Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

Sistema de Apoio à Decisão na Compra de Passagens Aéreas

Trabalho de Graduação

Autor: *Rodrigo Carlos de Albuquerque Calegario*
Orientador: *Luciano Barbosa*
Área: *Ciência dos Dados*

Recife,
28 de junho de 2018

Dedico este trabalho aos meus pais, por todo suporte que me deram no período da minha graduação, ao meu irmão, que sempre confiou nas minhas competências, e à minha namorada, que me apoiou em todos os meus momentos difíceis.

Agradecimentos

Primeiramente, gostaria de agradecer a Deus que sempre me deu força e determinação para vencer todos os desafios da minha vida. Em segundo, gostaria de agradecer à minha família, por nunca duvidar da minha capacidade. Em especial, gostaria de agradecer à minha mãe, Francisca de Fátima Carlos de Albuquerque Calegario, e ao meu pai, Carlos Alberto Calegario, por terem me dado todas as condições necessárias para eu entrar na faculdade, vencer meus desafios durante o curso e colher junto a mim as vitórias alcançadas nos anos vividos na Universidade Federal de Pernambuco. Ao meu irmão, Filipe Carlos de Albuquerque Calegario, agradeço por ter me dado força em todos o momentos que considerei difíceis no curso de Ciência da Computação.

Também gostaria de agradecer minha à namorada, Maria Clara Müller de Andrade, por sempre acreditar em mim, mesmo nas horas em que nem eu acreditei. Além disso, gostaria de agradecer aos meus amigos do PET-Informática, que me proporcionaram momentos incríveis durante toda a minha graduação e que com certeza sempre estarão do meu lado. Agradeço também aos amigos que conquistei no Centro de Informática, em especial a Arthur, Danilo, Milena, Paulo, Thiago e Walber, por sempre me ajudarem a concluir todos os desafios das disciplinas cursadas.

Não poderia deixar de agradecer ao meu orientador, Luciano Barbosa, por ter me ensinado e me apoiado em todo o processo de desenvolvimento deste projeto. Agradeço também a todos o professores e funcionários do Centro de Informática, por me proporcionarem um ambiente acolhedor e enriquecedor.

Por fim, gostaria de agradecer a todos que aqui não citei, mas que fizeram total diferença em minha jornada na graduação. Sem vocês eu não teria alcançado todas as vitórias na minha vida pessoal e profissional.

O problema são problemas demais se não correr atrás da maneira certa de solucionar.

—CHICO SCIENCE

Resumo

Com o aumento da quantidade de informações disponíveis na internet, provenientes de sites de compras, ferramentas que dão suporte para a decisão do consumidor na hora da compra vêm se tornando algo cada vez mais popular. Uma área em que um sistema de apoio à decisão tem sido de grande importância é na venda de passagens aéreas. Um desafio para essas ferramentas é fornecer um ambiente de busca completo, para que o consumidor possa visualizar todas as possibilidades a fim de fazer a melhor compra de acordo com as suas necessidades. O objetivo deste trabalho é desenvolver uma ferramenta que auxilie, de forma eficiente, a busca do melhor preço de promoções de passagens aéreas, através de dados coletados em um site especializado. Nesses dados, foram aplicadas técnicas de remoção de outliers, para que, em seguida, fossem disponibilizados para uma interface de visualização e para a criação de modelos gerados a partir de métodos de aprendizado supervisionado, visando à predição de preços promocionais de passagens aéreas.

Palavras-chave: preço promocional, passagens aéreas, limpeza de dados, visualização de dados, aprendizado supervisionado, predição

Abstract

With the increasing amount of information available on the internet from shopping sites, tools that support the decision of the consumer at the time of purchase have become something increasingly popular. One area in which a decision support system has been of great importance is in the sale of air tickets. A challenge for these tools is to provide a complete search environment so that the consumer can visualize all the possibilities in order to make the best purchase according to his needs. The aim of this work is to develop a tool to efficiently assist the search for the best price of airfare promotions through data collected on a specialized website. In these data, outliers removal techniques were applied, so that they were made available to a visualization interface and to the creation of models generated from supervised learning methods, aiming the prediction of promotional prices of air tickets.

Keywords: promotional price, air tickets, data cleaning, data visualization, supervised learning, prediction

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Fazer compras na web tornou-se uma tarefa fácil e acessível para os consumidores, e realizar uma busca pelo melhor preço é uma prioridade para quem procura economizar. Nesse contexto, as companhias aéreas viram uma grande oportunidade na venda de passagens aéreas online, facilitando a vida dos consumidores e aumentando a procura por esse serviço. Porém, os preços das passagens variam muito e saber a hora certa de comprar tornou-se um desafio (Todesco, Lovadine, de Andrade Januário Bettini, & Vassallo, 2008). Muito sites oferecem ao consumidor a opção de buscar preços das passagens aéreas de diferentes companhias, mas poucos deles instruem o consumidor se os valores das passagens apresentados são razoáveis ou não.

Sendo assim, ferramentas que possibilitem encontrar o valor razoável de forma organizada e simples estão se tornando cadas vez mais populares. Um bom exemplo de uma ferramenta de apoio à decisão de compras online (N. Marcelo, 2003), na área de passagens aéreas, é o Decolar[1] que sumariza a busca de vários sites de companhias aéreas em um único local, oferecendo ao usuário uma melhor experiência na hora da compra. Sites com ainda mais detalhes possibilitam uma busca mais refinada, como por exemplo o Google Flight[2] e o Kayak[3], que fornecem mapas e gráficos que mostram a variação dos preços das passagens em diferentes datas.

O objetivo do trabalho é criar uma ferramenta web para dar suporte a usuários interessados em comprar passagens aéreas, sendo que, através de uma interação com um mapa, o usuário obtém informações sobre o preço promocional de uma passagem aérea. Os dados utilizados na ferramenta são coletados do site do Melhores Destinos[4] que oferece promoções de passagens aéreas na web. Depois os dados são tradados utilizando técnicas de limpeza de dados para o caso de valores estranhos. Os dados tratados são disponibilizados a partir de uma API que será consumida pela aplicação web. Nessa aplicação, haverá a possibilidade de escolher a origem da viagem em um mapa e todos os possíveis destinos. As informações do preço médio das promoções das passagens aéreas ficam disponibilizadas no mapa. Caso seja escolhido um destino, informações específicas do preço da passagem serão disponibilizadas, como por exemplo, o melhor dia da semana para a compra e o melhor mês para viajar.

O segundo objetivo deste trabalho é desenvolver um modelo de predição, que com as informações de origem e destino da viagem, juntamente com a data de ida e da volta, informe um possível valor promocional de uma passagem aérea. Será necessário utilizar técnicas de aprendizado de máquina nos dados.

A estrutura deste trabalho está organizada em 6 Capítulos. O primeiro Capítulo apresentado

[1] www.decolar.com
[2] www.google.com/flights
[3] www.kayak.com.br
[4] www.melhoresdestinos.com.br

é introdução, seguido pelo Capítulo 2 onde são apresentadas referências bibliográficas e trabalhos acadêmicos e ferramentas existentes no mercados relacionadas ao projeto. No Capítulo 3, são mostrados as técnicas utilizadas para desenvolver todas as seções do trabalho. No Capítulo 4, são apresentados os resultados obtidos com os dados e modelos de predição. No Capítulo 5, é feito um estudo do caso de uso que descreve o funcionamento da ferramenta de visualização de dados. Por fim, é apresentada a conclusão do projeto.

CAPÍTULO 2

Fundamentação Teórica

Neste Capítulo, serão apresentados conhecimentos que foram utilizados no projeto. Inicialmente, são apresentados conceitos utilizados na extração de informações na web, como as estruturas em que os dados podem ser encontrados e os formatos com que eles são extraídos. Em seguida, são explicadas técnicas de limpeza de dados para a retirada de *outliers*. Também é apresentado o conceito de aprendizado supervisionado e alguns dos seus diferentes métodos. Ao final, foram discutidos trabalhos acadêmicos relacionados ao tema do projeto, assim como ferramentas já existentes no mercado.

