avaliação geral sobre a ferramenta DSAdvisor
Usuário 1	Potencial didático exorbitante. Com um professor apresentando inicialmente a ferramenta para os alunos, seria uma ótima forma de não só atrair o interesse como introduzir a classe em tópicos mais avançados.
Usuário 2	Bastante complexa, precisa da ajuda de um especialista para conseguir utilizar. Para novatos será um problema.
Usuário 3	Boa.
Usuário 4	Interessante, principalmente para pessoas iniciantes em Ciência de Dados.
Usuário 5	Eficiente e bem descritiva.
Usuário 6	Muito boa.
Usuário 7	Ajuda bem no geral, pois automatiza processos nas etapas iniciais.
Usuário 8	Atende ao que é proposto
Usuário 9	Excelente
Usuário 10	Considero a experiência muito interessante, possui um bom workflow de decisão das métricas a serem utilizadas.

Figura 75 – Avaliação geral pelos usuários especialistas. Fonte: Autor.

	Idade	Gênero	Profissão	Você trabalha ou estuda em alguma área relacionada à Ciência de Dados?
Entrevistado 1	21 a 29	Feminino	Analista de projetos de T.I	Sim, estudo em um curso de área relacionada (Computação, Estatística, Engenharia da Computação e entre outros)
Entrevistado 2	21 a 29	Masculino	Cientista de Dados	Sim, estudo em um curso de área relacionada (Computação, Estatística, Engenharia da Computação e entre outros)
Entrevistado 3	30 a 39	Masculino	Professor universitário	Sim, trabalho há 5 anos ou menos
Entrevistado 4	21 a 29	Feminino	Cientista de Dados	Sim, trabalho há 5 anos ou menos
Entrevistado 5	21 a 29	Masculino	Cientista de Dados	Sim, trabalho há 5 anos ou menos
Entrevistado 6	30 a 39	Masculino	Professor universitário	Sim, trabalho há 5 anos ou menos
Entrevistado 7	21 a 29	Feminino	Desenvolvedor	Sim, estudo em um curso de área relacionada (Computação, Estatística, Engenharia da Computação e entre outros)
Entrevistado 8	21 a 29	Masculino	Desenvolvedor	Sim, estudo em um curso de área relacionada (Computação, Estatística, Engenharia da Computação e entre outros)
Entrevistado 9	30 a 39	Masculino	Desenvolvedor	Não
Entrevistado 10	21 a 29	Masculino	Cientista de Dados	Sim, trabalho há 5 anos ou menos
Entrevistado 11	30 a 39	Masculino	Desenvolvedor	Não
Entrevistado 12	21 a 29	Masculino	Desenvolvedor	Sim, estudo em um curso de área relacionada (Computação, Estatística, Engenharia da Computação e entre outros)
Entrevistado 13	21 a 29	Masculino	Advogado	Sim, trabalho há 5 anos ou menos
Entrevistado 14	21 a 29	Masculino	Desempregado	Não
Entrevistado 15	21 a 29	Masculino	Desenvolvedor	Não
Entrevistado 16	21 a 29	Masculino	Aluno de Pós-Graduação	Não
Entrevistado 17	21 a 29	Masculino	Desenvolvedor	Não
Entrevistado 18	21 a 29	Feminino	Desenvolvedor	Não
Entrevistado 19	21 a 29	Masculino	Desenvolvedor	Não
Entrevistado 20	21 a 29	Feminino	Aluna de Pós-Graduação	Sim, estudo em um curso de área relacionada (Computação, Estatística, Engenharia da Computação e entre outros)

Figura 76 – Perfil dos participantes. Fonte: Autor.

Perfil do Participante - Pesquisa de usabilidade

Prezado(a) participante, você está sendo convidado(a) a participar de uma pesquisa de usabilidade para uma ferramenta desenvolvida pelo aluno de mestrado José Augusto Câmara Filho do programa de pós-graduação Mestrado e Doutorado em Ciência da Computação(MDCC) da Universidade Federal do Ceará - UFC.

O objetivo desta pesquisa é identificar perfis de usuários em Ciência de Dados para em seguida participarem de um experimento remoto usando a ferramenta DSAdvisor.

Sua participação é voluntária e anônima.

A duração é de aproximadamente 20 minutos.

Qualquer dúvida no preenchimento do questionário você pode enviar e-mail para augustocam95@gmail.com

Gostaríamos de deixar claro que:
- apenas os responsáveis pela pesquisa terão acesso aos dados;
- os dados serão usados apenas para fins acadêmicos; e
- a divulgação dos dados será de forma anônima, preservando a privacidade de todos os participantes.

Desde já agradeço a colaboração. Sua participação é essencial no sucesso deste trabalho!

