UNIVERSIDADE DO ESTADO DO
RIO DE JANEIRO

INSTITUTO POLITÉCNICO
GRADUAÇÃO EM ENGENHARIA
DE COMPUTAÇÃO

Leonardo Simões

Aplicativo web para análise de dados

Nova Friburgo
2021

UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO

INSTITUTO POLITÉCNICO
GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Leonardo Simões

Aplicativo web para análise de dados

Trabalho de conclusão de curso apresentado como pré-requisito para obtenção do título de Engenheiro de Computação, ao Departamento de Modelagem Computacional, do Instituto Politécnico, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Bernardo Sotto-Maior Peralva

Nova Friburgo
2021

UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO
INSTITUTO POLITÉCNICO - CURSO DE ENGENHARIA DE COMPUTAÇÃO

Reitor: Ricardo Lodi Ribeiro
Vice-Reitor: Mario Sérgio Alves Carneiro
Diretor do Instituto Politécnico: Angelo Mondaini Calvão
Coordenadora de Curso: Ana Cristina Fontes Moreira

Banca Avaliadora Composta por: Prof. Dr. Bernardo Sotto-Maior Peralva (Orientador)
Prof. Dr. Roberto Pinheiro Domingos
Prof. Dr. Guilherme de Melo Baptista Domingues

Ficha elaborada pelo autor através do
Sistema para Geração Automática de Ficha Catalográfica da Rede Sirius - UERJ

S593 Simões, Leonardo
Aplicativo web para análise de dados / Leonardo Simões. - 2021.
60 f.

Orientador: Bernardo Sotto-Maior Peralva
Trabalho de Conclusão de Curso apresentado à Universidade do Estado do Rio de Janeiro, Instituto Politécnico, para obtenção do grau de bacharel em Engenharia da Computação.

1. Ciência de dados - Monografias. 2. Análise de dados - Monografias. 3. Aplicativo web - Monografias. 4. Python - Monografias. I. Peralva, Bernardo Sotto-Maior. II. Universidade do Estado do Rio de Janeiro. Instituto Politécnico. III. Título.

CDU 004.41

Endereço: UERJ - IPRJ, Rua Bonfim, 25 - Prédio 5, Vila Amélia. CEP 28625-570 - Nova Friburgo - RJ - Brasil.

Leonardo Simões

Leonardo Simões

Aplicativo web para análise de dados

Trabalho de conclusão de curso apresentado como pré-requisito para obtenção do título de Engenheiro de Computação, ao Departamento de Modelagem Computacional, do Instituto Politécnico, da Universidade do Estado do Rio de Janeiro.

Aprovado em 12 de março de 2021.

Banca examinadora:

Prof. Dr. Bernardo Sotto-Maior Peralva
Instituto Politécnico - UERJ

Prof. Dr. Roberto Pinheiro Domingos
Instituto Politécnico - UERJ

Prof. Dr. Guilherme de Melo Baptista Domingues
Instituto Politécnico - UERJ

Nova Friburgo
2021

DEDICATÓRIA

Dedico este trabalho à minha família, que me incentivou a cursar uma graduação.

AGRADECIMENTOS

Agradeço primeiramente a Deus por toda ajuda diante das dificuldades e desafios durante a vida e a graduação.

Aos professores, colegas e amigos que me apoiaram e ajudaram de alguma forma a superar os obstáculos presentes durante o tempo que cursei minha graduação.

RESUMO

SIMÕES, L. *Aplicativo web para análise de dados*. 2021. 60 f. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) - Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2021.

Tipicamente, para analisar de dados de forma eficiente é necessário o uso de softwares instalados previamente no dispositivo, como por exemplo, o Microsoft Power BI. Alternativamente, pode-se optar por ambientes e linguagens de programação, como RStudio, Python e Jupyter Notebook, que também tornam necessários conhecimentos programação. Desta forma, o presente trabalho tem como objetivo o desenvolvimento de um aplicativo *web* que possibilite a realização do processo de análise dos dados presentes em arquivo de texto, em formato csv, tsv ou txt, carregado pelo usuário. Através da aplicação desenvolvida neste trabalho qualquer usuário com acesso a internet e o conjunto de dados pode realizar a análise destes dados de forma eficiente, de modo que as etapas são realizadas de forma semiautomatizada e personalizada conforme as opções selecionadas.

Palavras-chave: Ciência de dados. Análise de dados. Aplicativo web. Python.

ABSTRACT

SIMÕES, L. *Web application for data analysis*. 2021. 60 f. Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) - Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2021.

Typically, in order to efficiently perform data analysis, specific software tools are required and installed in a device, such as the Microsoft Power BI. Alternatively, programming languages may also be chosen, where data analysis features are available for use as it is the case for RStudio, Python or Jupyter Notebook, for instance. However, for the latter, some programming skills is required. Therefore, this work presents a web application that carries out some important data analysis operations. It only requires from the user a text file (in txt, csv or tvs formats) which is loaded into the developed tool. Through the proposed application, any user with internet access, and a data set, is able to extract relevant information from the data in an automatic and customized way, according to the available options previously selected by the user.

Keywords: Data science. Data analysis. Web application. Python.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

UERJ	Universidade do Estado do Rio de Janeiro
IPRJ	Instituto Politécnico do Rio de Janeiro

SUMÁRIO

INTRODUÇÃO

O volume de dados gerados pela humanidade tem crescido de forma exponencial durante a última década, principalmente devido a avanços tecnológicos, como aumento da velocidade de acesso à internet, e melhorias em dispositivos móveis, assim como também, a disseminação das redes sociais e crescimento do *e-commerce*.

A área de ciência de dados também cresceu e evoluiu bastante com estes adventos, tanto em meio acadêmico quanto em negócios. A ciência de dados é considerada com uma intersecção de outras três grandes áreas, a engenharia de *software*, matemática e a área de negócios. A análise de dados é uma subárea da ciência de dados, que se foca na parte de exploração dos dados, visualização e aplicação estatística.

Com o aumento do volume de dados, empresas, pesquisadores e pessoas interessadas buscam utilizá-los para agregar maior valor a negócios, realizar descobertas, prever resultados e tendências, confirmar hipóteses. A análise de dados tem sido aplicada às mais diversas áreas de negócios, como *marketing*, administração, contabilidade e medicina. Por exemplo, na medicina, tem sido empregado o Registro Eletrônico de Saúde (RES) onde cada paciente tem seu próprio registro digital, incluindo informações demográficas, e dados gerais sobre a evolução de doenças como o histórico médico, alergias, resultados de exames laboratoriais e etc. Tais registros são armazenados e compartilhados com segurança através de sistemas de informação, estando disponíveis para provedores do setor público e privado. Desta forma, os profissionais de saúde podem atualizar os registros e implementar mudanças ao longo do tempo sem burocracia e sem risco de duplicação de dados.

Em um outro exemplo, agora na advocacia, uma aplicação prática é a identificação de tendências de decisões que pode ser muito útil, por exemplo, quando um grupo de condôminos entra com uma ação judicial, reclamando que a construtora não cumpriu uma cláusula contratual. Pela análise de casos semelhantes, o juiz proferirá uma sentença favorável aos condôminos ou à construtora? É uma resposta que pode ser encontrada analisando qual é a predisposição da Justiça brasileira nestes casos.

Portanto, sistemas que fornecem informações sobre uma base de dados são bastante úteis de uma maneira geral, inclusive para o usuário que desconhece tecnologias de programação e teorias de análises de dados simples, como modelos de regressão linear.

Objetivos

O objetivo deste trabalho é criar uma aplicação web para análise de dados

estruturados e tabulares de forma interativa e semiautomatizada, de modo que, facilite o processo e possa ser realizada por uma pessoa sem conhecimentos em programação, e sem necessidade de instalar algum aplicativo.

O público-alvo consiste em qualquer pessoa com acesso à internet e com um conjunto de dados bem organizado e no formato de arquivo de texto. Ao contrário dos *softwares* mais convencionais, o usuário precisa apenas marcar alguns caixas de seleção e clicar nas opções desejadas para fazer a maior parte da análise. Apenas para a parte de consulta personalizada o usuário precisa digitar algo.

Visita guiada

No Capítulo 1, é apresentada uma fundamentação teórica sobre os conceitos envolvidos em uma análise de dados, e também informações sobre as tecnologias utilizadas no desenvolvimento do aplicativo. No Capítulo 2, é mostrado como os conceitos e tecnologias foram utilizados para a construção da aplicação web. No Capítulo 3, são exibidos os resultados, através de imagens e comentários, da execução da aplicação sobre um conjunto de dados escolhido. Ao final, é apresentada uma conclusão sobre o projeto desenvolvido.

1 FUNDAMENTAÇÃO TEÓRICA

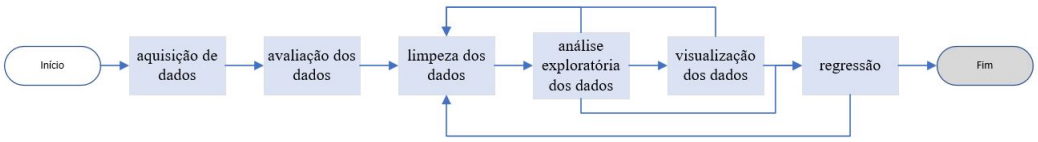
Este capítulo apresenta os fundamentos teóricos sobre o processo de análise de dados e informações sobre ambientes de programação usados, tais como a linguagem de programação *Python* e seus módulos.

1.1 Análise de dados

O fluxo de uma análise de dados não é sempre o mesmo para todas as situações, a divisão em etapas pode ser diferente, a ordem das operações pode mudar, e algumas destas operações podem não ser feitas. A mudança do fluxo da análise pode variar devido a vários aspectos, mas principalmente, por convenção, tipo de problema tratado e formato dos dados.[1]

O fluxo de análise de dados adotado neste trabalho foi definido pelas etapas de aquisição de dados, avaliação dos dados, limpeza dos dados, análise exploratória dos dados, visualização dos dados e regressão. Este fluxo geralmente é linear, mas para as etapas após a de limpeza, pode haver um retorno para uma das etapas anteriores até a de limpeza.

Figura 1 – Fluxo de análise de dados



Fonte: O Autor (2020).

As etapas de aquisição de dados, avaliação de dados e limpeza de dados compõem o processo de preparação de dados.[1]

1.1.1 Preparação dos dados

Como sugerido pelo nome, o processo de preparação de dados é responsável por garantir que os dados estejam prontos para serem usados de maneira eficiente, sem problemas de qualidade, em uma análise exploratória ou em uma aplicação de *machine learning*.[1]

A etapa de aquisição de dados consiste nas diversas formas de coleta de dados para a análise.[1] Algumas destas formas de coleta incluem uso de banco de dados, web scraping, download e leitura de arquivos.[1] Para este aplicativo, a etapa de aquisição de dados é feita através do upload de um arquivo de texto, no formato csv, tsv ou txt, com separador de vírgula, ponto e vírgula, espaço ou tabulação.

A etapa de avaliação de dados consiste em examinar o conjunto de dados a fim de identificar as principais características e problemas de qualidade dos dados.[1] As características avaliadas são as dimensões, número de linhas e colunas, os nomes das colunas e os seus respectivos tipos. Os problemas de qualidade verificados são a presença de linhas duplicadas, a quantidade e porcentagem de valores ausentes, representados por valor nulo, para cada coluna.

A etapa de limpeza de dados consiste em resolver os problemas de qualidade de dados detectados na etapa anterior e remoção de dados indesejados.[1] Para remover dados indesejados, normalmente, opta-se por remover colunas inteiras do conjunto ou por remover linhas cujo valor para determinada coluna esteja ausente. Para linhas cujo valores de alguma coluna numérica esteja ausente, além de excluí-las, pode se optar por preenche-las com algum valor significante, como 0, mínimo, máximo, média ou mediana. Nesta etapa, as linhas duplicadas, além da primeira ocorrência, são excluídas.

1.1.2 Análise exploratória dos dados

A etapa de análise exploratória de dados consiste em extrair informações relevantes em relação aos dados considerados.[1] Nesta etapa, as estatísticas descritivas são usadas para caracterizar cada coluna numérica do conjunto de dados. As principais estatísticas usadas para colunas numéricas são média, mediana, quantidade, mínimo, máximo, desvio padrão, quartil inferior e quartil superior. Para as colunas não numéricas, há a contagem de valores únicos. Usando agrupamentos por colunas e consultas personalizadas ao conjunto de dados, é possível obter melhores caracterizações dos dados.

1.1.3 Visualização gráfica dos dados

As visualizações de dados são feitas através de gráficos gerados em um plano de uma ou duas dimensões e podem caracterizar uma, duas ou até três colunas ao mesmo tempo em uma mesma figura.[2] Os gráficos considerados neste trabalho foram histograma, gráfico de barras, gráfico de caixa (BoxPlot), gráfico de dispersão (pontos), gráfico de pizza, gráfico de violino, gráfico de barras agrupadas para 2 variáveis, gráfico de barras agrupadas para 3 variáveis, pairplot, mapa de calor das correlações.

Um histograma é um gráfico de barras que mostra a distribuição de uma variável numérica, de preferência de valor real, ao invés de inteira, no eixo x, enquanto que o eixo y indica sua frequência absoluta.[3]

Um gráfico de barras é um gráfico que mostra a frequência de uma variável qualitativa através de barras, onde o comprimento da barra define um valor numérico, como a frequência, e a base é distinta para cada valor da variável.[4]

Um gráfico de caixa, ou boxplot, ilustra de forma forma gráfica o resumo dos cinco números, ou seja, mínimo, quartil inferior, mediana, quartil superior e máximo, de uma variável numérica. Os quartis são ilustrados em uma caixa, das quais partem duas linhas, uma até o mínimo e outra até o máximo.[5] O mínimo e máximo são calculados usando as medidas dos quartis, e medidas abaixo desse mínimo ou acima deste máximo são exibidos como pontos e considerados possíveis outliers.