2.1 Extração de Informações na Web

Obter informações na web pode ser algo difícil, já que a disposição da informação dentro dos sites dificulta a leitura das máquinas e até das pessoas, prejudicando possíveis estudos como base nessas informações. Muitas vezes, as informações se encontram de forma dispersa dentro dos sites, como por exemplo, na forma de textos corridos, sendo chamados de dados não estruturados. Para utilizar esses dados é preciso extraí-los da página web de origem.

Dessa forma, são utilizadas técnicas de extração de informação, as quais podem variar a depender da organização dos dados dentro dos sites e formato de extração desses dados. Quanto à organização dos dados, esses podem se encontrar em formatos organizacionais diferentes (Adaniya & Proença, 2009), a saber:

- Sem estrutura: textos livres, sem nenhuma estrutura, onde estão contidas as informações de interesse. Na maioria das vezes, essas informações são escritas em linguagem natural;
- Semi-estruturado: mesmo apresentando uma estrutura clara, esses dados não podem ser manipulados, como por exemplo, uma tabela de informações de um produto na internet;
- Estruturado: documentos com estrutura clara e de fácil manipulação. Esse tipo de dados é facilmente encontrado em bancos de dados.

Já para o formato de extração dos dados, observa-se uma variação na aplicação de regras para recuperar as informações presentes nos locais de interesse. Esses formatos podem ser divididos em:

- Manual: extração onde, para cada página, os dados serão extraídos manualmente, se tornando pouco eficientes e pouco escaláveis;

- Supervisionada: com base em um grupo de sites previamente conhecidos, o programa de extração irá aprender a extrair as informações de interesse;
- Semi-supervisionada: para cada grupo de sites similares são aplicadas regras que identificam locais com os dados de interesse;
- Não supervisionada: tendo-se páginas de sites com o mesmo *template*, usa-se um algoritmo que analisa as páginas fornecidas e aprende o padrão para extração dos dados. Esse procedimento pode ser muito custoso computacionalmente.

A partir do conhecimento do formato da estrutura dos dados nas páginas dos sites de interesse e escolhendo-se um formato de extração dos dados, as informações relevantes são recuperadas e os dados encontrados são estruturados para futuras análises.

2.2 Limpeza de Dados

Para uma boa análise de dados, é necessário que os mesmos estejam com uma boa qualidade. Pontos que não representam a distribuição real dos dados podem prejudicar resultados das análises. Aplicar técnicas de limpeza de dados pode ajudar em futuras análises e possibilitar o uso de métodos e algoritmos de previsão (de Jonge & van der Loo, 2013). Porém, deve-se observar qual técnica de limpeza deverá ser aplicada para cada conjunto de dados, pois a má utilização pode acarretar dados errados. Dessa forma, algumas técnicas de limpeza de dados serão discutidas nesta Seção.

2.2.1 *Outliers*

Outliers são valores observados com baixa frequência dentro de um conjunto de dados, que se diferem muito dos demais . Esses valores podem atrapalhar futuras análises, distorcendo os resultados. Para evitar prejuízos nas análises, aconselha-se retirar todos os *outliers* do conjunto de dados. Como existe uma variação muito grande de definições para identificar se um valor é um *outlier* ou não, métodos estatísticos são utilizados para o reconhecimentos desses valores anormais (de Jonge & van der Loo, 2013). Esses métodos são divididos em dois grupos, os métodos univariados, os quais analisam cada variável separadamente, e os métodos bivariados, que avaliam a relação entre duas variáveis. Neste projeto, foram usados apenas métodos univariados e são explicados dois desses métodos nas próximas seções.

2.2.1.1 MAD_e *method*

O MAD_e *method* que utiliza a mediana e a MAD (*Median Absolute Deviation*) (2.1) para identificar um intervalo de confiança, onde qualquer valor fora desse intervalo pode ser considerado um *outlier* (Seo, 2002). Por utilizar a mediana e a MAD, esse método sofre menos influência dos valores extremos de um grupo de dados, dando a ele uma maior robustez. O MAD_e *method* é definido na Equação 2.2:

$$MAD = mediana(x_i - mediana(X)) \tag{2.1}$$

onde x_i são os valores do conjunto X

$$MAD_e = mediana(X) \pm 1,483 \times MAD(X) \tag{2.2}$$

onde X é um conjunto de valores.

2.2.1.2 *Tukey's method*

O *Tukey's method* (*boxplot*), onde são mostradas em um gráfico as informações sobre uma variável de um conjunto de dados, sendo essas informações o valor mais baixo, o valor mais alto, o valor da mediana, o valor do primeiro quartil (Q1) e o valor do terceiro quartil (Q3). A distância entre os quartis é chamada de *Inter Quartile Range* (IQR). Com o valor do IQR são criadas as cercas internas e externas de um *boxplot*. A cerca interna fica a uma distância de 1,5IQR para baixo de Q1 e para cima de Q3. A cerca externa fica a uma distância 3IQR para baixo de Q1 e para cima de Q3. Dessa forma, todos os valores que se encontram entre a cerca interna e a cerca externa são possíveis *outliers*, enquanto os que se encontram além da cerca externa são considerados pelo método *outliers*. Por utilizar os quartis e a mediana, o *tukey's method* é menos sensível a valores extremos. Um exemplo da representação do *boxplot* se encontra na Figura 2.1.

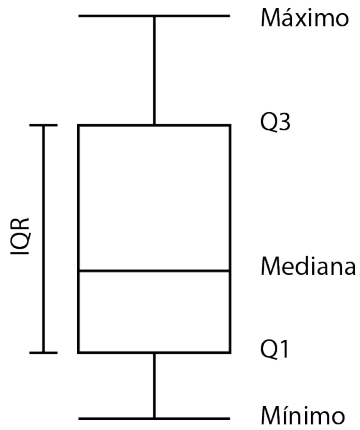


Figura 2.1: Exemplo da representação do *boxplot* onde estão destacados os valores máximo, primeiro quartil (Q1), mediana, terceiro quartil (Q3) e mínimo. Também está destacado o *Inter Quartile Range* (IQR). Fonte: adaptação de Marques (2015)

2.3 Aprendizado de Supervisionado

O aprendizado de máquina tem o objetivo de permitir que os computadores utilizem técnicas e algoritmos para que possam aprender sem que sejam explicitamente programados (Marques, 2015). Esse aprendizado pode ser usado para classificar grupos ou predizer valores dentro de um conjunto de dados. Há várias subcategorias para o aprendizado de máquina e uma delas é o aprendizado supervisionado.

Configura-se um aprendizado supervisionado, quando é fornecido um conjunto de dados contendo os valores de entrada e saída conhecidos que auxiliam na construção de um modelo usado para classificar grupos ou predizer valores de saída desconhecidos (Marques, 2015). Para modelos que classificam grupos, busca-se, a partir da entrada dos dados, separá-los em categorias distintas. Já para os modelos que buscam predizer valores, busca-se encontrar uma função contínua que, através da entrada de dados, seja possível entregar como saída um valor contínuo (Camilo & da Silva, 2009). Neste projeto, foi utilizado o aprendizado supervisionado para predição de preços de promoções de passagem aérea. A seguir, são apresentados alguns métodos de regressão utilizados neste trabalho

2.3.1 *Random Forest*

Random Forest é um classificador que, a partir de um grupo de árvores de decisão, entrega a média das predições dessas árvores como resposta. As árvores de decisão, utilizadas nesse classificador, são classificadoras de instâncias desconhecidas. A sua construção é feita a partir do nó raiz onde todo o conjunto de treinamento é considerado. Os dados em cada nó são particionados escolhendo-se o melhor atributo tendo como base um critério em que os dados fiquem melhor organizados nos novos nós criados. O critério de parada (criação de uma folha) ocorre quando a partição gerada não tem nenhum dado ou quando há um pequena quantidade de dados nessa partição. Essa folha é considerada um classe da instância, sendo necessário percorrer toda a árvore até uma folha para obter um classificação de uma instância desconhecida. Com pequenas mudanças no conjunto de treinamento, novas árvores são geradas fornecendo um classificador completamente diferente e são essas árvores diferentes que o *Random Forest* usa para gerar seu classificador (Marques, 2015).