*Obrigatório

Termo de Consentimento

O(a) senhor(a) irá acessar a pesquisa, assim que concordar com este documento. Sua participação é importante, porém você não deve participar contra a sua vontade. Leia atentamente as informações abaixo e faça qualquer pergunta que desejar, para que todos os procedimentos desta pesquisa sejam esclarecidos.

Garanto que esta pesquisa não oferece nenhum risco de natureza física ou psicológica para o(a) senhor(a). Também garanto-lhe a privacidade, para sua maior segurança, será mantido sigilo em relação ao seu nome e/ou quaisquer outros aspectos que possam vir a identificá-lo(a), e as informações utilizadas neste estudo possuirão a única finalidade de colaborar com a presente dissertação de mestrado bem como a divulgação em relatórios e revistas científicas.

O consentimento para a participação é uma escolha livre e esta participação poderá ser interrompida a qualquer momento, caso o(a) senhor(a) precise ou deseje.

E ainda, para participar da mesma, não será oferecido nenhum valor ao (a) senhor (a). Portanto, nesta pesquisa, sua participação é totalmente voluntária.

1. *

Marcar apenas uma oval.

◯ Li e concordo com os termos

Opinião sobre ferramentas de Ciência de Dados

2. Nome *

3. Idade *

Marcar apenas uma oval.

- ◯ 17 ou menos
- ◯ 18 a 20
- ◯ 21 a 29
- ◯ 30 a 39
- ◯ 40 a 49
- ◯ 50 a 59
- ◯ 60 ou mais

4. Gênero *

Marcar apenas uma oval.

- ◯ Feminino
- ◯ Masculino
- ◯ Prefiro não dizer
- ◯ Outro: _____

5. E-mail *

_____

6. Em uma escala de 0 a 10, onde 0 seria usuário iniciante (precisa de ajuda para a maioria das tarefas) e 10 seria usuário avançado (consegue realizar a maioria das ações que deseja realizar), como você se considera em relação a proficiência em atividades relacionadas a Ciência de Dados? *

Marcar apenas uma oval.

	0	1	2	3	4	5	6	7	8	9	10	
preciso de ajuda para a maioria das tarefas	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	usuário avançado

7. Você trabalha ou estuda em alguma área relacionada à Ciência de Dados? *

Marcar apenas uma oval.

- ◯ Sim, trabalho há mais de 5 anos
- ◯ Sim, trabalho há 5 anos ou menos
- ◯ Sim, estudo em um curso de área relacionada (Computação, Estatística, Engenharia da Computação e entre outros)
- ◯ Não

8. Já utilizou alguma ferramenta voltada para Ciência de Dados, tais como Orange Tool, Weka, Knime, RapidMiner ou qualquer outra? *

Marcar apenas uma oval.

- ◯ Sim, uso rotineiramente
- ◯ Sim, mas com pouca frequência
- ◯ Sim, algumas vezes apenas por curiosidade ou experimentação
- ◯ Não

9. Caso já tenha utilizado alguma outra ferramenta voltado para Ciência de Dados, por favor nos informe qual!

_____

_____

_____

_____

_____

10. Já utilizou alguma linguagem de programação(Python, Matlab, R, por exemplo) para lhe auxiliar em atividades relacionadas a Ciência de Dados? *

Marcar apenas uma oval.

- ◯ Sim, uso rotineiramente
- ◯ Sim, mas com pouca frequência
- ◯ Sim, algumas vezes apenas por curiosidade ou experimentação
- ◯ Não

11. Caso já tenha utilizado alguma outra linguagem de programação para realizar tarefas de Ciência de Dados, por favor nos informe qual(is) você já utilizou.

_____

Voluntários para próxima etapa da pesquisa

Você tem interesse de participar de uma pesquisa de usabilidade em uma ferramenta chamada "DSAdvisor" voltada para atividades preditivas? Esta pesquisa será realizada de forma remota, tudo o que precisa é que tenha um computador, internet banda larga/fibra para acessar o link da ferramenta e possa realizar uma chamada de voz com o pesquisador durante 10-15 minutos. Nesta chamada de voz será realizado um tour pela ferramenta para exibição de todas as funcionalidades existentes na aplicação para realização de uma tarefa preditiva de um conjunto de dados sintético.

12. Gostaria de participar da próxima etapa da pesquisa? *

Marcar apenas uma oval.

- ◯ Sim
- ◯ Não

13. Caso tenha interesse em participar da pesquisa por favor nos indique 3 dias e horários para participar desta pesquisa.

Exemplo: 7 de janeiro de 2019

14. Hora

Exemplo: 08h30

15.

Exemplo: 7 de janeiro de 2019

16. Hora

Exemplo: 08h30

17.