Um gráfico de pizza é um gráfico que consiste em uma região circular que se divide em regiões que expressam a frequência de valores de uma variável categórica.[6] Este tipo de gráfico facilita a visualização da frequência de valores em proporções ou porcentagens.

Um gráfico de dispersão é construído com duas variáveis numéricas, uma no eixo x e outra no eixo y, e formados por pontos com coordenadas correspondentes.[7] Através de sua visualização, pode se supor uma correlação caso a maioria dos pontos aparentemente se situam em torno de uma reta diagonal imaginária.

Um gráfico de violino é um gráfico, semelhante ao gráfico de caixa, que exibe a distribuição de uma variável numérica em relação a sua frequência ou a outra variável numérica.[8] O formato do gráfico lembra um violino, e os quartis da variável numérica podem ser ilustrados na figura.

Um gráfico de barras agrupadas para 2 variáveis é uma variação do gráfico de barras de modo que um dos eixos exibe a frequência absoluta, enquanto o outro eixo possui as combinações de valores de duas variáveis qualitativas. Um gráfico de barras agrupadas para 3 variáveis é semelhante a um gráfico de barras agrupadas para 2 variáveis, exceto pelo fato de que ao invés de representar a frequência absoluta em um dos eixos, este eixo representa uma variável numérica.[1]

O pairplot, ou "gráfico em pares", é um tipo de gráfico onde há uma disposição de gráficos em uma figura de modo posicional, semelhante a uma matriz. Nesta suposta matriz, os gráficos na diagonal são histogramas, e os demais são gráficos de dispersão.[9]

Um mapa de calor utiliza uma escala de cores, normalmente indica em uma legenda, para representar a intensidade de uma variável ou medida.[10] Uma matriz de correlação é aquela em que exibe o valor das correlações entre duas variáveis, indicadas por índices na matriz, sendo que os seus elementos presentes na diagonal principal possuem valores iguais a 1, enquanto os demais possuem valores entre -1 e 1 inclusive. O mapa de calor de correlações é uma junção de um gráfico de calor e de uma matriz de correlação.[11]

1.1.4 Regressão linear e regressão logística

As regressões são técnicas de aprendizado de máquina (machine learning) supervisionado, utilizadas para predições de valor de uma variável, ou coluna, a

partir dos valores de outras.[1] As regressões se apresentam como alternativas às estatísticas inferenciais e outros métodos de machine learning. As regressões resultam em coeficientes que associam valores das variáveis preditoras e uma constante a um valor para a variável predita através de uma função.[12]

Dois tipos de regressão são a regressão linear e a regressão logística. Para a regressão linear, a variável predita possui um valor numérico contínuo, enquanto que, para a regressão logística, a variável predita possui um valor numérico discreto binário (0 ou 1).

1.2 **Regressão linear**

Análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas a fim de predizer uma outra variável.[13] Esta abordagem é utilizada em diversas áreas como ciências biológicas, sociais e etc.

Através da regressão é possível realizar diversas estimativas, como estimar vendas de um produto baseando-se na relação entre vendas e gastos com publicidade. Também é possível estimar a performance de um funcionário através da relação performance e testes de aptidão e etc.

1.3 **Relação entre Variáveis**

Relações entre duas variáveis podem ser relações funcionais ou estatísticas.[14] Esta seção será responsável por apresentar ambas.

1.3.1 Relação Funcional entre Duas Variáveis

A relação funcional entre duas variáveis é expressada matematicamente de forma que se *X* é uma variável independente e *Y* uma variável dependente,[14] a relação funcional será dada por:

$$Y = f(X) \tag{1}$$

Onde, dado um *X*, *f* resultará no valor correspondente de *Y*.

Um simples exemplo pode ser dado através da relação funcional:

$$Y = 2X \tag{2}$$

X e *Y* podem significar qualquer coisa, por exemplo, *Y* o valor em dólares e *X* a quantidade de unidades vendidas, onde o preço unitário seria de 2 dólares. Com isso teríamos os seguintes dados:

Tabela 1 – Unidades Vendidas.

Unidades Vendidas	Valor
25	50
75	150
130	260

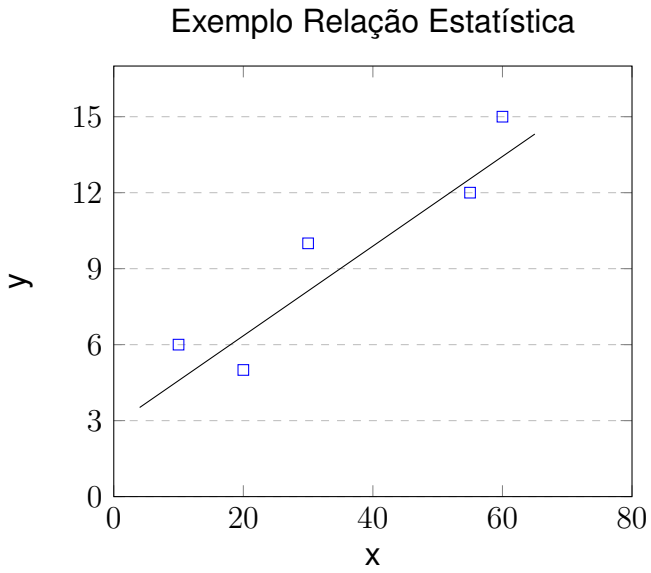
Fonte: O Autor (2021).

1.3.2 Relação estatística entre Duas Variáveis

Relação estatística não é tão simples e direta como a relação funcional, isto é, fazendo analogia a um gráfico, dificilmente será obtido uma função que "passa"por todos os seus pontos.[14]

O gráfico a seguir exemplifica este fato:

Figura 2 – Exemplo - Relação Estatística



Fonte: O Autor (2021).

Percebe-se que apesar do perfil linear, não é possível estabelecer uma única reta que passa por todos os pontos, portanto, pode-se obter uma reta onde o somatório dos erros seja mínimo, apesar de não perfeito, tal abordagem propicia bons resultados.

1.4 **Modelos de Regressão**

Análise de Regressão foi desenvolvida por Francis Galton durante o século XIX. Ele estudou a relação de altura entre pais e filhos e concluiu que filhos de pais altos e baixos tendem para a média do grupo.[15]

Regressão pode ser definida como a forma de se expressar variáveis em uma relação estatística, e sua análise tem por objetivo descrever, controlar ou predizer resultados de um problema.[13] A construção de um modelo de regressão é dado pelas escolhas das variáveis de escopo e pela função, que geralmente é linear, quadrática ou obtida empiricamente através dos dados coletados.

Além disso, é importante mencionar que a relação estatística entre duas variáveis não implica em dependência entre elas, não importa o quão forte seja esta esta relação. Em outras palavras, correlação não implica causalidade. De forma geral, a análise de regressão provê informações sobre padrões casuais, no entanto, o real significado de cada padrão deve ser feito caso a caso.

1.5 **Modelo Simples de Regressão Linear**

O modelo de regressão linear é dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{3}$$

onde Y_i é i-ésimo o valor aproximado, β_0 e β_1 parâmetros, X_i é o i-ésimo valor da variável conhecida, ϵ_i o i-ésimo erro aleatório de média 0 e variância σ^2, e $i = 1, ..., n$.[13]

Este modelo possui este nome devido ao fato de haver apenas uma variável de interesse, parâmetros são lineares e a variável é elevada apenas a potência primeira. Este tipo de modelo também é conhecido como modelo de primeira ordem. Além disso, o nome variável aleatória de Y_i se dá ao fato de Y_i ser a soma de uma parte constante ($\beta_0 + \beta_1 X_i$) e a soma de uma parte aleatória (ϵ_i)

Os parâmetros β_0 e β_1 são chamados de coeficientes de regressão. β_0 é a intercepção com *Y*, logo β_0 fornece a probabilidade da distribuição de *Y* quando $X = 0$, no entanto, tal coeficiente não possui um significado aplicável quando o modelo não aceita $X = 0$. Já β_1 indica a mudança da média da probabilidade de *Y* por cada unidade aumentada em *X*.[15]

1.6 **Dados para a Análise de Regressão**

Geralmente não sabemos os valores de β_0 e β_1 a priori, por isso deve-se encontrá-los com base em dados relevantes, e tais dados podem ser classificados em dados Observacionais e dados Experimentais.[14]

Dados observacionais são dados obtidos de estudos não experimentais, tais estudos não fazem controle das variáveis de interesse.[14] Um exemplo seria o estudo

da relação entre idade de um funcionário (*X*) e o número de dias doente no ano (*Y*). Este estudo é observacional, pois a idade á uma informação não controlada.

Dados experimentais são obtidos quando não há informação a priori, e então, um ambiente é preparado a fim de que seja possível coletar tais dados baseando-se em amostras aleatórias. Um exemplo seria treinar dois grupos de funcionários, o primeiro por 2 semanas e o segundo por 4 semanas. Após, faz-se a análise de ambos grupos a fim de checar qual grupo teve a melhor performance, ou seja, obtendo os resultados experimentalmente.

Dados experimentais baseados em amostras aleatórias tendem a prover informações mais sólidas a respeito da relação causa e efeito se comparado aos Dados ocupacionais.

1.7 **Estimativa da Função de Regressão**

1.7.1 Método dos Mínimos Quadrados

O Método dos Mínimos Quadrados (MMQ) para dados (X_i, Y_i), considera o desvio de Y_i em relação ao seu valor esperado:

$$Y_i - (\beta_0 + \beta_1 X_i) \tag{4}$$

O objetivo deste método é encontrar boas estimativas para os parâmetros β_0 e β_1 a fim de que a diferença ao quadrado de Y_i em relação ao seu valor esperado seja mínima (*Q*). Comumente utiliza-se a notação de b_0 para β_0 e b_1 para β_1.

$$Q = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 \tag{5}$$

Ambos b_0 e b_1 podem-se ser encontrados numericamente e analiticamente, onde este, é viável apenas para modelos de regressão matematicamente não complexos.

Desta forma, tais coeficientes podem ser encontrados através:

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \tag{6}$$

$$b_0 = \frac{1}{n}(\sum Y_i - b_1 \sum X_i) = \bar{Y} - b_1 \bar{X} \tag{7}$$

onde $\bar{X}$ e $\bar{Y}$ são as médias de X_i e Y_i respectivamente.

Exemplo: Suponha um teste feito com pessoas de diferentes idades a fim de resolverem um problema complexo. Este teste registra o número tentativas até a pessoa desistir, e com isso montou-se a seguinte tabela:

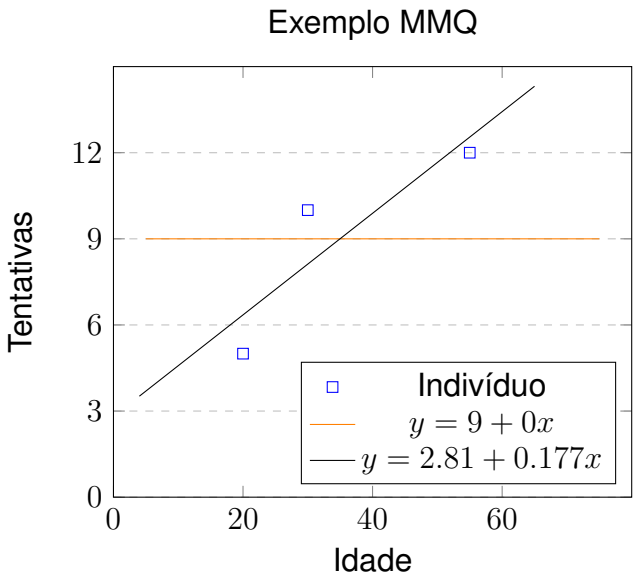
Tabela 2 – Número de tentativas.

Indivíduo	Idade	Nº de tentativas
1	20	5
2	30	10
3	55	12

Fonte: O Autor (2021).

Baseando-se nesta tabela, foi possível plotar o seguinte gráfico:

Figura 3 – Exemplo - MMQ



Fonte: O Autor (2021).

Pelo perfil dos resultados dos indivíduos, percebe-se que uma reta pode ser uma boa aproximação, no entanto, há uma infinidade de opções de reta, e por isso, deve-se buscar a que possui o menor erro possível.

Por exemplo, foi selecionada uma reta qualquer $y = 9 + 0x$ demonstrada em laranja no gráfico acima. Claramente percebe-se que tal reta não é uma boa aproximação, no entanto, a reta $y = 2.81 + 0.177x$ em preto demonstra uma aproximação bem melhor. De fato, esta última aproximação foi obtida através do MMQ, por isso sua aproximação ótima. Seguem os passos para tal aproximação.

Primeiro, calcula-se as médias de cada conjunto.

$$\bar{X} = \frac{20 + 30 + 55}{3} => \bar{X} = 35 \tag{8}$$

$$\bar{Y} = \frac{5+10+12}{3} => \bar{Y} = 9 \tag{9}$$

Aplicando $\bar{X}$ e $\bar{Y}$ na eq. 6 tem-se

$$\begin{aligned} b_1 &= \frac{(20-35)(5-9)+(30-35)(10-9)+(55-35)(12-9)}{(20-35)^2+(30-35)^2+(55-35)^2} \\ b_1 &= \frac{60-5+60}{225+25+400} => b_1 = 0.177 \end{aligned} \tag{10}$$

Logo aplicando eq.10 na eq.7 tem-se:

$$b_0 = 9 - 0.17735 => b_0 = 2.81 \tag{11}$$

Obtendo-se por fim a equação utilizada inicialmente

$$y = 2.81 + 0.177 * x \tag{12}$$

1.7.2 Erro mínimo do MMQ

Como exposto no início deste capítulo, o objetivo do MMQ é obter boas estimativas para os parâmetros b_0 e b_1 a fim de que a diferença ao quadrado de Y_i em relação ao seu valor esperado seja mínima (*Q*).