2.3.2 *Support Vector Regression*

Support Vector Regression (SVR) é uma técnica de aprendizagem supervisionada, que, a partir de uma entrada de dados de treinamento reconhece padrões, permitindo assim a criação de modelo regressão. Esse modelo pode ser criado separando classes distintas através de um hiperplano (H1), sendo esses dados separados linearmente ou não. O hiperplano se utiliza de dois outros hiperplanos paralelos (H11 e H12). Para maximizar a generalização do classificador, busca-se encontra um hiperplano que maximize as áreas entre os hiperplanos paralelos, sendo essas áreas chamadas de margens (Marques, 2015). Um exemplo da construção do modelo SVR encontra-se na Figura 2.2.

2.3.3 Regressão Linear

A Regressão Linear busca a relação entre uma variável dependente com uma ou mais variáveis independentes através de uma equação que tenta explicar as variações entre as variáveis. Essa equação pode ser obtida através de um diagrama de dispersão, um gráfico que mostra o comportamento da variável dependente em função da variável independente, podendo esse comportamento aparecer de diversas formas: linear, quadrática, cúbica etc... . Os pontos no gráfico não se ajustaram com perfeição à curva do modelo proposto, havendo uma certa distân-

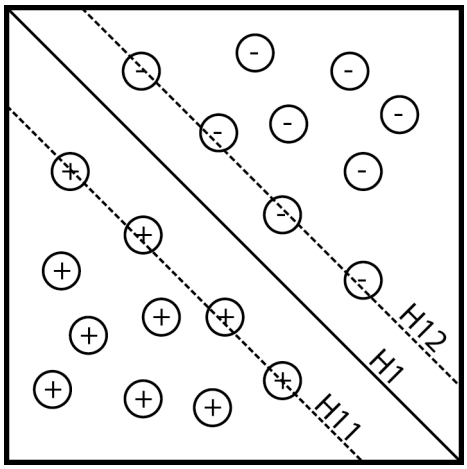


Figura 2.2: Exemplo de representação do modelo de SVR onde estão destacados o principal hiperplano (H1) e os hiperplanos de referência (H2 e H3). Fonte: adaptação de Marques (2015)

cia entre eles, já que os fenômenos usados nesse modelo não são matemáticos e sim fenômenos que acontecem ao acaso. Dessa forma, busca-se encontrar a curva que melhor se ajuste aos pontos do diagrama de dispersão. Para que o modelo escolhido tenha o menor erro possível, os pontos do diagramas têm que apresentar a menor distância possível em relação aos pontos da curva. Para isso, utiliza-se os Método dos Mínimos Quadrados (MMQ), o qual busca minimizar a soma de quadrados da distância dos pontos do giagramas e os pontos da curva estimada (Peternelli, 2004).

2.4 Trabalhos Relacionados

Nesta Seção, serão revisados trabalhos acadêmicos e ferramentas existentes no mercado, que auxiliam o consumidor na compra de passagens aéreas na internet.

2.4.1 Trabalhos Acadêmicos

No artigo de Todesco et al. (2008), foi feito um estudo na variação dos preços das passagens aéreas após grandes mudanças no setor de aviação brasileiro. O artigo teve como objetivo central saber em qual tipo de viagem as companhias aéreas disponibilizam os melhores descontos. Tendo em vista que vários fatores afetam os preços das passagens aéreas, os pesquisadores escolheram os que tinham maior relevância e os que tinham informações disponíveis na internet. Alguns desses fatores são a companhia aérea escolhida para voar, o horário escolhido para o embarque, os aeroportos de origem e destino, dentre outros. Após aplicarem um modelo de regressão múltipla nos dados coletados nos sites das próprias companhias, os pesquisadores observaram que os melhores preços das passagens aéreas são aplicados para voos feitos durante a madrugada e que os preços mais elevados ocorrem para voos com horários no início da noite e que possuem conexões.

Outro artigo encontrado, que fala sobre o melhor momento para comprar uma passagem,

foi o de Etzioni, Tuchinda, Knoblock, and Yates (2003). Nele, os autores desenvolveram um estudo inicial na área da mineração de dados sobre passagens aéreas com intenção de mostrar a grande variação dos preços das passagens e que é possível economizar dinheiro na hora da compra. Para esse estudo, foram coletados dados de duas rotas de viagens nos Estados Unidos a partir de um extrator de dados que aplicava técnicas de aprendizado de máquina para criar regras de extração. Com as informações das viagens armazenadas em bancos de dados, os pesquisadores aplicaram métodos de classificação, como *Ripper* (*Rule Learning*), *Q-learning* e *Time Series*, para gerar um algoritmo que fornecesse como resposta se o momento da compra da passagem era ideal ou não. Ao fim dos experimentos, os pesquisadores concluíram que apesar da eficiência dos algoritmos não ser muito alta, devido à falta de variáveis disponibilizadas pelas companhias aéreas, a economia financeira proporcionada era muito interessante, tornando-se assim viável o estudo na área de predição de preços de passagens aéreas.

2.4.2 Ferramentas Existentes no Mercado

O mercado de passagens aéreas evoluiu muito desde a consolidação da internet (Todesco et al., 2008). Muitas companhias aéreas viram os benefícios na comercialização de passagens online e abriram mais um canal de vendas para os clientes. As agências de viagens, que antes necessitavam de um espaço físicos para funcionar, também viram uma grande oportunidade no comércio digital e desenvolveram sistemas onde o consumidor pode explorar, de forma simples, diversas opções de viagens comparando preços entre diferentes companhias e diferentes horários.

Sites como o Decolar, Submarino Viagens[1], Melhores Destinos[2], Mundi[3], Expedia[4], dentre outros, têm sistemas parecidos para buscar passagens aéreas. A partir da entrada das informações da cidade de origem do usuário, do destino desejado, da data de início e fim da viagem (Figura 2.3), esses sites oferecem uma vasta opções de passagens aéreas das companhias aéreas que fazem o percurso escolhido, com variação da hora da saída e chegada do voo e uma grande variação de preços (Figura 2.4).

O usuário pode aplicar filtros para encontrar a passagem que atenda melhor às suas expectativas. Filtros como variação do preço da passagem, número de conexões ou escalas de voo, companhia aérea favorita e duração da viagem, são algumas das opções que o usuário encontra na maioria dos sites (Figura 2.5).

Mas ficar preso em uma busca com data específica, onde para comparar dias diferentes da viagem é necessário começar uma nova busca, se tornou um novo problema. Novamente, empresas viram a oportunidade de oferecer ferramentas mais atrativas, sendo agora possível comprar o preço das passagens aéreas em diferentes datas. Contudo, as melhorias nas ferramentas de buscas de passagens aéreas não acabam com essa nova funcionalidade. Melhorar a experiência que o consumidor pode ter na hora de comprar uma passagem possibilitou o desenvolvimento de sites onde o usuário pode interagir com um mapa, algo mais lúdico para uma viagem, para ver os possíveis destinos a partir da sua origem e ter uma noção do preço da

[1] www.submarinoviagens.com.br
[2] www.melhoresdestinos.com.br
[3] www.mundi.com.br
[4] www.expedia.com.br

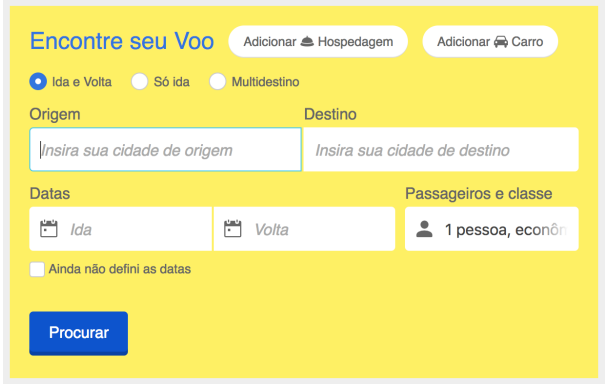


Figura 2.3: Área de busca de um do site da Decolar

passagem para cada um desses destinos.