Exemplo: 7 de janeiro de 2019

18. Hora

Exemplo: 08h30

19. Caso prefira, pode deixar seu numero de telefone para facilitar a marcação desta segunda etapa da pesquisa.

20. Observações e/ou sugestões

Introdução

Prezado(a) participante, você está sendo convidado(a) a participar de uma pesquisa de usabilidade para uma ferramenta desenvolvida pelo aluno de mestrado José Augusto Câmara Filho do programa de pós-graduação Mestrado e Doutorado em Ciência da Computação(MDCC) da Universidade Federal do Ceará - UFC.

O objetivo desta pesquisa é avaliar a usabilidade da ferramenta "DSAdvisor".

Sua participação é voluntária e anônima.

A duração é de aproximadamente 20 minutos.

Qualquer dúvida no preenchimento do questionário você pode enviar e-mail para augustocam95@gmail.com.

Gostaríamos de deixar claro que:
- apenas os responsáveis pela pesquisa terão acesso aos dados.
- os dados serão usados apenas para fins acadêmicos.
- e a divulgação dos dados será de forma anônima, preservando a privacidade de todos os participantes.

Desde já agradeço a colaboração. Sua participação é essencial no sucesso deste trabalho.

*Obrigatório

Termo de Consentimento

O(a) senhor(a) irá acessar a pesquisa, assim que concordar com este documento. Sua participação é importante, porém você não deve participar contra a sua vontade. Leia atentamente as informações abaixo e faça qualquer pergunta que desejar, para que todos os procedimentos desta pesquisa sejam esclarecidos.

Garanto que esta pesquisa não oferece nenhum risco de natureza física ou psicológica para o(a) senhor(a). Também garanto-lhe a privacidade, para sua maior segurança, será mantido sigilo em relação ao seu nome e/ou quaisquer outros aspectos que possam vir a identificá-lo(a), e as informações utilizadas neste estudo possuirão a única finalidade de colaborar com a presente dissertação de mestrado bem como a divulgação em relatórios e revistas científicas.

O consentimento para a participação é uma escolha livre, e esta participação poderá ser interrompida a qualquer momento, caso o(a) senhor(a) precise ou deseje.

E ainda, para participar da mesma, não será oferecido nenhum valor ao (a) senhor (a). Portanto, nesta pesquisa, sua participação é totalmente voluntária.

1. Clique na opção para afirmar a leitura e a concordância dos termos *

   Marcar apenas uma oval.

   ◯ Li e concordo com os termos

2. Clique na opção abaixo para concordar em participar dessa entrevista *

   Marcar apenas uma oval.

   ◯ Concordo em participar

Questionário para avaliação da realização de tarefas preditivas

Prezado(a) participante, você está sendo convidado(a) a participar de uma pesquisa de usabilidade para uma ferramenta desenvolvida pelo aluno de mestrado José Augusto Câmara Filho do programa de pós-graduação Mestrado e Doutorado em Ciência da Computação(MDCC) da Universidade Federal do Ceará - UFC.

O objetivo desta pesquisa é avaliar a realização de tarefas preditivas com o sistema DSAdvisor.

Sua participação é voluntária e anônima.

A duração é de aproximadamente 20 minutos.

Qualquer dúvida no preenchimento do questionário você pode enviar e-mail para augustocam95@gmail.com

Gostaríamos de deixar claro que:
- apenas os responsáveis pela pesquisa terão acesso aos dados;
- os dados serão usados apenas para fins acadêmicos; e
- a divulgação dos dados será de forma anônima, preservando a privacidade de todos os participantes.

Desde já agradeço a colaboração. Sua participação é essencial no sucesso deste trabalho.

*Obrigatório

1. *

Marcar apenas uma oval.

◯ Afirmo que li e concordo com os termos da pesquisa (apresentados no formulário inicial)

Termo de Consentimento

O(a) senhor(a) irá acessar a pesquisa, assim que concordar com este documento. Sua participação é importante, porém você não deve participar contra a sua vontade. Leia atentamente as informações abaixo e faça qualquer pergunta que desejar, para que todos os procedimentos desta pesquisa sejam esclarecidos.

Garanto que esta pesquisa não oferece nenhum risco de natureza física ou psicológica para o(a) senhor(a). Também garanto-lhe a privacidade, para sua maior segurança, será mantido sigilo em relação ao seu nome e/ou quaisquer outros aspectos que possam vir a identificá-lo(a), e as informações utilizadas neste estudo possuirão a única finalidade de colaborar com a presente dissertação de mestrado bem como a divulgação em relatórios e revistas científicas.

O consentimento para a participação é uma escolha livre e esta participação poderá ser interrompida a qualquer momento, caso o(a) senhor(a) precise ou deseje.

E ainda, para participar da mesma, não será oferecido nenhum valor ao (a) senhor (a). Portanto, nesta pesquisa, sua participação é totalmente voluntária.