Sabe-se do Cálculo Diferencial que mínimos e máximos podem ser obtidos partindo-se de pontos críticos, ou seja, onde as derivadas são iguais a zero. Sendo assim, para obter o erro mínimo da estimativa deve-se minimizar b_0 e b_1 e consequentemente fazer o uso das derivadas parciais para estas variáveis.

Logo o valor mínimo de b_0 é encontrado da seguinte forma:

$$\begin{aligned} &\frac{\partial Q}{\partial b_0} = \frac{\partial(\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2)}{\partial b_0} = 0 \\ &2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)(-1) = 0 \\ &-\sum_{i=1}^{n}Y_i + \sum_{i=1}^{n}b_0 + \sum_{i=1}^{n}b_1X_i = 0 \\ &nb_0 = \sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}b_1X_i \\ &b_0 = \frac{\sum_{i=1}^{n}Y_i}{n} - b_1\frac{\sum_{i=1}^{n}X_i}{n} \\ &b_0 = \bar{Y} - b_1\bar{X} \end{aligned} \tag{13}$$

E b_1:

$$
\begin{aligned}
&\frac{\partial Q}{\partial b_1} = \frac{\partial(\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2)}{\partial b_1} = 0 \\
&2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(-X_i) = 0 \\
&\sum_{i=1}^{n}(-X_i Y_i + b_0 X_i + b_1 X_i^2) = 0 \\
&\sum_{i=1}^{n} -X_i Y_i + \sum_{i=1}^{n} b_0 X_i + \sum_{i=1}^{n} b_1 X_i^2 = 0 \\
&\sum_{i=1}^{n} -X_i Y_i + \sum_{i=1}^{n}(\bar{Y} - b_1 \bar{X}) X_i + \sum_{i=1}^{n} b_1 X_i^2 = 0 \\
&\sum_{i=1}^{n} -X_i Y_i + \sum_{i=1}^{n} \bar{Y} X_i - \sum_{i=1}^{n} b_1 X_i \bar{X} + \sum_{i=1}^{n} b_1 X_i^2 = 0 \\
&-\sum_{i=1}^{n} b_1 X_i \bar{X} + \sum_{i=1}^{n} b_1 X_i^2 = \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} \bar{Y} X_i \\
&b_1(-\sum_{i=1}^{n} X_i \bar{X} + \sum_{i=1}^{n} X_i^2) = \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} \bar{Y} X_i \\
&b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}
\end{aligned}
\tag{14}
$$

Uma forma de avaliar a eficiência do modelo de regressão linear é através do score R^2, que possui um valor entre 0 e 1, sendo este diretamente proporcional a sua eficácia [16]. Um jeito de usar este score é calculá-lo utilizando os dados de treinamento.

$$
R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{15}
$$

1.8 **Regressão logísitica**

A regressão logística é uma técnica de aprendizado de máquina supervisionado para classificação binária. Assim, a regressão logística é usada para prever um valor númerico inteiro, contido em um conjunto discreto de dois elementos, normalmente {0,1}, a partir de uma função sigmóide que representa o modelo.[15]

O modelo da regressão logística é representado por uma sigmóide cujo expoente do número de Euler é uma função da soma de outras variáveis, multiplicadas por seus respectivos coeficientes, e de uma constante.[17] Os valores gerados por essa função se encontrarão em um intervalo entre 0 e 1, e o resultado será determinado comparando o valor obtido pela sigmóide com um valor limite, geralmente 0,5, e se estiver acima ou

for igual ao valor limite a classificação será 1, e abaixo será 0.

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \tag{16}$$

Uma métrica de eficácia dos resultados de uma classificação, como a regressão logística, é a acurácia. A acurácia mede quantidade relativa de acertos sobre a quantidade total de previsões, logo quanto maior a acurácia melhor será a avaliado o modelo segundo esta métrica.[18]

$$Acurácia = \frac{\#acertos}{\#total} \tag{17}$$

A escolha de variáveis preditoras normalmente é feita utilizando os valores de correlação entre variáveis. A correlação linear pode ser medida através do coeficiente de pearson. Esta correlação varia entre -1 e 1, de modo que quanto mais próximo seu valor absoluto seja de 1, maior será a tendência de que a relação entre as duas variáveis analisadas seja linear.[19] Para confirmar que a relação entre duas varáveis seja linear, recomenda-se que, além da verificação do valor de pearson, seja feita uma análise de um gráfico de dispersão entre estas variáveis.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{18}$$

Quanto maior o módulo do coeficiente de correlação entre duas variáveis mais pode-se prever uma destas usando a outra. Em modelos preditivos de uma variável, usa-se outras variáveis para prever seu valor, desejando-se assim, que os coeficientes de correlação entre a variável predita e cada uma das suas variáveis preditoras seja de módulo próxima igual a 1, e que os módulos dos coeficientes de correlação entre as variáveis preditoras umas com as outras não sejam muito próximo de 1.

1.9 **Ambiente de programação**

Uma aplicação web, assim como a desenvolvida neste trabalho, é implementada em pelo menos uma linguagem de programação, testada primeiramente em um servidor local, e depois, seu código é armazenado em um servidor web de modo que fique acessível para determinado público por meio de um endereço web, URL, acessado pelo navegador.

Para esta aplicação a linguagem de programação utilizada foi o Python, devido a sua versatilidade e eficiência na área de Ciência de Dados. Toda a parte do trabalho referente a aplicação web foi carregada para um diretório no Github. A partir do código carregado no GitHub, o serviço do Heroku foi usado para implantar e manter a execução da aplicação na web. A hospedagem do código no GitHub é opcional, poderia ser feita direta para o Heroku, mas desta forma o código não estaria público para visualização.

1.9.1 Linguagem Python

A linguagem Python é uma linguagem de programação multiparadigma e interpretada que tem sido muito popular e considerada de fácil aprendizado. Quanto ao campo da ciência de dados, o Python tem se apresentado como a linguagem mais utilizada, sendo seguida por R, Julia e algumas outras.[20]

Alguns dos tipos primitivos de dados na linguagem python são int64, float64, str e bool.[21] O tipo int64 representam números inteiros. O tipo float64 representam números reais, aqueles com uma parte decimal além da inteira. O tipo str, ou string, representa cadeias de um ou mais caracteres, usadas para letras, palavras, frases e textos. O tipo bool representa um valor verdadeiro (True) ou falso (False).

Os principais módulos, ou bibliotecas, da linguagem Python para Ciência de dados são pandas, numpy, matplotlib, seaborn e scikit-learn.[22]

O numpy oferece estruturas de dados do tipo numpy arrays, que são vetores n-dimensionais semelhantes as listas nativas do Python, mas com desempenho muito superior.[23, 24] O pandas oferece, principalmente, duas estruturas de dados importantes e muito utilizadas neste trabalho, que são as series e dataframes.[25] As séries são um tipo de vetor unidimensional indexado por um índice que não precisa ser necessariamente numérico.[26] Os dataframes são estruturas de dados bidimensionais semelhantes a uma planilha com rótulos para linhas e colunas, de modo que, cada linha e cada coluna são series.[27] Os dados do tipo str, em dataframes ou series, são definidos como sendo do tipo object. Os valores de series e dataframes são estruturas do tipo numpy array.

O matplotlib, principalmente graças o seu submódulo pyplot, é o módulo python fundamental para se trabalhar com gráficos e visualizações, e possui uma fácil integração com o pandas e o numpy.[28] O seaborn é um módulo de visualização derivado do matplotlib que é conhecido pela polidez de seus gráficos e compactação de código.[29] Apesar do fato que o matplotlib possa gerar todos, ou quase todos os gráficos que o seaborn, este último tem uma preferência por grande parte da comunidade, ainda assim ambos costumam ser usados juntos. Neste caso, o matplotlib costuma ser usado para gerar e personalizar as propriedades das figuras em si, enquanto que o seaborn é responsável pelo gráficos plotados nestas figuras.

O scikit-learn é um módulo para aprendizado de máquina, e contém implementações de algoritmos de regressão, classificação, agrupamento, seleção de modelos e redução de dimensionalidade.[30] Para este trabalho, foram considerados as implementações da regressão linear e da regressão logística.

O streamlit é um módulo python para criação de aplicativos web personalizado de ciência de dados e aprendizado de máquina.[31] O streamlit possui suporte amplo aos outros módulos python aqui citados, e é relativamente mais fácil de criar aplicativos deste tipo do que seus concorrentes. No streamlit, tanto o front-end quanto o back-end

são em código Python, podendo ser feitos nos mesmos arquivos.

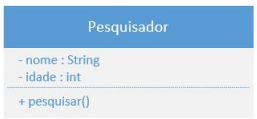
1.9.2 Orientação a objetos

A orientação objeto é um paradigma de programação muito utilizado na construção de softwares comerciais, aplicações web e mobile. A modelagem de um software orientado a objetos pode ser feita a partir de uma descrição textual do problema ou da solução, a princípio, cada palavra substantivada seria considerada uma classe, cada palavra adjetivada seria um atributo, e cada ação seria um método.[19] Uma classe é um tipo robusto de dados que possui atributos, que são variáveis, e métodos, que são funções que representam ações ou comportamentos. Um objeto pode ser uma variável que não é do tipo primitivo, mas de uma classe.

A UML é uma linguagem gráfica de projeto de software, que possui diversos tipos de diagramas, dentre eles, o diagrama de classes.[32] Em UML, uma classe é representada por um retângulo dividido em três partes, a superior que contém o nome da classe, a do meio que contém a indicação dos atributos, e a inferior que contém a indicação de métodos. Em um diagrama de classes, as classes podem estar ligadas por uma linha simples, que indica relacionamentos como a posse e uso de um objeto de uma classe por outra classe.[33]

Um exemplo de classe seria o de pesquisador, considerando suas características de nome e idade, possuiria dois atributos com estes nomes, e realizaria ações de pesquisa, modelada pelo método pesquisar.

Figura 4 – Exemplo - Classe Pesquisador



Fonte: O Autor (2020).

1.9.3 Repositórios e servidores

O GitHub é uma plataforma online para repositórios de projetos de programação com algumas funcionalidades de rede social.[34] O GitHub funciona junto com o sistema de controle de versão Git, permitindo também a colaboração e compartilhamento de código.

O Heroku é uma plataforma para armazenamento de aplicações em nuvem com suporte amplo de linguagens.[35] Através do Heroku, uma aplicação web pode ser implantada e mantida, necessitando de uma conta e a ações de carregamento do código e pequenas configurações.

2 DESENVOLVIMENTO

Este capítulo descreve o desenvolvimento da aplicação web para análise de dados referente a este trabalho. Primeiramente, é feita a descrição do código-fonte criado, de como foi feita a estruturação do software e aplicação da linguagem Python e seus módulos. Ao final, há a descrição da implementação de alterações para deploy da aplicação em um servidor web.

2.1 Ferramentas

As ferramentas computacionais utilizadas para o desenvolvimento e teste do software foram uma IDE, uma linguagem de programação e seus pacotes, e um navegador web.

A linguagem de programação Python utilizada foi a 3.8.5. A IDE(Integrated Development Environment), ou Ambiente de Desenvolvimento Integrado, utilizado para editar e executar o código-fonte foi o PyCharm Professional Edition 2020.2.3. O navegador utilizado foi o Google Chrome versão 87.0.4280.88 de 64 bits. Os pacotes Python utilizados foram adicionados e executados em um ambiente virtual isolados das versões instaladas no sistema operacional.

Tabela 3 – Módulos python utilizados e suas versões

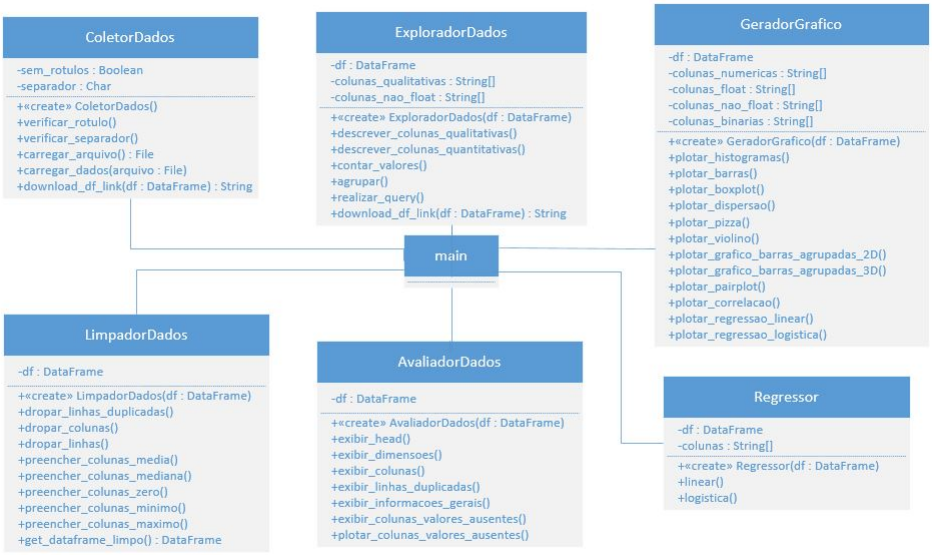
Módulo	Versão
numpy	1.19.4
pandas	1.1.4
matplotlib	3.3.2
seaborn	0.11.0
streamlit	0.71.0
scikit-learn	0.23.2

Fonte: O Autor (2021).

2.2 Código-Fonte

O código-fonte do aplicativo foi estruturado de forma que o arquivo main.py é o centro do programa, sendo responsável por chamar as demais funções, ou métodos, delegadas as outras partes do projeto, além de conter a implementação da barra laterais. Cada etapa do processo de análise de dados possui uma classe correspondente com métodos que implementam suas funcionalidades desejadas.

Figura 5 – Diagrama de classes da aplicação



Fonte: O Autor (2020).