A experiência de explorar o mapa para achar a viagem de interesse é encontrada em sites como o Google Flights, Kayak e Skyscanner[5]. Neles, o mapa mostra em destaque os destinos mais procurados a partir da localização atual do usuário e os melhores preços de passagens (Figura 2.6).

Outras ferramentas também estão disponíveis nesses sites. Gráficos barras que mostram a variação dos preços em datas futuras (Figura 2.7), ajudando a encontrar os dias com preços mais baixos. Calendários com os melhores preços para cada dia do mês (Figura 2.8), tornam mais fácil a visualização dos preços em relação às datas da viagem e aos dias das semana. O Google Flights oferece uma tabela onde é possível ver os preços das passagens aéreas para cada par de datas do início e fim da viagem (Figura 2.9).

Como podemos observar, a forma de visualizar as informações dos preços das passagens aéreas pode variar muito. Há alternativas de buscas mais diretas ou mais exploratórias. Cada vez mais, essas ferramentas auxiliam a tomada de decisão na hora da compra. Porém, a maioria dos sites para compra de passagens aéreas ainda restringem suas buscas a uma data específica ou a um único destino para a viagem, dificultando a comparação entre elas. Nos sites que permitem uma exploração livre dos dados, é possível observar uma baixa performance por conta do grande volume de dados, como observado no Google Fligths. Apesar de todas essas alternativas, ainda há muito a se evoluir no comércio digital de passagens aéreas para que o usuário possa tomar a melhor decisão na sua compra.

[5] www.skyscanner.com.br

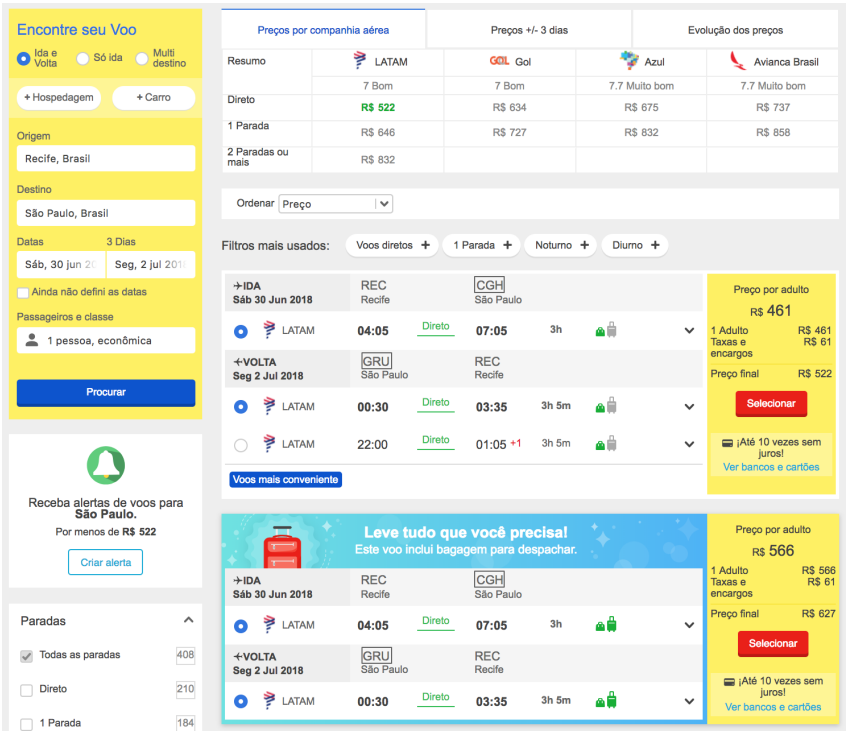


Figura 2.4: Resultado da busca de uma viagem entre Recife e São Paulo no site da Decolar

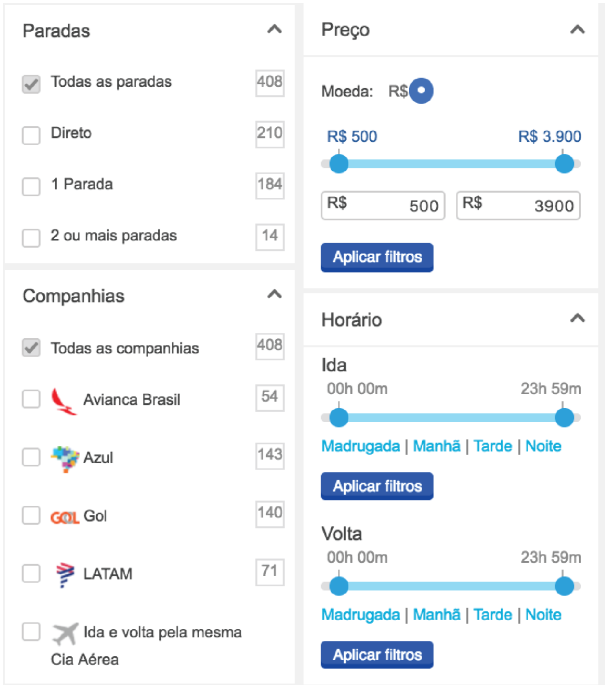


Figura 2.5: Filtros de quantidade de paradas, companhias aéreas, variação de preço e variação de horário encontrados no site da Decolar

Figura 2.6: Mapa fornecido pelo site do Kayak para a exploração do preço das passagens aéreas

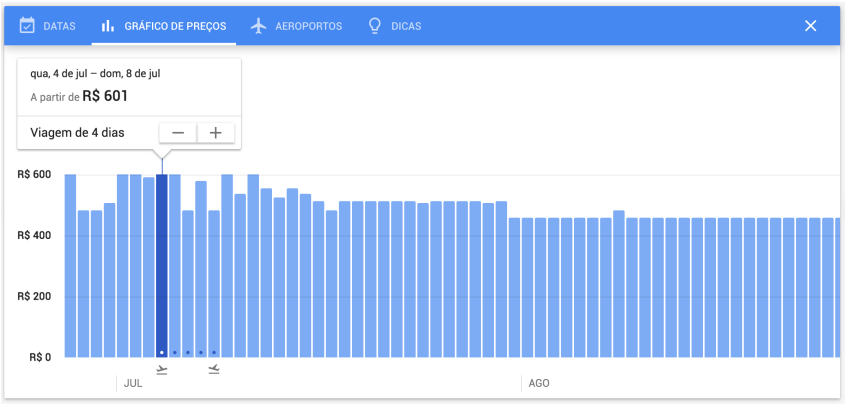


Figura 2.7: Gráficos barras que mostram a variação dos preços nos meses de julho e agosto de 2018 fornecidos pelo site do Google Flights para uma viagem entre Recife e São Paulo