2. *Marcar apenas uma oval.*

- ◯ Li e concordo com os termos

Identificação

3. Nome *

_____

4. Email *

_____

Avaliação da ferramenta DSAdvisor

A seguir serão apresentadas algumas questões relacionadas à ferramenta (sistema) DSAdvisor. Todas essas perguntas referem-se à utilização da ferramenta DSAdvisor na realização de tarefas preditivas.

5. Eu acho que gostaria de usar esse sistema com frequência. *

Marcar apenas uma oval.

- ◯ Discordo Totalmente
- ◯ Discordo
- ◯ Não estou decidico
- ◯ Concordo
- ◯ Concordo totalmente

6. Eu acho o sistema desnecessariamente complexo. *

Marcar apenas uma oval.

- ◯ Discordo Totalmente
- ◯ Discordo
- ◯ Não estou decidico
- ◯ Concordo
- ◯ Concordo totalmente

7. Eu achei o sistema fácil de usar. *

Marcar apenas uma oval.

- Discordo Totalmente
- Discordo
- Não estou decidico
- Concordo
- Concordo totalmente

8. Eu acho que precisaria de ajuda de uma pessoa com conhecimentos técnicos para usar o sistema. *

Marcar apenas uma oval.

- Discordo Totalmente
- Discordo
- Não estou decidico
- Concordo
- Concordo totalmente

9. Eu acho que as várias funções do sistema estão muito bem integradas. *

Marcar apenas uma oval.

- Discordo Totalmente
- Discordo
- Não estou decidico
- Concordo
- Concordo totalmente

10. Eu acho que o sistema apresenta muita inconsistência. *

Marcar apenas uma oval.

- Discordo Totalmente
- Discordo
- Não estou decidico
- Concordo
- Concordo totalmente

11. Eu imagino que as pessoas aprenderão como usar esse sistema rapidamente. *

Marcar apenas uma oval.

- Discordo Totalmente
- Discordo
- Não estou decidico
- Concordo
- Concordo totalmente

12. Eu achei o sistema atrapalhado de usar. *

Marcar apenas uma oval.

- Discordo Totalmente
- Discordo
- Não estou decidico
- Concordo
- Concordo totalmente

13. Eu me senti confiante ao usar o sistema. *

Marcar apenas uma oval.

- Discordo Totalmente
- Discordo
- Não estou decidico
- Concordo
- Concordo totalmente

14. Eu precisei aprender várias coisas novas antes de conseguir usar o sistema. *

Marcar apenas uma oval.

- Discordo Totalmente
- Discordo
- Não estou decidico
- Concordo
- Concordo totalmente

Qual a probabilidade de você recomendar a DSAdvisor para um colega ou amigo?

15. Escolha uma nota para indicar o quanto recomendaria a ferramenta proposta.

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Nem um pouco provável	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Extremamente provável

Opinião geral sobre a DSAdvisor

A seguir serão apresentadas algumas perguntas relacionadas a sua experiência geral do teste.

16. Quais os pontos POSITIVOS sobre a funcionalidade de realizar tarefas preditivas. *

17. Quais os pontos NEGATIVOS sobre a funcionalidade de realizar tarefas preditivas. *

18. Qual a sua avaliação geral sobre a ferramenta DSAdvisor *

19. Como você julgaria a funcionalidade do "Best fit" para estimar as distribuições mais próximas das variáveis numéricas contínuas do conjunto de dados? *

Marcar apenas uma oval.

	1	2	3	4	5	
Nada útil	◯	◯	◯	◯	◯	Muito útil

20. Como você julgaria a funcionalidade de identificação de valores anômalos (Outlier Detection) presentes no conjunto de dados? *

Marcar apenas uma oval.

	1	2	3	4	5	
Nada útil	◯	◯	◯	◯	◯	Muito útil

21. A utilização de uma ferramenta web para realizar tarefas preditivas é uma vantagem a ser considerada em comparação às ferramentas que precisam ser instaladas? *

Marcar apenas uma oval.

- ◯ Discordo totalmente
- ◯ Discordo
- ◯ Não sei opinar
- ◯ Concordo
- ◯ Concordo Totalmente

22. A DSAdvisor possibilita gerar um arquivo contendo todas as informações acerca das decisões tomadas durante a sua utilização, com o objetivo de assegurar a reprodutibilidade. Essa funcionalidade é relevante quando se deseja realizar tarefas preditivas. *

Marcar apenas uma oval.

- ◯ Discordo totalmente
- ◯ Discordo
- ◯ Não sei opinar
- ◯ Concordo
- ◯ Concordo Totalmente

Feedback sobre o experimento

Caso tenha interesse em deixar algum feedback sobre o experimento, em qualquer aspecto, por favor informe na pergunta a seguir.

23. Opiniões e sugestões sobre o experimento