A classe ColetorDados é responsável pelo upload do arquivo que contém o conjunto de dados e por sua transformação em um objeto do tipo dataframe. O método "verificar_rotulo()"exibe uma opção a ser marcada quando o arquivo não possui rótulos na primeira linha. O método "verificar_separador()"exibe uma caixa de seleção que indica o caracter que separa os dados no arquivo, sendo vírgula a opção padrão, e ponto e vírgula, um espaço e tabulação as demais opções. O método "carregar_arquivo()"fornece uma opção para upload de arquivo csv, tsv ou txt através de um botão que abre uma janela de exploração do sistema operacional para seleção do arquivo. O método "carregar_dados"recebe o arquivo carregado pelo método "carregar_arquivo"para transformar os dados em um objeto do tipo dataframe, usando o separador selecionado através do método "verificar_separador", e caso a opção do método "verificar_rotulo()"tenha sido marcado, os rótulos serão enumerados a partir de 1 e com prefixo "x". O método "download_df_link"recebe um dataframe como parâmetro e retorna um link para download deste dataframe como um arquivo csv.

A classe AvaliadorDados é responsável por exibir informações referentes a disposição e qualidade dos dados, de forma automática. O método "exibir_head"exibe o cabeçalho e as primeiras 10 linhas do conjunto de dados. O método "exibir_dimensoes"exibe uma frase informando as dimensões, linhas e colunhas do conjunto de dados. O método "exibir_colunas"exibe uma frase informando os nomes de todas as colunas do conjunto de dados. O método "exibir_linhas_duplicadas"exibe uma frase informando a quantidade de linhas duplicadas no conjunto de dados. O método "exibir_informacoes_gerais"exibe uma figura de uma tabela que indica o nome, tipo, quan-

tidade de valores ausentes e percentual de valores ausentes para cada coluna do conjunto de dados. O método "exibir_colunas_valores_ausentes"exibe uma frase informando os nomes das colunas do conjunto de dados que possuem valores ausentes. O método "plotar_colunas_valores_ausentes"exibe gráfico de barras horizontal que indica a proporção entre valores ausentes para colunas que os possuem.

A classe LimpadorDados é responsável por possibilitar a limpeza do conjunto de dados, resolvendo os problemas de qualidade de dados identificados conforme opções desejadas pelo usuário. Os métodos desta classe são "dropar_linhas_duplicadas", "dropar_colunas", "dropar_linhas", "preencher_colunas_media", "preencher_colunas_mediana", "preencher_colunas_zero", "preencher_colunas_minimo", "preencher_colunas_maximo", e e suas repectivas ações são, remoção de linhas duplicadas, remover colunas, remover linhas com valores ausentes, preencher os valores ausentes de colunas com suas médias, medianas, zero, mínimo e máximo. Apenas o método "dropar_linhas_duplicadas"é executado automaticamente, os demais são executados de acordo com as opções de colunas escolhidas pelo usuário em caixas de seleção múltipla.

A classe ExploradorDados é responsável pela análise exploratória dos dados limpos, fornecendo descrições das colunas, exibição de contagem de valores e agrupamento do conjunto de dados por colunas selecionadas, filtragem por um consulta(query) digitada. O método "descrever_colunas_qualitativas"faz a descrição de colunas não numéricas exibindo uma tabela que apresenta os nomes destas colunas, os seus tipos e suas quantidades de valores únicos. O método "descrever_colunas_quantitativas"descreve colunas numéricas exibindo uma tabela que, há para cada coluna, as medidas de quantidade, média, desvio padrão, mínimo, quartil 1 (25%), mediana - quartil 2 (25%), quartil 3 (75%) e máximo.

Ainda em relação a classe ExploradorDados, o método "agrupar"faz o agrupamento do conjunto de dados por determinadas colunas, selecionadas em uma caixa de seleção múltipla pelo usuário, colocando-as como índices de linhas em uma nova tabela, na qual, as colunas serão as mesmas medidas usadas na descrição de colunas numéricas para as demais colunas do conjunto original. O método "realizar_query"recebe uma consulta em um campo de texto a ser preenchido pelo usuário e exibe o conjunto de dados resultantes desta consulta logo abaixo. O texto da consulta deve ser um sentença lógica usando os nomes de colunas, operadores aritméticos (+, -, *, /, %), operadores relacionais (<, >, ==, <>) e operadores lógicos (&, |). O método "contar_valores", para as colunas selecionadas em uma caixa de seleção múltipla pelo usuário, exibe tabelas de uma linha de células, cujas colunas reprentam cada valor único e cada célula sua respectiva quantidade no conjunto de dados. O método "download_df_link"é semelhante ao método homônimo em ColetorDados, mas desta vez só é usado internamente por outros métodos da mesma classe, o "agrupar"e "realizar_query".

A classe GeradorGrafico é responsável por exibir as opções de seleções de colunas para cada tipo de gráfico e seus resultados plotados em figuras correspondentes. Ao se criar um objeto desta classe, os seus atributos colunas_numericas,. colunas_float, colunas_nao_float são inicializados contendo nomes das colunas que são numéricas (int64 e float64), com valores de números reais (float64), e que não estão com valores de números reais (float64), respectivamente. Os gráficos unidimensionais, que usam apenas uma coluna do conjunto de dados, são histogramas, de barras, de pizza, boxplot que são tratados pelos métodos, "plotar_histogramas", "plotar_barras", "plotar_pizza", "plotar_boxplot". Os gráficos de dispersão, violino, de barras agrupadas 2D, de barras agrupadas 3D e pairplot são referentes aos métodos gráficos de "plotar_dispersao", "plotar_violino", "plotar_barras_agrupadas_2D", "plotar_barras_agrupadas_2D"e "plotar_pairplot". No método "plotar_correlacao"é exibido um mapa de calor com os valores das correlações entre as colunas numéricas, não havendo nenhuma opção de seleção de tais colunas. Os métodos "plotar_regressao_linear"e "plotar_regressao_logistica"plotam gráficos semelhantes aos de dispersão, mas com retas de regressão linear e curva de função sigmoide respectivamente, para cada par de variáveis selecionados.

A classe Regressor será responsável por exibir as opções referentes as colunas utilizadas para variáveis preditoras e variável predita. Ao criar um objeto do tipo Regressor, o atributo df conterá apenas as colunas numéricas e linhas com valores preenchido a partir do dataframe passado como parâmetro, o atributo colunas serão as colunas do atributo df. O método "linear"será correspondente a regressão linear, e o método "logistica"a regressão logística. Em ambos os métodos, será gerada uma caixa de seleção múltipla para as variáveis preditoras, uma caixa de seleção para a variável predita, e um botão para treinar a regressão correspondente. Como resultados serão exibidos os coeficientes de regressão dispostos em uma tabela e o score R2 do modelo em relação aos dados de treinamento.

2.3 **Deploy**

Após o término da construção do aplicativo e testes manuais executados localmente no computador do autor, o aplicativo deve ser colocado para execução na web para acesso público.

Para que o aplicativo seja devidamente implantado através do Heroku, 3 arquivos adicionais foram criados: requirements.txt, setup.sh e Procfile. O arquivo requirements.txt contém as correspondências de cada módulo pytthon utilizado no projeto com sua respectiva versão. O arquivo setup.sh é um shell script que cria um subdiretorio chamado de streamlit, onde serão armazenados dois arquivos de extensão .toml, credentials.toml que contém o e-mail do autor e da conta de usuário do Heroku, e config.toml que indica configurações de portas de conexão usada. O arquivo Procfile

não possui extensão, seu conteúdo indica que o escopo do aplicativo é web, que o arquivo setup.sh deve ser executado como shell script e que o arquivo python main.py deve ser executado pelo comando "streamlit run".

Primeiramente o código foi carregado para um diretório no GitHub do autor[36], servindo como portfólio e armazenamento intermediário para a implantação. Então foi feito deploy do aplicativo usando o Heroku, com o nome de "analisador-dados"para o aplicativo e o upload de código de forma manual através do diretório do GitHub mencionado. Finalmente, o aplicativo pode ser acessado por um link[37].

3 **RESULTADOS**

Neste capítulo serão exibidos os resultados da execução da aplicação para um determinado conjunto de dados, incluindo tabelas e gráficos. Algumas limitações e particularidades também são mostradas para este caso específico. Durante a execução do aplicativo, foi usado o modo de visualização padrão, e não o modo wide.

3.1 **O conjunto de dados utilizado**

O conjunto de dados é uma amostra dos dados dos passageiros do titanic, fornecido pelo Kaggle[38] em uma de suas páginas de competições, a chamada "Titanic - Machine Learning from Disaster"[39].

As colunas, ou variáveis, do conjunto de dados são survival, pclass, sex, age, sibsp, parch, ticket, fare, cabin, embarked. Survival indica se o passageiro sobreviveu, sendo 0 para não e 1 para sim. Pclass indica a classe da passagem, sendo 1, 2 ou 3. Sex indica o gênero do passageiro, sendo "male"para masculino e "female"para feminino. Age indica a idade do passageiro em anos. Sibsp indica o número de irmãos/cônjuges a bordo. Parch indica o número de pais/filhos a bordo. Ticket indica o número do bilhete. Fare indica o preço da passagem. Cabin indica o número da cabine. Embarked indica o porto de embarque, C para Cherbourg, Q para Queenstown e S para Southampton.

3.2 **Exemplo de análise de dados usando o aplicativo**

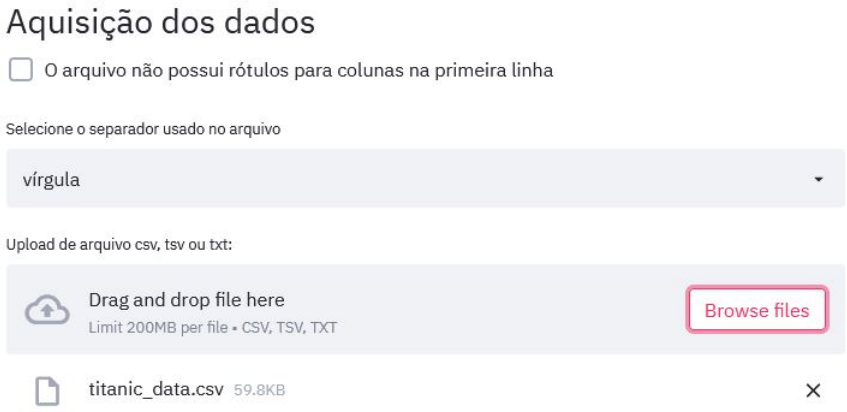
Ao acessar o aplicativo pelo endereço de URL[37], uma tela inicial será carregada no navegador. A tela inicial contém um título, nome do autor e uma breve descrição, além da seção de aquisição dos dados.

Figura 6 – Tela inicial do aplicativo

Fonte: O Autor (2020).

Para este conjunto de dados, na etapa de aquisição dos dados, a opção que pergunta se o conjunto de dados não possui rótulos para colunas na primeira linha deve permanecer desmarcada. O separador selecionado deve permanecer como sendo vírgula. Para realizar o upload do arquivo contendo o conjunto de dados, o botão "Browser Files"foi acionado, e o arquivo foi selecionado pela nova janela de navegação aberta.

Figura 7 – Etapa aquisição dos dados



Fonte: O Autor (2020).

No início da etapa de avaliação dos dados, é exibido uma amostra com as 10 primeiras linhas e o cabeçalho do conjunto de dados, e outras informações relacionadas.

O conjunto de dados foi medido com 891 linhas e 12 colunas, e sem nenhuma linha duplicada. Os nomes das colunas são passengerid, survived, pclass, name, sex, age, sibsp, parch, ticket, fare, cabin e embarked, das quais, cabin, age e embarked possuem valores ausentes, ou seja, com valores NaN ("Not a Number").

Figura 8 – Início da avaliação dos dados

Avaliação dos dados

O cabeçalho e as primeiras linhas do dataset são:

	passengerid	survived	pclass	name	sex	age	sibsp
0	2	1	1	Cumings, Mrs. John Bra…	female	38	1
1	3	1	3	Heikkinen, Miss. Laina	female	26	0
2	4	1	1	Futrelle, Mrs. Jacques…	female	35	1
3	5	0	3	Allen, Mr. William Hen…	male	35	0
4	6	0	3	Moran, Mr. James	male	NaN	0
5	7	0	1	McCarthy, Mr. Timothy J	male	54	0
6	8	0	3	Palsson, Master. Gosta…	male	2	3
7	9	1	3	Johnson, Mrs. Oscar W …	female	27	0
8	10	1	2	Nasser, Mrs. Nicholas …	female	14	1
9	<						>

As dimensões do dataset são 891 linhas e 12 colunas.

As colunas são: passengerid, survived, pclass, name, sex, age, sibsp, parch, ticket, fare, cabin, embarked.

A quantidade de linhas duplicadas é 0.

As colunas com valores ausentes são: cabin, age, embarked.

Fonte: O Autor (2020).

Ainda na avaliação dos dados, é exibida uma tabela cujas colunas são "Colunas", "Tipo", "Valores Ausentes"e "Percentual Faltante", que representa o nome de cada coluna do conjunto de dados com as suas respectivas informações de tipo (int64, float64 ou object), quantidade de valores ausentes e percentual de valores ausentes.

A contagem absoluta e relativa de valores ausentes maiores que zero, são, respectivamente, 177 e 19,90% para "age", 687 e 77,10% para "cabin", 2 e 2% para "embarked".