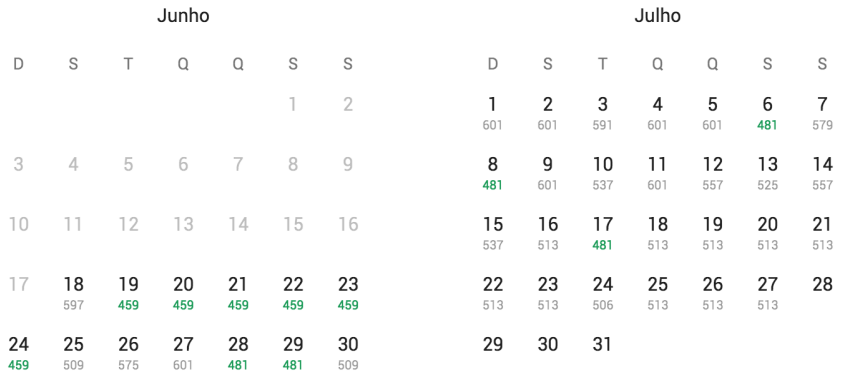
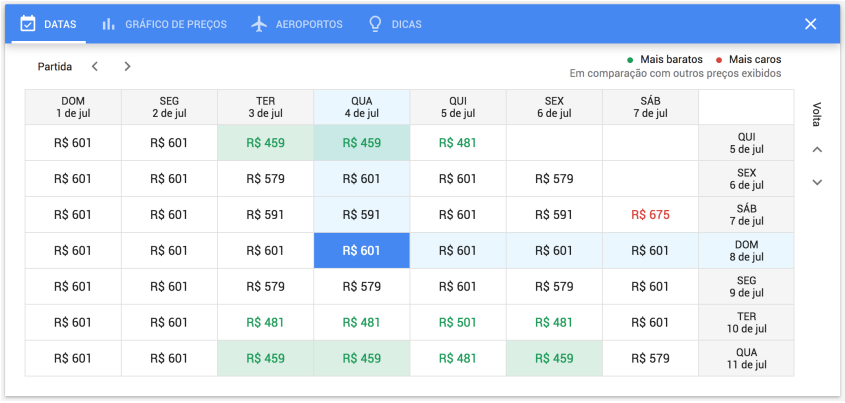


Figura 2.8: Calendários com os melhores preços para os meses de junho e julho de 2018 fornecidos pelo site do Google Flights para uma viagem entre Recife e São Paulo



DOM 1 de jul	SEG 2 de jul	TER 3 de jul	QUA 4 de jul	QUI 5 de jul	SEX 6 de jul	SÁB 7 de jul	
R$ 601	R$ 601	R$ 459	R$ 459	R$ 481			QUI 5 de jul
R$ 601	R$ 601	R$ 579	R$ 601	R$ 601	R$ 579		SEX 6 de jul
R$ 601	R$ 601	R$ 591	R$ 591	R$ 601	R$ 591	R$ 675	SÁB 7 de jul
R$ 601	R$ 601	R$ 601	R$ 601	R$ 601	R$ 601	R$ 601	DOM 8 de jul
R$ 601	R$ 601	R$ 579	R$ 579	R$ 601	R$ 579	R$ 601	SEG 9 de jul
R$ 601	R$ 601	R$ 481	R$ 481	R$ 501	R$ 481	R$ 601	TER 10 de jul
R$ 601	R$ 601	R$ 459	R$ 459	R$ 481	R$ 459	R$ 579	QUA 11 de jul

Figura 2.9: Preços das passagens aéreas para cada par de datas de início e fim da viagem fornecidos pelo site do Google Flights para uma viagem entre Recife e São Paulo

CAPÍTULO 3

Solução

Para a criação de uma ferramenta web de apoio à decisão na compra de passagens aéreas e para a predição do preço dessas passagens, dividiu-se o projeto em quatro partes. Inicialmente, os dados são coletados do site Melhores Destinos através de um extrator que coleta os dados referentes às promoções de passagens aéreas disponíveis no site, armazenado esses dados em arquivos CSV (*Comma-Separated Values*). Com os dados salvos, eles são analisados e técnicas de limpeza de dados são aplicadas para a remoção de *outliers*, pois esses valores podem distorcer futuros resultados. Os dados tratados são usados para a predição de preços de promoções de passagens aéreas ou para a exploração em uma interface web. Todas as partes do projeto são explicadas nas seções deste capitulo e podem ser visualizadas na Figura 3.1

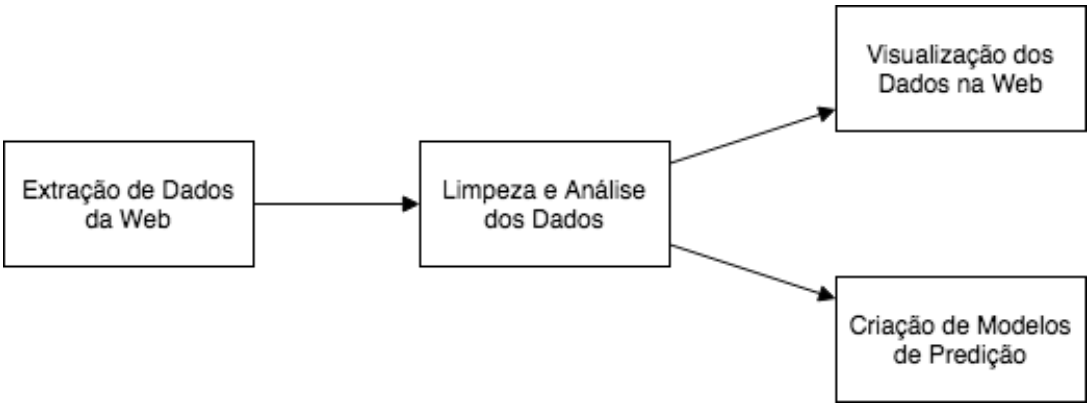


Figura 3.1: Partes do projeto e na ordem de sua execução

3.1 Coleta dos Dados

Para poder criar um sistema de apoio à decisão da compra de passagens aéreas promocionais, é necessário obter dados relacionados às viagens. Para isso buscou-se encontrar um local que fornecesse informação de passagens aéreas promocionais como a origem e o destino da viagem, a data que essa viagem ia ocorrer e o valor da mesma. Após buscas por sites que pudessem fornecer essas informações, observou-se que o site Melhores Destinos tem informações necessárias, facilitando a extração dos dados. Para ilustrar a organização dos dados no site, aplicou-se o filtro para uma viagem entre São Paulo e Recife e obteve-se as datas de ida e volta da viagem, a companhia aérea que fará o translado e o valor da passagem aérea como mostrado na Figura 3.2.

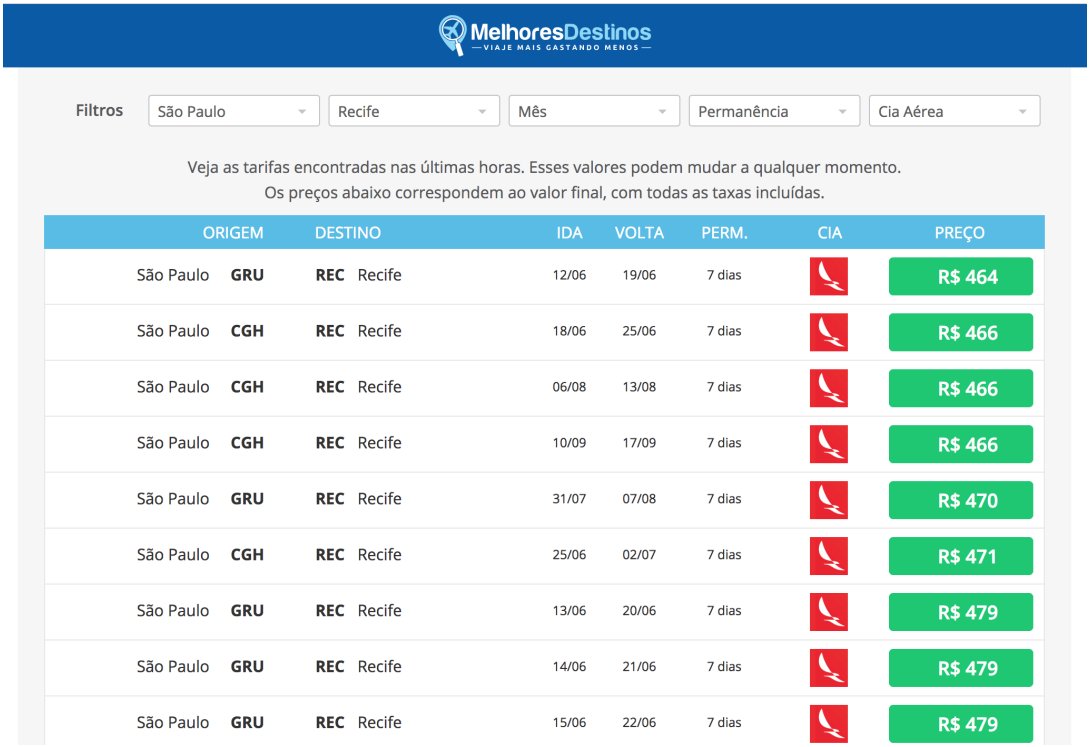


Figura 3.2: Informações das viagem quando se escolhe a cidade de São Paulo como origem da viagem e Recife como destino no site Melhores Destinos

Já que todos os dados foram retirados do mesmo site pode-se aplicar um extrator de dados (*wrapper*) mais focado e com maior eficiência. Detectando-se a região com os dados desejados, que visualmente no site encontram-se em uma tabela, criou-se um parse sobre o DOM *tree*. O DOM (Document Object Model) é um modelo HTML (HyperText Markup Language) constituído por uma árvore (*tree*) com objetos HTML. Já o *parse*, um algoritmo que recebe como entrada um objeto e, aplicando-se regras, entrega um novo objeto criado com as informações da entrada. No caso deste projeto, utilizou-se um parse na DOM *tree* para extrair as informações das passagens e criar novos objetos com essas informações.

Após a extração das informações das passagens aéreas, esse dados foram armazenados em arquivos logs com formato CSV para serem processados posteriormente. Essa rotina de extração aconteceu todos os dias, em um horário fixo no período de 01 de agosto de 2016 a 27 de janeiro de 2018. Todo o log da extração fica armazenado em uma máquina virtual.

3.2 Análise e Limpeza dos Dados

Com os dados extraídos do site Melhores Destinos, foram analisadas as informações em cada uma das variáveis inicialmente escolhida, sendo elas dia da postagem da promoção cidade origem, cidade destino, preço da passagem, dia do início da viagem, dia da volta da viagem e a companhia aérea.

Para a primeira verificação feita nos dados, buscou-se saber se havia alguma informação faltando em alguma das variáveis. A ausência de informações deve ser resolvida ou retirando os dados da viagem onde há uma variável faltando ou atribuindo um valor para essa variável. A técnica a ser utilizada depende da quantidade de dados disponível ou da consequência de inserir dados não reais no conjunto original. Caso haja poucos dados disponíveis para análise, recomenda-se utilizar técnicas de imputação de dados onde é feita uma previsão dos valores faltantes. No caso do conjunto de dados de promoções de passagens aéreas utilizado neste projeto, nenhum dado estava faltando, dessa forma não foi necessário aplicar nenhuma técnica para dados ausentes.

Na segunda verificação que foi feita, buscou-se por *outliers* (já explicado na Subseção 2.2.1) no conjunto de dados. A única variável onde inicialmente foi possível buscar por valores estranhos foi o campo de preço da passagem aérea, já que esse é o único valor numérico presente no início da análise. Dessa forma, para verificar se havia valores estranhos, analisou-se um resumo dos preços a partir dos valores máximo, mínimo, primeiro quartil, terceiro quartil, média e a mediana do preço das promoções. Os valores obtidos podem ser encontrados na Tabela 3.1. Percebe-se que o valor máximo (R$ 3.927.539,00) mostra-se muito estranho para o preço de uma passagem aérea. Dessa forma, é interessante aplicar técnicas de remoção de *outliers*. Porém, sabe-se que a maior parte das viagens tem origem em cidades brasileiras (99,98% das viagens), tendo como destinos tanto cidades brasileiras (36,84% das viagens) quanto cidades fora do Brasil (63,16% das viagens), e que o valor das promoções de passagens aéreas para viagens nacionais e internacionais tem um intervalo de preço diferente. Espera-se que a variação dos preços das promoções de viagens feitas dentro do Brasil seja menor do que a variação dos preços de viagens feitas para cidades de outros países. Para separarmos os tipos das viagens, criou-se um novo campo no conjunto de dados que informaria se a viagem é nacional ou internacional. Para isso, foi preciso buscar uma lista com o nome de todas as cidades brasileiras e verificar se o destino da viagem está dentro dessa lista. Caso o destino se encontre dentro da lista, assume-se que a viagem é nacional; caso contrário, a viagem é tida com internacional. Com essa informação, o conjunto de dados foi dividido em dois (conjunto de viagens nacionais e conjunto de viagens internacionais) e aplicou-se o método MADe (mais explicações na Subsubseção 2.2.1.1) para encontrar o intervalo de confiança desses dois conjuntos. Os valores máximo e mínimo do intervalo de confiança de cada um dos conjuntos são de R$ 5.012,00 e R$ 1.064,00 para viagens internacionais e R$ 1.557,00 e R$ 240,00 para viagens nacionais. Observa-se que os valores dos intervalos de confiança responderam às expectativas, já que os valores máximo e mínimo do intervalo das viagens internacionais são maiores do que os das viagens nacionais. Os valores que se encontram fora desses intervalos, em cada um dos conjuntos, foram retirados do conjunto de dados final do projeto. Uma outra alternativa que poderia ter sido utilizada para encontrar *outliers* seria observar os preços referentes às viagens com origem e destino distintos.

Outra informação interessante que foi obtida a partir das variáveis existentes no conjunto de dados foi a duração da viagem onde subtraiu-se a data do início da viagem do término da viagem, obtendo a quantidade de dias entre essas duas datas. Novamente, analisou-se o resumo das informações dessa nova variável, para verificar se havia valores estranhos. Esse resumo se encontra na Tabela 3.1. Valores negativos para a duração da viagem mostram-se

muito equivocados. Dessa forma, foram retirados todos os valores menores do que um dia de duração.

Ainda observando as datas disponíveis no conjunto de dados, pôde-se criar uma nova variável. Foi possível descobrir quanto tempo há entre o dia da postagem da promoção e o dia do início da viagem, subtraindo-se essas duas datas. Esse campo mostra-se interessante para ser usado em futuras análises, pois, juntamente com o preço das promoções, pode-se verificar a variação do preço entre viagens onde a promoção foi postada muito antes do início da viagem ou mais próxima da data de embarque. Assim como foi feito com todas as variáveis numéricas, analisou-se o seu resumo dos valores (Tabela 3.1). Novamente, foram observados valores negativos ou zero para a quantidade de dias entre a postagem da promoção e o início da viagem. Como esses valores não representam a realidade, eles foram retirados do conjunto final dos dados.

Novos campos também foram adicionados para auxiliar em futuras análises ou para serem apresentados na visualização dos dados. As informações do dia da semana e mês da data da postagem da promoção foram extraídas para duas novas variáveis. Da mesma forma, foram extraídas para duas variáveis os valores do dia da semana e mês da data de início da viagem. Essas variáveis foram adicionadas ao conjunto de dados finais deste projeto.

A última variável adicionada ao conjunto de dados final, foi a distância entre a origem e destino da viagem. Para isso, foi necessário obter as coordenadas geográficas das cidades de origem e destino. Essas coordenadas foram obtidas utilizando a API do *Google Maps Geocoding*, que, ao se fornecer o nome de uma cidade, retorna a sua latitude e longitude (Maps, 2018), dentre outras informações que não foram usadas neste projeto. Com os valores das coordenadas, utilizou-se a Fórmula de *Haversine*, que, a partir da latitude e longitude de duas coordenadas em uma esfera, fornece a distância entre os dois pontos (Hijmans, Williams, & Vennes, 2017). Essa fórmula foi utilizada a partir da biblioteca *geosphere* da linguagem R. Para todo o par de origem e destino, foi encontrada uma distância e armazenada em uma nova variável.