Figura 9 – Informações gerais das colunas

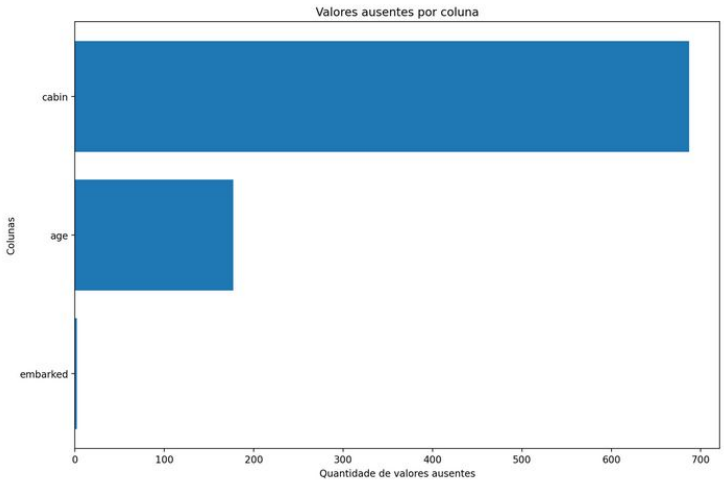
Informações gerais das colunas

	Coluna	Tipo	Valores Ausentes	Percentual Faltante
0	passengerid	int64	0	0
1	survived	int64	0	0
2	pclass	int64	0	0
3	name	object	0	0
4	sex	object	0	0
5	age	float64	177	0.1990
6	sibsp	int64	0	0
7	parch	int64	0	0
8	ticket	object	0	0
9	fare	float64	0	0
10	cabin	object	687	0.7710
11	embarked	object	2	0.0020

Fonte: O Autor (2020).

Através do histograma, pode-se acentuar visualmente a quantidade de valores ausentes para as colunas. A coluna "embarked"possui poucos, e quase nenhum valor nulo, enquanto "age"possui uma quantidade consideravelmente importante, e "cabin"possui uma quantidade extremamente grande em comparação as outras e também ao número total de linhas do conjunto de dados.

Figura 10 – Histograma horizontal de valores ausentes

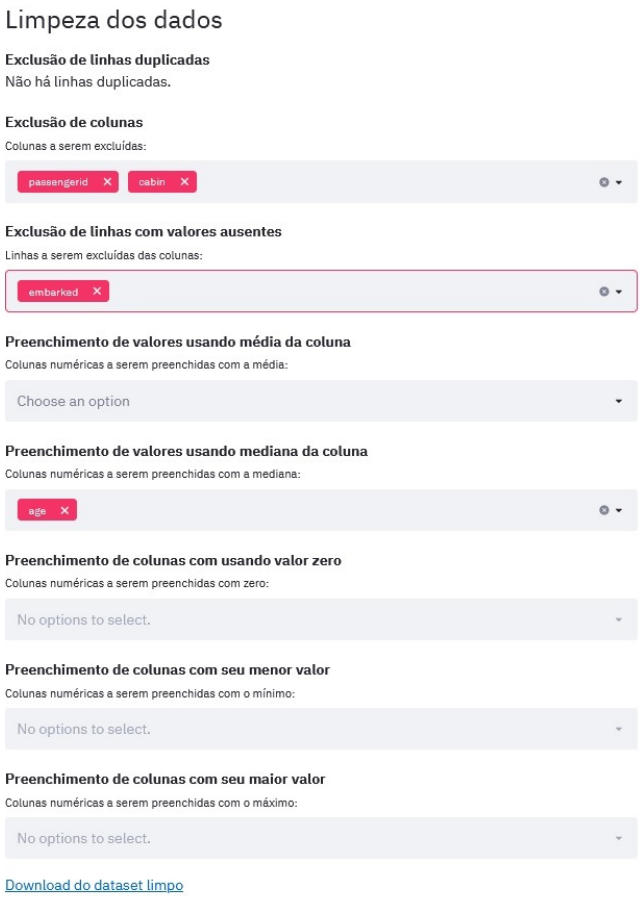


Fonte: O Autor (2020).

Na etapa de limpeza dos dados, os problemas de ausência de alguns valores

são tratadas. A coluna "cabin"foi excluída por apresentar uma quantidade muito grande de valores ausentes. A coluna "passengerid"também foi excluída, porque, como seu nome indica, apresenta apenas informação do id do passageiro, que não será relevante para análise adiante. A coluna "embarked"possui valores ausentes apenas em duas linhas, que são excluídas. Os valores ausentes para a coluna "age"são preenchidos com a mediana, um valor que não deve prejudicar os passos seguintes.

Figura 11 – Opções de limpeza dos dos dados



Fonte: O Autor (2020).

Analisando a tabela de descrição de colunas não numéricas, percebe-se que cada passageiro possui um nome distinto, há três regiões de embarque, 680 passagens e dois gêneros definidos para os passageiros.

Figura 12 – Descrição colunas não numéricas

Análise Exploratória dos dados

Descrição das colunas não numéricas

	Coluna	Tipo	Valores Únicos
0	name	object	889
1	sex	object	2
2	ticket	object	680
3	embarked	object	3

Fonte: O Autor (2020).

Quanto a coluna "age", ou seja, idade em anos, o valor médio é aproximadamente 29, a pessoa mais jovem registrada tem menos de um ano, e a mais velha 80. Quanto a coluna "fare", o valor médio é aproximadamente 32,1, e o maior valor 512,3292. As colunas "survived"e "pclass", possui valores discretos entre 0 e 1, ou 1 e 3, respectivamente. Analisando "sibsp"e "parch", pelo menos, 50% e 75% dos passageiros não possuem respectivamente, nenhum irmãos/cônjuge e de pais/filhos a bordo.

Figura 13 – Descrição colunas numéricas

Descrição de colunas numéricas

	survived	pclass	age	sibsp	parch	fare
quantidade	889	889	889	889	889	889
média	0.3825	2.3116	29.3152	0.5242	0.3825	32.0967
desvio padrao	0.4863	0.8347	12.9849	1.1037	0.8068	49.6975
mínimo	0	1	0.4200	0	0	0
quartil 1 (25%)	0	2	22	0	0	7.8958
mediana - quartil 2 (50%)	0	3	28	0	0	14.4542
quartil 3 (75%)	1	3	35	1	0	31
máximo	1	3	80	8	6	512.3292

Fonte: O Autor (2020).

Para as colunas "sex", "pclass", "survived"e "embarked"foi realizada uma contagem de valores. Quanto a "sex", 577 dos passageiros são "male"e 312 são "female". Em relação a "pclass", vê-se que 491 foram de 3, 214 de 1 e 184 para 2. Através dos valores de survived, sendo 0 para não sobrevivente e 1 para sobrevivente, percebe-se que 549 dos passageiros não sobreviveram e 340 sobreviveram. Por meio da contagem de "embarked", observa-se que o porto com maior número de embarques foi em Spara

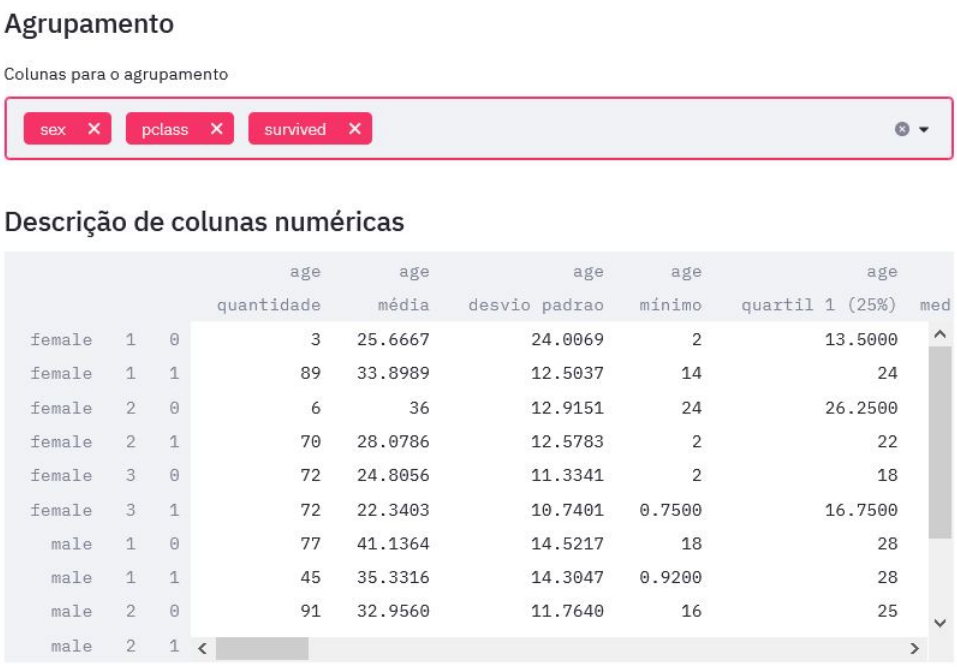
Southampton(S) com 644 passageiros, enquanto o menor foi em Queenstown(Q) com 77 passageiros, e Cherbourg(C) teve 168 passageiros.

Figura 14 – Contagem Valores

Fonte: O Autor (2020).

Foi realizado um agrupamento por "sex", "pclass"e "survived". Através deste agrupamento percebe-se que

Figura 15 – Agrupamento colunas



Fonte: O Autor (2020).

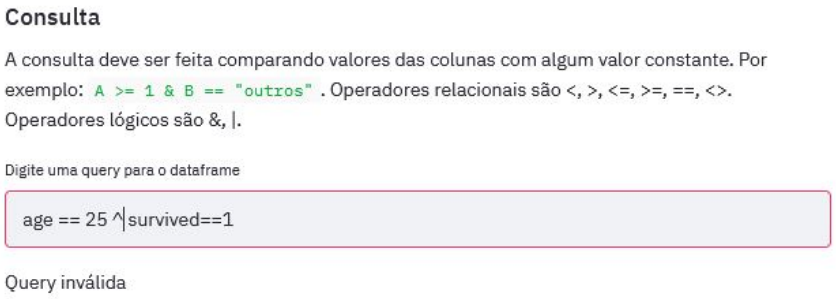
Para realizar uma consulta válida foi usada a sentença de idade igual a 25 e sobrevivente, que retorno algumas linhas.

Figura 16 – Consulta válida

Fonte: O Autor (2020).

Para uma consulta inválida, semelhante a consulta anterior, mas com o símbolo ˆ no lugar de &, não é exibido nenhum conjunto de dados ou tabela, apenas a mensagem "Query inválida".

Figura 17 – Consulta inválida



Fonte: O Autor (2020).

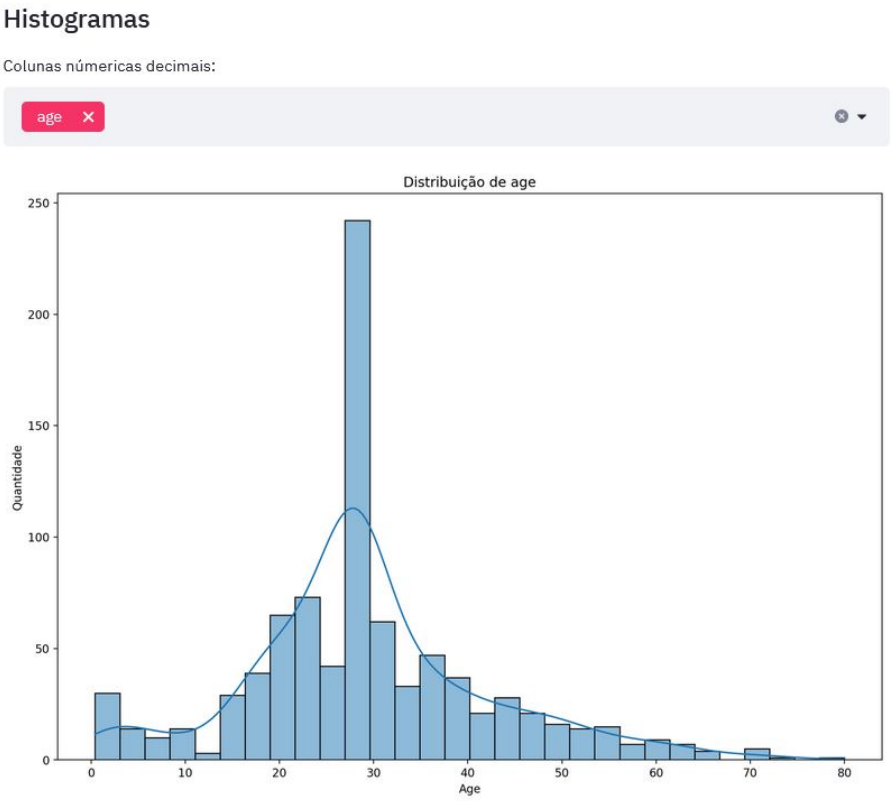
Para as etapas de visualização e regressão, todas as opções de gráficos e tipos de regressão foram marcadas na barra lateral.

Figura 18 – Barra lateral com opções marcadas

Fonte: O Autor (2020).

Um histograma foi plotado para a variável "age". A distribuição de "age"é unimodal com uma cauda à direita maior do que à esquerda. A mediana, aproximadamente entre 27 e 30 anos, apresentam uma quantidade muito grande de amostras, devido ao fato também do preenchimento de valores ausentes realizado na etapa de limpeza dos dados. A faixa de idades dos passageiros é entre 0 e 80 anos.