Ao final de todos os processos de limpeza, foram adicionadas novas variáveis ao conjunto de dados ficando disponíveis a origem, destino, preço, companhia aérea, data da postagem da promoção, data do início da viagem, data do fim da viagem, duração da viagem, dia da semana da viagem, mês da viagem, dia da semana da postagem, mês da postagem e quantidades de dias antes da viagem.

3.3 Predição de Preços

Esta Seção tem como objetivo explicar a escolha das variáveis utilizadas nos métodos geradores dos modelos para a predição dos preços promocionais de passagens aéreas.

3.3.1 Escolha das Variáveis

Para minimizar os erros nos métodos utilizados para prever os valores promocionais das passagens aéreas, foram feitas análises nos dados para identificar as variáveis mais importantes dentro de todo o conjunto de dados.

Tabela 3.1: Resumo dos valores mínimo, primeiro quartil, mediana, média, terceiro quartil e máximo das variáveis preço da promoção, duração da viagem e dias antes da viagem sem terem recebido técnicas de limpeza

Resumo	Preço (R$)	Duração (dias)	Dias antes da viagem
Mínimo	59	-364	-30
Primeiro Quartil	787	7	52
Mediana	1551	9	91
Média	1960	10	110
Terceiro Quartil	2606	14	155
Máximo	3927539	350	358

O primeiro passo para encontrar as variáveis mais importantes foi agrupar os dados com base nas duplas de cidade origem e destino para descobrir as duplas que mais foram postadas durante o ano, já que, dessa forma, existiria uma melhor distribuição das informações no ano. Após a contagem, observou-se que a maioria das duplas de cidades representavam viagens internacionais. Para que fossem observadas também, viagens com destinos brasileiros, foram escolhidas as informações de viagens de cinco duplas de cidades com destino internacional e cinco duplas de cidades com destino nacional. Dessa forma, foram criados arquivos CSV para cada dupla de viagens, com as informações de mês e dia da semana da postagem da promoção, mês e dia da semana da viagem, preço da promoção, quantidade de dias antes da viagem cuja promoção foi postada e a duração da viagem. Essas informações foram tidas como as mais importantes para auxiliar na previsão de preços promocionais de passagens aéreas. Foram criados também dois outros conjuntos de dados, sendo um com todos dados de viagens das cinco duplas de cidades para viagens nacionais e o outro grupo com os dados das cinco duplas de cidades com destinos internacionais. Dessa forma, houve a comparação se a predição dos preços de promoções de passagens aéreas seria mais eficiente com os dados de uma única viagem ou se seria mais eficiente com um grupo de viagens. Vale ressaltar que, para esses dois conjuntos de dados, foram adicionadas duas novas variáveis, a cidade origem e a cidade destino da viagem.

Todos os conjuntos de dados utilizados nos métodos de predição tiveram que ser separados em dois outros conjuntos, sendo que um serviria para o treinamento dos modelo de predição e o outro conjunto serviria para o teste do modelo. Para isso, os conjuntos de duplas de viagens e os do tipo da viagem (nacional e internacional) foram separados em quantidade de 70% dos dados para treino e 30% dos dados para teste. Para fazer essa separação, foi encontrada a data de postagem da promoção que divide os conjuntos de dados na proporção desejada. As duplas de cidades escolhidas e as quantidades de dados nos conjuntos de treino e teste encontram-se na Tabela 3.2

3.3.2 Técnicas Utilizadas

Para se aplicar os métodos de predição previamente explicados na Seção 2.3, foi necessário utilizar a biblioteca scikit-learn para a linguagem de programação Python. Essa biblioteca de

Tabela 3.2: Tamanho dos conjuntos de treino e teste usados nos métodos de predição

Viagem (Origem - Destino)	Conjunto de Treino	Conjunto de Teste
Campo Grande - São Paulo	10.217	3.999
Cuiabá - São Paulo	11.689	3.828
Florianópolis - Rio de Janeiro	10.019	4.071
Goiânia - Rio de Janeiro	10.018	3.963
Salvador-São Paulo	11.250	4.345
Rio de Janeiro - Miami	14.994	6.354
Rio de Janeiro - Orlando	14.812	5.495
São Paulo - Madri	13.254	5.064
São Paulo - Miami	17.493	7.292
São Paulo - Nova York	13.946	5.751

código aberto oferece vários algoritmos de classificação, regressão e agrupamento da área de aprendizado de máquina (Fabian Pedregosa, 2011).

Como todos os dados estavam em arquivos CSV, foi necessário utilizar a biblioteca Pandas da linguagem Python para fazer a leitura dos arquivos. A biblioteca Pandas oferece estruturas de dados que permite uma manipulação rápida, flexível e expressiva de dados relacionais ou classificáveis de maneira fácil e intuitiva (Pandas, 2018).

Com os dados disponíveis para serem utilizados no código Python, foram aplicados os conjuntos de dados de treino nos métodos *Random Forest*, SVR e Regresão Linear para a criação dos modelos de predição. Todos esses métodos foram importados da biblioteca scikit-learn para serem usados no código em Python. Após a criação dos modelos, foram aplicados os dados de teste na função de predição, fornecida por cada um dos modelos, para receber como resultado os valores preditos dos preços de promoções de passagens aéreas de cada conjunto de dados. Esse valores preditos foram utilizados para descobrir o MRE (*Mean Relative Error*) e o MSE (*Mean Squared Error*), dados representados pelas Equações 3.1 e 3.2.

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y} \tag{3.1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.2}$$

onde Y representa os valores reais e $\hat{Y}$ representa os valores preditos.

3.4 Visualização dos Dados

Para auxiliar o usuário na tomada de decisão no momento da compra da passagens aéreas, criou-se uma ferramenta que auxilia a visualização e exploração dos dados referentes às promoções. A partir da interação com um mapa, o usuário pode obter todas as informações disponíveis pelo sistema, através de tabelas e gráficos.

Foi necessário desenvolver uma infraestrutura para disponibilizar os dados das passagens aéreas na internet. Todos os dados tratados foram armazenados usando a tecnologia do MongoDB, um banco de dados não relacional. Foi criada uma API para disponibilizar os dados salvos no banco, para isso, sendo utilizado o NodeJS para criar toda a estrutura necessária. Por fim, os dados das promoções de passagens aéreas foram consumidos por uma ferramenta web que disponibiliza os dados através de um mapa e gráficos que auxiliam na exploração das informações das viagens. Para isso, foi utilizada React, uma biblioteca JavaScript para criação de componentes. É possível observar todo o pipeline dessa parte do projeto na Figura 3.3. Dessa forma, esta Seção mostra toda a infraestrutura necessária para dar suporte à interface da visualização dos dados das promoções de passagens aéreas.

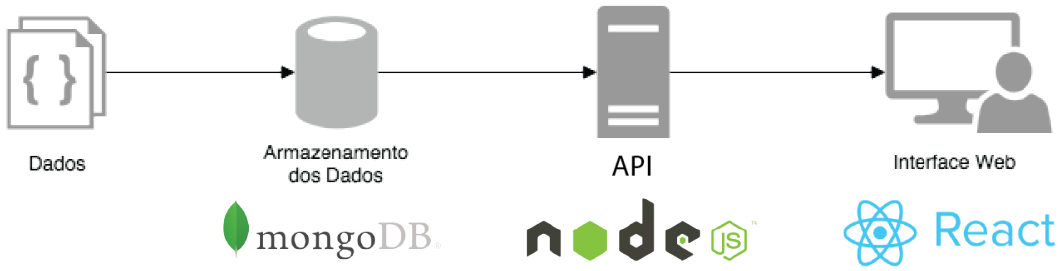


Figura 3.3: Pipeline com todo o processo necessário para a visualização dos dados de promoções de passagens aéreas através da interface web

3.4.1 Armazenamento dos Dados

Já tendo em mãos os dados pré-processados e com as informações que se deseja disponibilizar para o usuário, foi escolhido um banco de dados que agregasse na infraestrutura do sistema. O banco utilizado neste projeto deve suportar um grande volume de informações e permitir a criação de *querys* diversificadas. Dessa forma, foi escolhido o Mongo DB.