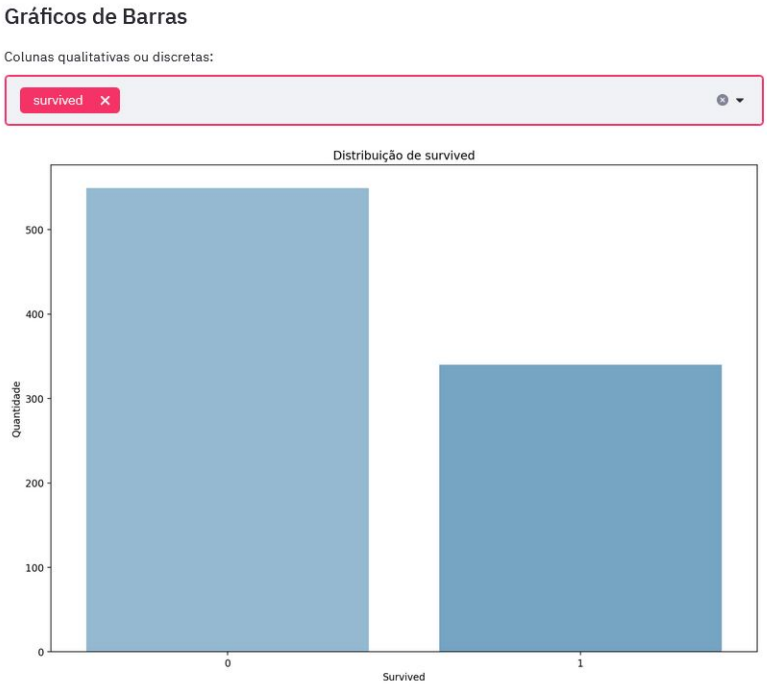
Figura 19 – Histograma de Age



Fonte: O Autor (2020).

Um gráfico de barras foi plotado para a variável "survived". Os possíveis valores para esta variável são 0 e 1, sendo 0 para não sobrevivente e 1 para sobrevivente. Há um número bem maior de não sobreviventes do que de sobreviventes. O conjunto de dados está desequilibrado em relação a esta variável, como se era esperado devido ao conhecimento do evento de naufrágio do titanic.

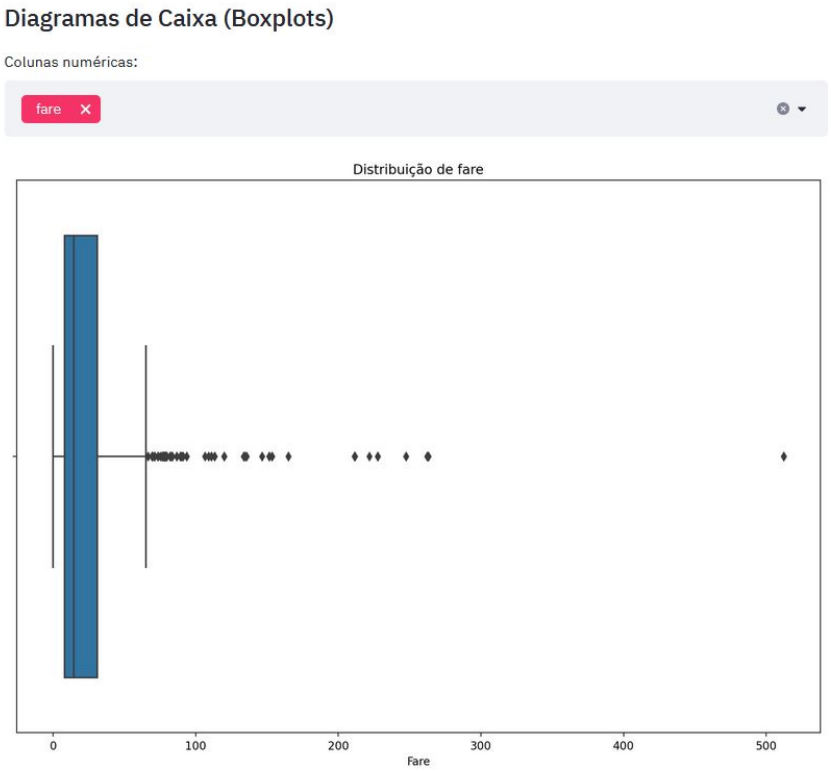
Figura 20 – Gráfico de barras para survived



Fonte: O Autor (2020).

Um diagrama de caixa, ou boxplot, foi plotado para a variável "fare". Como esperado o valor mínimo foi 0, já que alguns não pagaram. Há vários possíveis outliers.

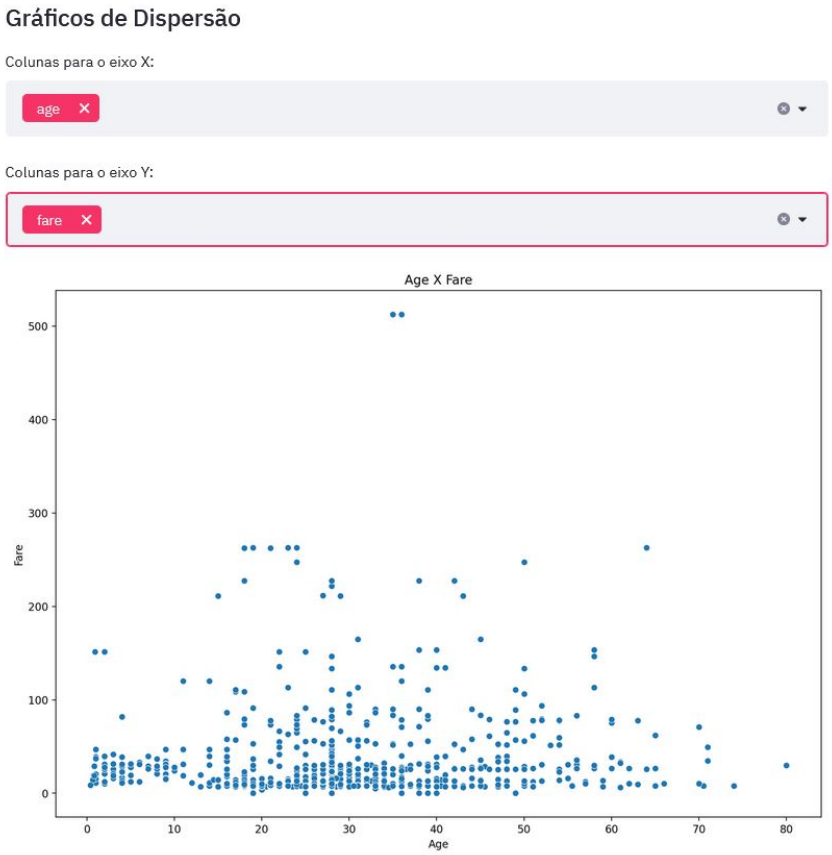
Figura 21 – Diagrama de caixa (boxplot) de fare



Fonte: O Autor (2020).

Um gráfico de dispersão, ou de pontos, com age no eixo x e "fare"no eixo y é plotado. Observando este gráfico percebe-se que não há uma forte correlação entre as duas variáveis. Os valores de "fare"parece se concentrar em sua maioria na faixa entre 0 e 100 para todas as idades amostradas.

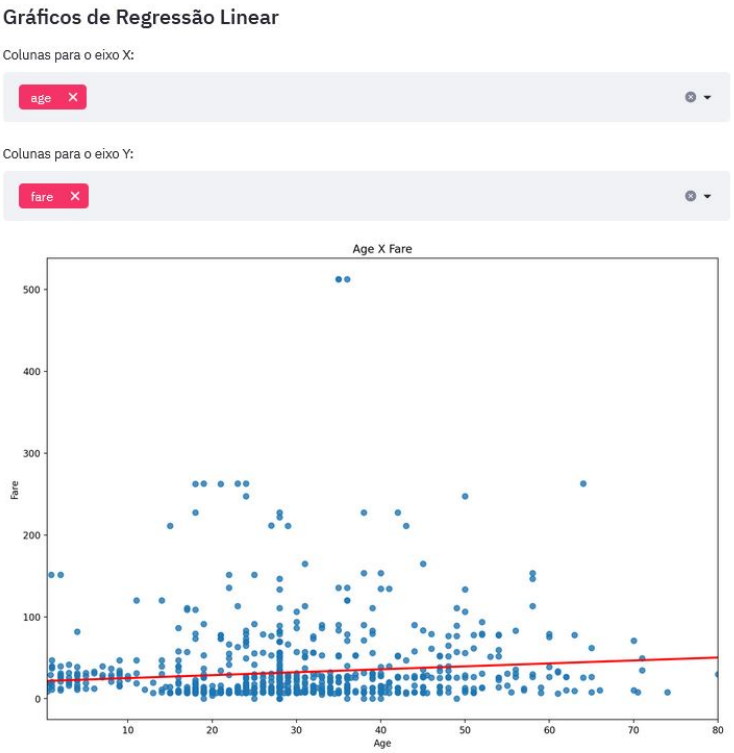
Figura 22 – Gráfico de dispersão entre age e fare



Fonte: O Autor (2020).

Um gráfico de regressão linear, com age no eixo x e "fare"no eixo y foi plotado. A reta de regressão se ajusta à área de maior concentração dos dados e mesmo assim, vários pontos se encontram distante desta. Parece haver uma correlação fraca e positiva entre estas duas variáveis.

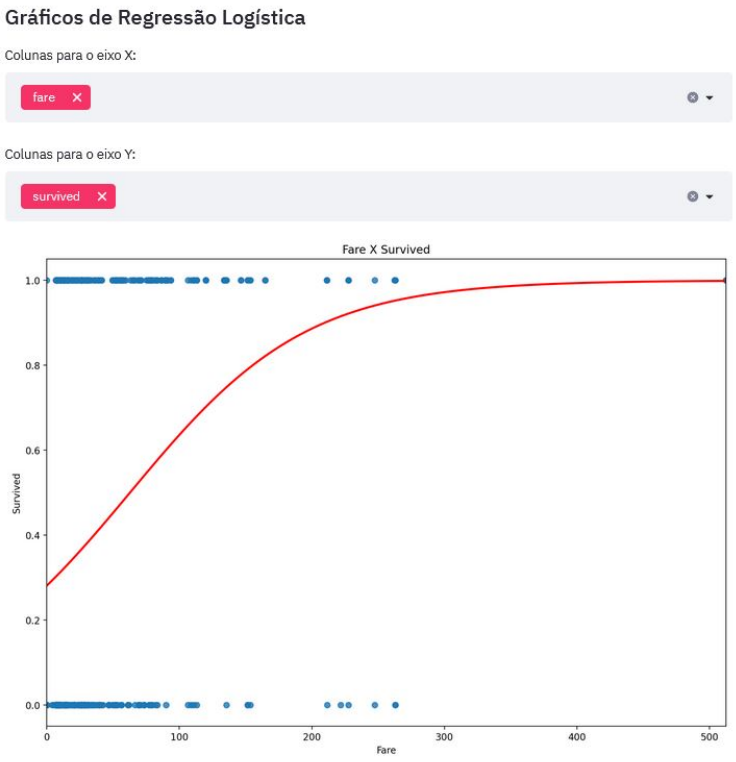
Figura 23 – Gráfico de regressão entre age e fare



Fonte: O Autor (2020).

Um gráfico de regressão logística, com "fare"no eixo x e survived no eixo y foi plotado. A reta de regressão não se ajustou bem aos dados, de modo que visivelmente não se pode separar os valores de "survived"a partir dos valores de "fare". Parece haver uma correlação fraca entre estas duas variáveis.

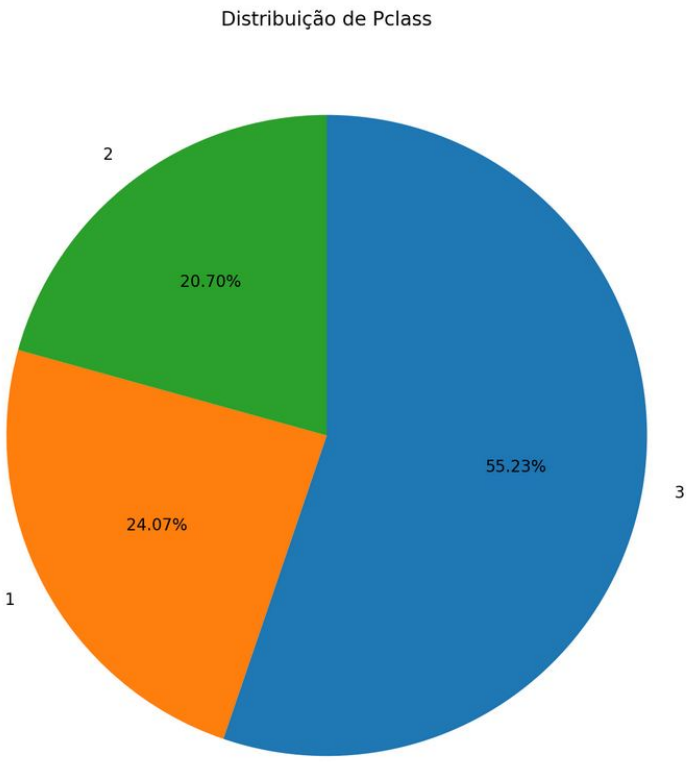
Figura 24 – Gráfico de regressão entre fare e survived



Fonte: O Autor (2020).

Um gráfico de pizza foi plotado para a variável "pclass". Através do gráfico percebe-se que mais da metade, cerca de 55,23 % dos passageiros pertencem a "pclass"3, enquanto que, 24,07% e 20,70% pertencem as "pclass"1 e 2, respectivamente.

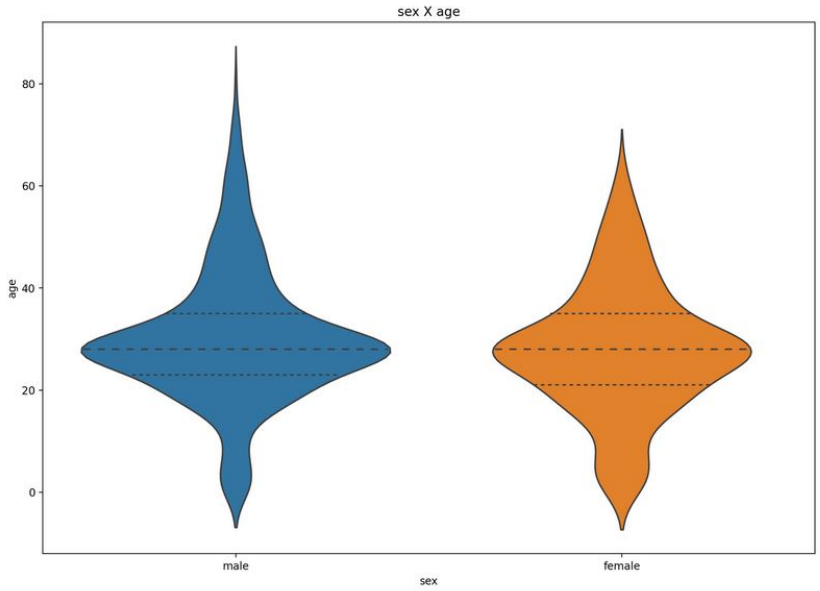
Figura 25 – Gráfico de pizza para pclass



Fonte: O Autor (2020).

Um gráfico de violino com "sex"no eixo x e "age"no eixo y foi plotado para analisar a distribuição da idade para cada um dos gêneros. As pessoas mais velhas, com aproximadamente 80 anos são do gênero masculino. Pessoas na faixa entre 0 e 15, são, em maior parte do sexo feminino.

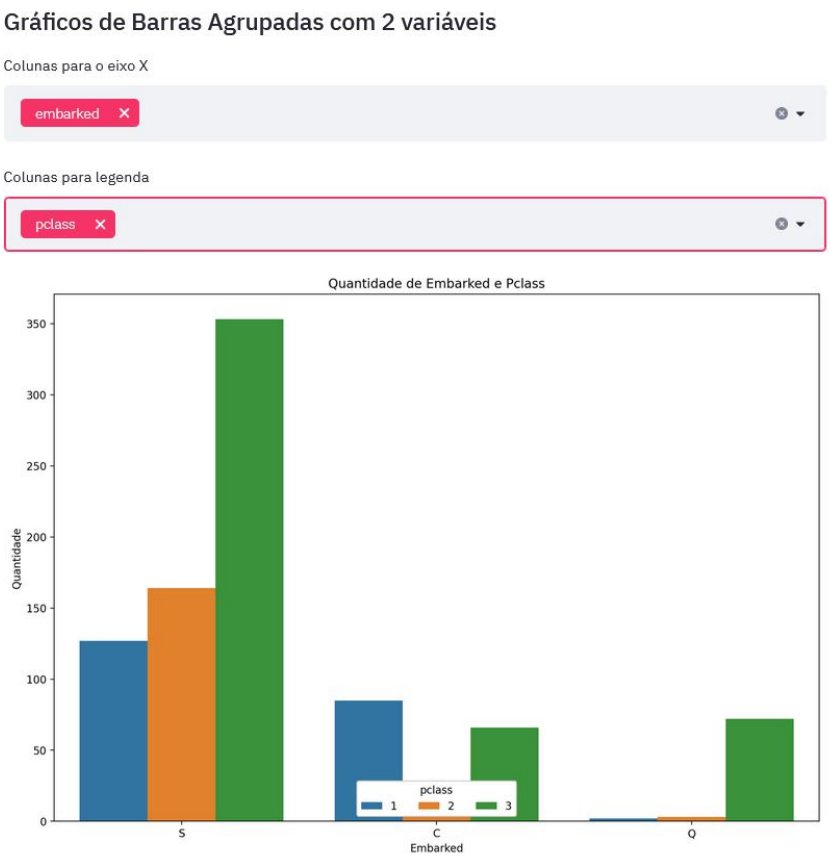
Figura 26 – Gráfico de Violino para sex e age



Fonte: O Autor (2020).

Um gráfico de barras 2D para "pclass"e "embarked"foi plotado, sendo o eixo y representado pela quantidade de passageiros. A maior parte dos passageiros embarcou em Southampton, dos quais a "pclass"3 contém mais do que as outras duas somadas. Para os que embarcaram em Cherbourg apenas uma parcela pequena foi da "pclasse"2. Para os que embarcaram em Queenstown, quase todos são de "pclass"3.

Figura 27 – Gráfico de barras 2D para pclass e embarked



Fonte: O Autor (2020).

Um gráfico de barras 3D para "pclass"e "embarked"no eixo x e fare no eixo y foi plotado. O local de embarque e "pclass"afetam efetivamente o valor de "fare". Os maiores valores de fare são da "pclass"1 e embarcaram em Cherbourg.

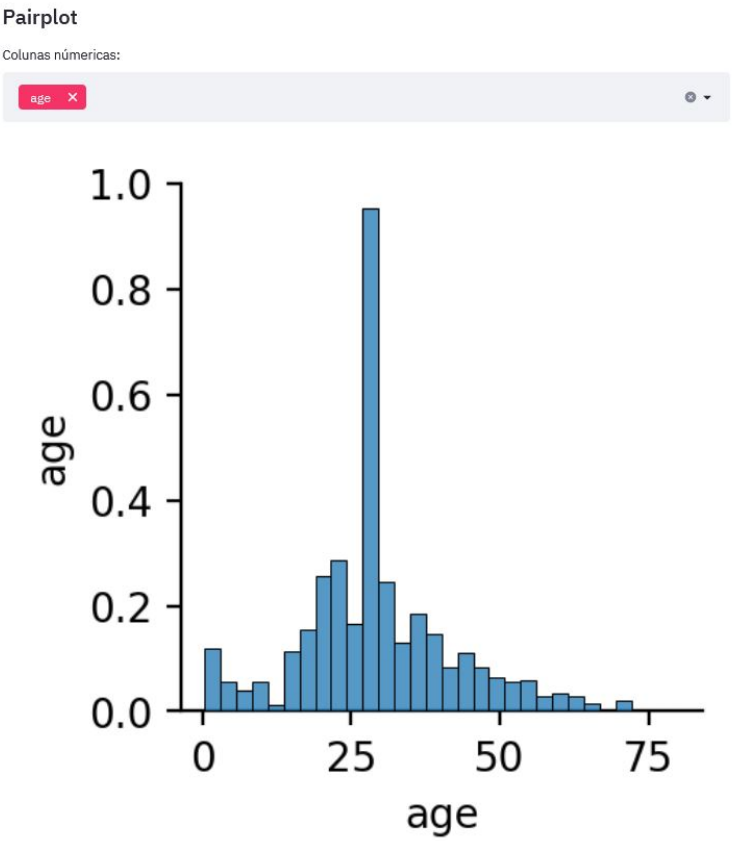
Figura 28 – Gráfico de barras 3D para pclass, embarked e fare

Fonte: O Autor (2020).

O pairplot foi realizado para três combinações diferentes, primeiro só para "age", depois para "age"e "fare", e por último para "age", "fare"e "sibsp".

O pairplot apenas com age apresenta um histograma, semelhante, mas não idêntico, ao plotado anteriormente.

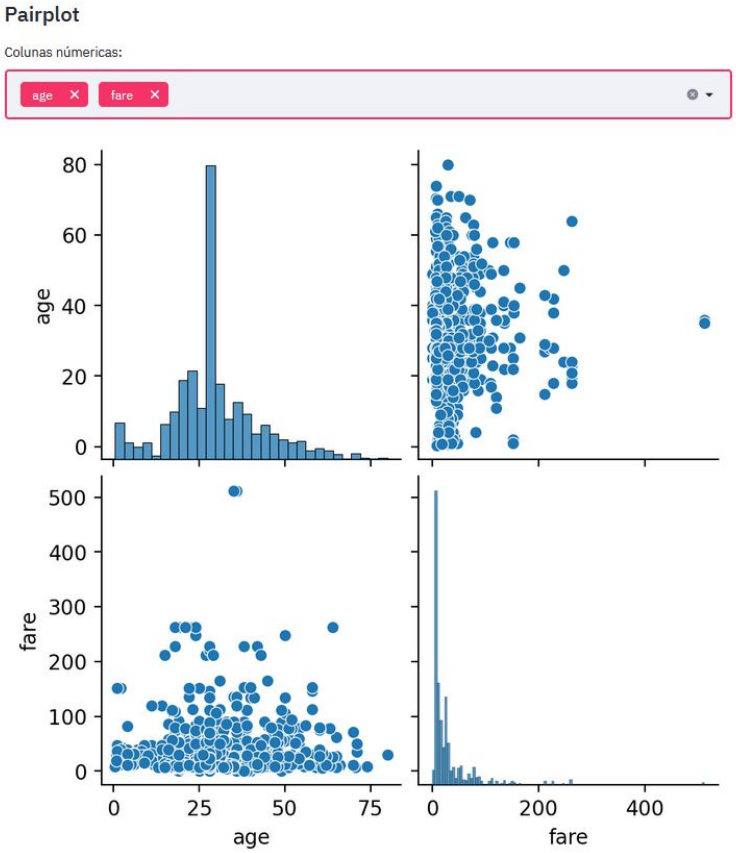
Figura 29 – Pairplot de age



Fonte: O Autor (2020).

O pairplot para duas variáveis apresenta quatro gráficos em uma figura. O primeiro gráfico da diagonal principal é o mesmo plotado anteriormente, mas em uma escala menor. O outro histograma na diagonal principalmente correspondente a "fare", indicando uma distribuição unimodal com cauda à direita. Os gráficos na diagonal secundária representam a dispersão entre "age"e "fare", sendo um deles semelhante ao que foi plotado antes, e transposto do outro. A dispersão exibe a tendência de não haver uma forte correlação entre as variáveis.

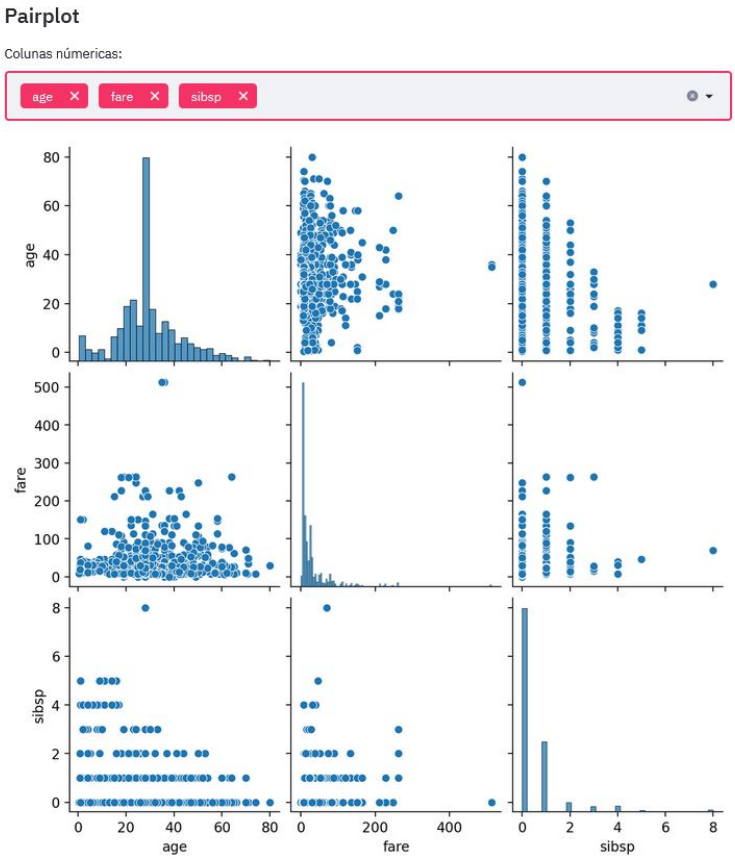
Figura 30 – Pairplot de age e fare



Fonte: O Autor (2020).

O pairplot para duas variáveis apresenta oito gráficos em uma figura. O histograma de "sibsp"indica que a variável assume valores discretos, indicando que a grande maioria dos passageiros possui nenhum ou até 1 filho a bordo. As dispersões de "sibsp"com "age"e "fare"sugerem que não haja correlação linear entre estas.

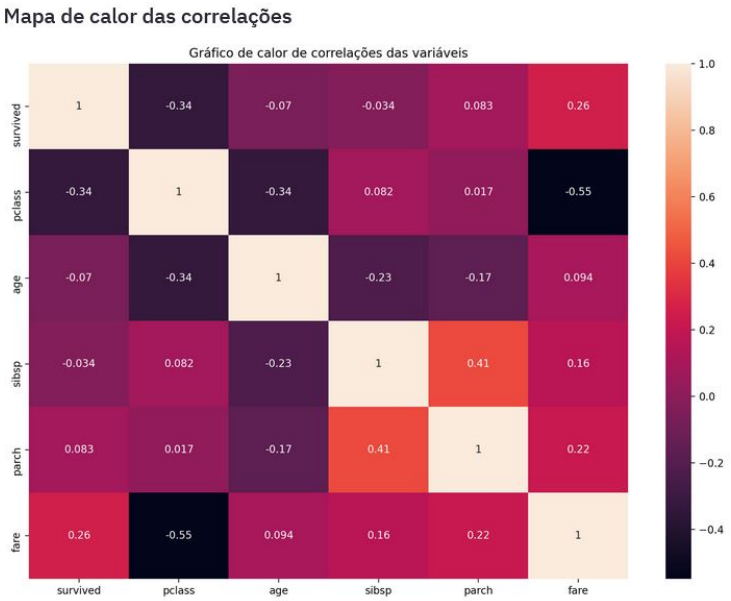
Figura 31 – Pairplot de age, fare e sibsp



Fonte: O Autor (2020).

Para exibir os valores de correlações entre as variáveis numéricas foi plotado um mapa de calor com a matriz de correlações. O menor valor em módulo exibido foi o de 0,017, que representa a força de correlação entre "parch"e "pclass". As duas correlações mais fortes se referem a "pclass"em relação a "age"e "fare", com módulos de 0,55 e 0,34, respectivamente.

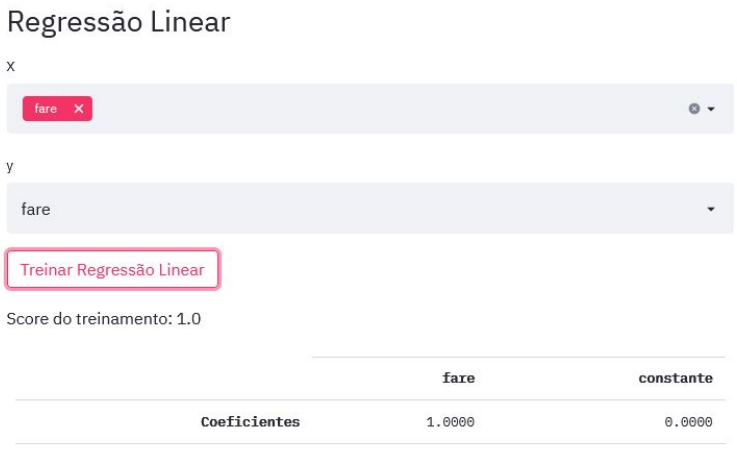
Figura 32 – Mapa de calor das correlações



Fonte: O Autor (2020).

A primeira regressão linear treinada foi apenas um teste para confirmação dos resultados, usando a variável "fare"para prever a si mesmo. Como esperado, o modelo apresentou o coeficiente 1 para "fare"e 0 para constante, e o score de treinamento foi de 1, indicando que o modelo explica 100% da variabilidade de "fare".

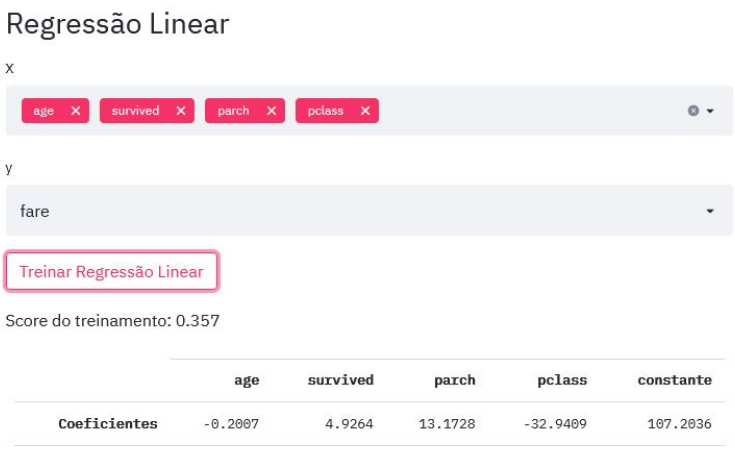
Figura 33 – Regressão linear - Fare



Fonte: O Autor (2020).

A segunda regressão linear visa prever o valor de "fare"em função de "age", "survived", "parch"e "pclass". Pelo score de treinamento de 0,357, o modelo explica 35,7% da variabilidade de "fare", que não é muito satisfatório no aspecto geral, mas é devido a força de correlação de "fare"com as outras variáveis.

Figura 34 – Regressão linear - Fare (2)

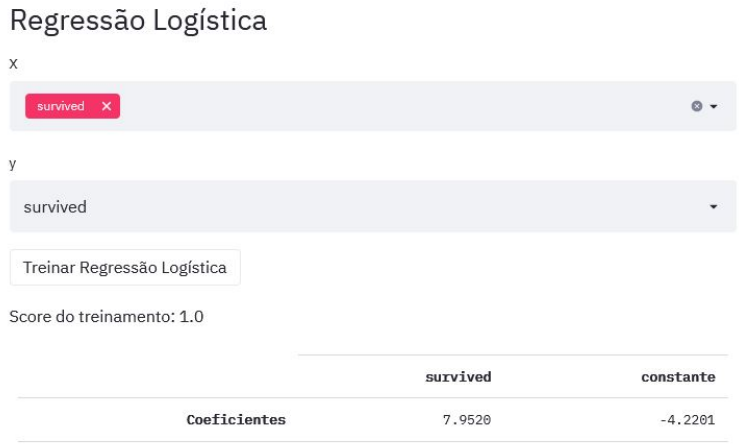


	age	survived	parch	pclass	constante
Coeficientes	-0.2007	4.9264	13.1728	-32.9409	107.2036

Fonte: O Autor (2020).

A primeira regressão logística treinada foi apenas um teste para confirmação dos resultados, usando a variável "survived"para prever a si mesmo. Como esperado, o score de treinamento foi de 1, indicando que o modelo explica 100% da variabilidade de "survived".

Figura 35 – Regressão Logística - survived (1)

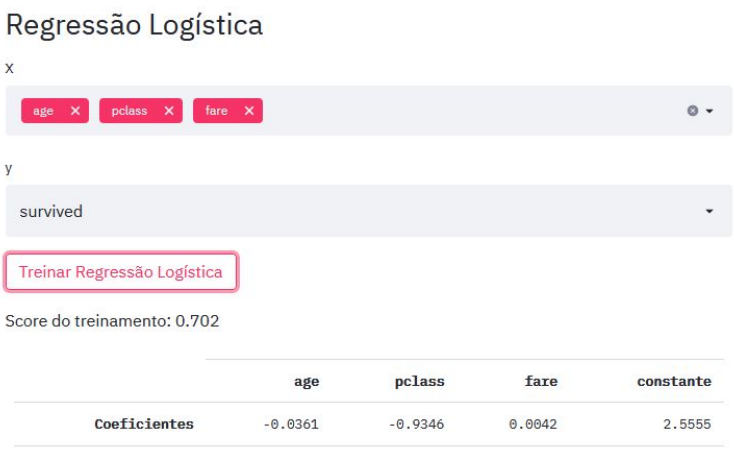


	survived	constante
Coeficientes	7.9520	-4.2201

Fonte: O Autor (2020).

A segunda regressão logística visa prever o valor de "survived"em função de "age", "pclass"e "fare". Pelo score de treinamento de 0,702, o modelo explica 70,2% da variabilidade de "survived", que é razoavelmente satisfatório no aspecto geral.

Figura 36 – Regressão Logística - survived (2)



	age	pclass	fare	constante
Coeficientes	-0.0361	-0.9346	0.0042	2.5555

Fonte: O Autor (2020).

CONCLUSÃO

A análise de dados transforma a enorme quantidade de dados adquiridos por diversos tipos de sensores em informações realmente úteis e valiosas. Analisar bem os dados gera uma grande vantagem competitiva e alavanca exponencialmente os resultados de empresas e profissionais interessados.

Este trabalho apresentou o desenvolvimento e operação de um aplicativo alternativo para realizar procedimentos básicos de análise de dados. A partir de um conjunto de dados e o acesso à internet, a ferramenta web desenvolvida propõe diversas funcionalidades ao usuário, que não precisa ter conhecimento sobre linguagens de programação e o uso do aplicativo não necessita a instalação em nenhum dispositivo.

Através dos resultados obtidos para os conjuntos de dados utilizados como exemplo neste trabalho, foi demonstrado que os aspectos teóricos e funcionalidades abordadas foram implementados com sucesso. Desta forma, o principal objetivo do trabalho foi alcançado, ou seja, o desenvolvimento de uma aplicação web para a análise de dados e a hospedagem em um servidor web para acesso público. A facilidade de uso também foi alcançada, uma vez que basta ao usuário fazer o carregamento do arquivo contendo a base de dados, e selecionar as opções desejadas para dispor das principais informações referentes ao conjunto de dados. Caso o usuário deseje analisar apenas um subconjunto dos dados iniciais, pode-se realizar uma consulta e descarregamento do subconjunto resultante e, em seguida, repetir a execução do aplicativo para este novo conjunto.

Como desdobramento desta ferramenta e possíveis trabalhos futuros, uma aplicação auxiliar que visa avaliar a eficiência dos modelos gerados é sugerida. Tal ferramenta faria uso do conceito de validação cruzada, dividindo a base de dados em treino e teste, podendo ainda otimizar parâmetros dos modelos estimados.

REFERÊNCIAS

1 UDACITY - Data Analyst Nanodegree Program. Disponível em: <https://www.udacity.com/course/data-analyst-nanodegree--nd002>. Acesso em: 2 jun. 2020.

2 INFOGRAM - O que é visualização de dados? Disponível em: <https://infogram.com/pt/pagina/visualizacao-de-dados>. Acesso em: 2 mar. 2021.

3 SCOTT, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization.* [S.l.: s.n.].

4 EVERGREEN, S. *Effective Data Visualization: The Right Chart for the Right Data.* [S.l.: s.n.], 2016. ISBN 9781506303055.

5 PORTALACTION - BoxPlot. Disponível em: <http://www.portalaction.com.br/estatistica-basica/31-boxplot>. Acesso em: 2 mar. 2021.

6 INFOGRAM - Gráfico de Pizza. Disponível em: <https://infogram.com/pt/criar/grafico-de-pizza>. Acesso em: 2 mar. 2021.

7 FM2S- O que é e para que serve o Gráfico de Dispersão? Disponível em: <https://www.fm2s.com.br/grafico-de-dispersao/>. Acesso em: 2 mar. 2021.

8 SEABORN-VIOLINPLOT. Disponível em: <https://seaborn.pydata.org/generated/seaborn.violinplot.html>. Acesso em: 2 mar. 2021.

9 SEABORN-PAIRPLOT. Disponível em: <https://seaborn.pydata.org/generated/seaborn.pairplot.html>. Acesso em: 2 mar. 2021.

10 CLEVELAND, W. S.; PRESS, H. *Visualizing Data*. [S.l.: s.n.], 1993. ISBN 0963488406.

11 TOWARDS Data Science. Better Heatmaps and Correlation Matrix Plots in Python. Disponível em: <https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec>. Acesso em: 2 mar. 2021.

12 HARRISON, M. *Machine Learning – Guia de Referência Rápida: Trabalhando com Dados Estruturados em Python*. [S.l.]: Novatec, 2020.

13 CHARNET, R. et al. *Análise de Modelos de Regressão Linear com aplicações*. [S.l.]: Unicamp, 1999.

14 FREEDMAN, D. A. *Statistical Models: Theory and Practice*. [S.l.]: Cambridge University Press, 2009.

15 MENARD, S. W. *Applied Logistic Regression*. [S.l.: s.n.], 2002. ISBN 9780761922087.

16 SCIKIT-LEARN - Linear Regression. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression>. Acesso em: 2 mar. 2021.

17 SCIKIT-LEARN - Logistic Regression. Disponível em: <https://ichi.pro/pt/qual-e-a-funcao-sigmoide-como-e-implementado-na-regressao-logistica-77981969140323>. Acesso em: 2 mar. 2021.

18 SCIKIT-LEARN - Logistic Regression. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression>. Acesso em: 2 mar. 2021.

19 UDACITY - Machine Learning Nanodegree Program. Disponível em: <https://www.udacity.com/course/machine-learning-engineer-nanodegree--nd009t>. Acesso em: 10 dez. 2020.

20 BUSINESS Over Broadway - Usage of Programming Languages by Data Scientists: Python Grows while R Weakens. 2020. Disponível em: <https://businessoverbroadway.com/2020/06/29/usage-of-programming-languages-by-data-scientists-python-grows-while-r-weakens/>. Acesso em: 2 mar. 2021.

21 EXPLORANDO TI - Tipos primitivos e suas aplicações no Python3. Disponível em: <https://www.explorandoti.com.br/tipos-primitivos-e-suas-aplicacoes-no-python3/>. Acesso em: 2 mar. 2021.

22 SAMPAIO, C. *Data Science Para Programadores. Um Guia Completo Utilizando a Linguagem Python*. [S.l.]: Ciência Moderna, 2018.

23 NUMPY - What is numpy? Disponível em: <https://numpy.org/doc/stable/user/whatisnumpy.html>. Acesso em: 2 mar. 2021.

24 OLIPHANT, T. E. *Guide to NumPy*. [S.l.: s.n.], 2006.

25 PANDAS - Docs. Disponível em: <https://pandas.pydata.org/docs/>. Acesso em: 2 mar. 2021.

26 PANDAS.SERIES. Disponível em: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.html>. Acesso em: 2 mar. 2021.

27 PANDAS.DATAFRAMES. Disponível em: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html>. Acesso em: 2 mar. 2021.

28 MATPLOTLIB. Disponível em: <https://matplotlib.org/>. Acesso em: 2 mar. 2021.

29 SEABORN. Disponível em: <https://seaborn.pydata.org/>. Acesso em: 2 mar. 2021.

30 PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

31 STREAMLIT. Disponível em: <https://streamlit.io/>. Acesso em: 2 mar. 2021.

32 LARMAN, C. *Utilizando UML e Padrões: Uma Introdução à Análise e ao Projeto Orientados a Objetos e ao Desenvolvimento Iterativo*. [S.l.]: Bookman, 2007.

33 DEVMEDIA. Orientações básicas na elaboração de um diagrama de classes. 2016. Disponível em: <https://www.devmedia.com.br/orientacoes-basicas-na-elaboracao-de-um-diagrama-de-classes/37224>. Acesso em: 3 mar. 2021.

34 HOSTINGER. O que é Github e Para Que é Usado? Disponível em: <https://www.hostinger.com.br/tutoriais/o-que-github>. Acesso em: 2 mar. 2021.

35 HEROKU - What is Heroku? Disponível em: <https://www.heroku.com/what>. Acesso em: 2 mar. 2021.

36 SIMõES, L. *UERJ-TCC-Analisador-Dados*. [S.l.]: GitHub, 2021. <https://github.com/leosimoes/UERJ-TCC-Analisador-Dados>.

37 APLICATIVO Web Analisador de Dados. Disponível em: <https://analisador-dados.herokuapp.com/>. Acesso em: 2 mar. 2021.

38 KAGGLE. Disponível em: <https://www.kaggle.com/>. Acesso em: 2 mar. 2021.

39 KAGGLE. Titanic - Machine Learning from Disaster. Disponível em: <https://www.kaggle.com/c/titanic/data>. Acesso em: 2 mar. 2021.