O Mongo DB é um banco não relacional que armazena documentos, com as informações dos dados, no formato BSON (*Binary JavaScript Object Notation*) (representação binária do JSON) e fornece uma grande velocidade em suas *querys* (Xplenty (2017)). A divisão dos dados é feita através de *collections* que armazenam documentos similares em um mesmo local. A título de comparação, as *collections* têm a mesma função que tabelas em bancos relacionais como o MySQL (Xplenty (2017)). O MongoDB tem uma fácil integração com projetos em NodeJS, principalmente ao utilizar-se a biblioteca Mongoose, que auxilia na criação de *schemas*, construção de *querys* e comunicação com o banco (Medlock (2017)).

Para este projeto foi necessária a criação de três *collections* para armazenar diferentes tipos de dados que serão consumidos pela API (*Application Programming Interface*). Uma das *collection* armazena documentos com o nome e as coordenadas geográficas de todas as cidades que servem como origem de uma viagem. Outra *collection* presente no banco armazena as informações básicas de uma viagem, tendo como principais campos as cidades origem e destino, o valor médio das promoções e a quantidade de suas ofertas. Além desses campos, também estão presentes as coordenadas geográficas das cidades da viagem. A última *collection* existente no banco armazena as informações mais detalhadas da viagem, que foram utilizadas para

construção das tabelas e gráficos na interface web (mais explicações na Subseção 3.4.3). Os documentos dessa *collection* têm em sua estrutura o nome e as coordenadas geográficas das cidades origem e destino, o valor médio das passagens promocionais, assim como a quantidade de passagens ofertadas. Também existem os campos de "mês da compra", "mês da viagem", "dias antes da viagem", "duração da viagem", "dia da semana de compra"e "dia da semana da viagem", sendo esses campos listas onde constam as informações de preço das passagens promocionais e de quantidade de promoções ofertadas.

3.4.2 API

A API tem como objetivo ser um servidor que recebe requisições HTTP em seus *endpoints* (rotas) e respondem com os dados em formato JSON. Essas requisições podem vir com parâmetros que serão utilizados em lógica de negócio dentro do servidor.

A montagem da API deste projeto tem como base NodeJS, uma plataforma para aplicações web que se utiliza do motor V8 da Google para interpretar JavaScript no lado do servidor (Santos (2016)). O NodeJS oferece nativamente bibliotecas que auxiliam no tratamento das requisições HTTP (*HyperText Transfer Protocol*), porém, a biblioteca Express facilita ainda mais esse tratamento. Através de algumas configurações simples, é possível criar *endpoints* para receber requisições HTTP de todos os tipos, como por exemplo requisições do tipo GET ou POST. Essa biblioteca precisa ser instalada através do NPM (*Node Package Manager*), ficando disponível no escopo do projeto.

No projeto, foram criados três *endpoints* que respondem por três solicitações distintas. Um dos *endpoints* tem a responsabilidade de fornecer todas as coordenadas das cidades que podem ser origem de uma viagem. Outro *endpoint* recebe como parâmetro o nome da cidade origem de uma viagem e envia como resposta todas as cidades que podem ser destinos daquela origem e o preço médio das promoções. O último *endpoint* recebe como parâmetro as cidades origem e destino de uma viagem e tem a responsabilidade de responder com as informações do percurso escolhido. Esse último *endpoint* ainda recebe um terceiro parâmetro que está relacionado com o tipo de informação a ser retornada, podendo esta ser completa ou resumida.

Para obter todas as informações necessárias às respostas dos *endpoints*, é necessário que haja uma comunicação com o banco. Para isso é preciso instalar, via NPM, a biblioteca do MongoDB. Dessa forma, uma conexão com o banco pode ser aberta e *querys* podem ser feitas para buscar os dados necessários que fazem parte das respostas dos *endpoints*. Para simplificar a conexão com o banco de dados MongoDB e a montagem das *querys*, foi instalada a biblioteca Mongoose através do NPM.

3.4.3 Interface Web

A interface de visualização foi desenvolvida para prover as seguintes interações com os dados:

- Interação 1 (I1): a partir da escolha de uma origem no mapa, mostrar as possíveis cidades destino e o valor médio das promoções para cada destino;
- Interação 2 (I2): ao escolher um destino, mostrar informações resumidas sobre os melhores momentos para comprar a passagem;

- Interação 3 (I3): explorar informações mais detalhadas sobre as mudanças nas ofertas e nos preços das promoções para a viagem escolhida.

As figuras 3.4 e 3.5 mostram as ferramentas usadas pelo usuário para realizar as interações.

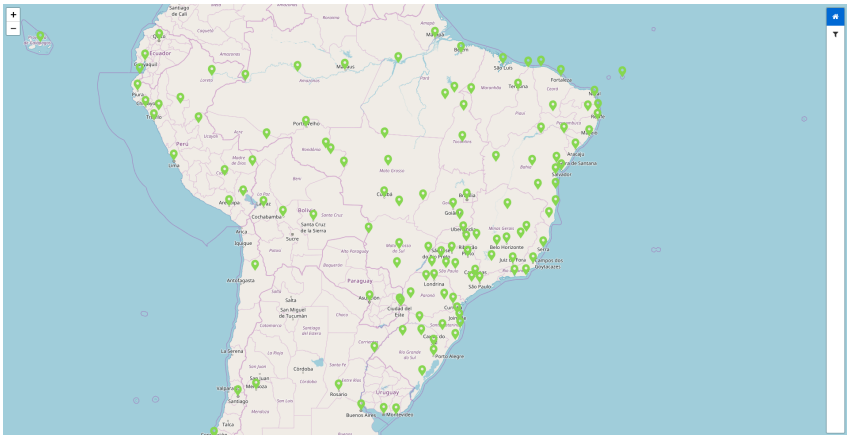


Figura 3.4: Mapa da tela inicial no sistema de visualização dos dados

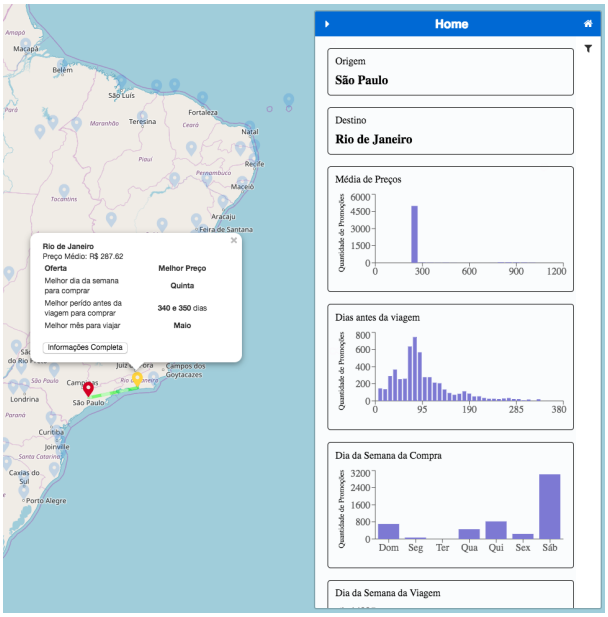


Figura 3.5: Tabela e gráficos disponíveis no sistema de visualização dos dados de promoções de passagens aéreas

Foi escolhido o ambiente web para disponibilizar a interface de visualização dos dados. Esse ambiente facilita a interação do usuário e retira a necessidade de instalação do sistema na máquina do usuário. Dessa forma, o sistema pode atingir uma maior quantidade de pessoas interessadas nas informações. Para disponibilizar a interface na web, utilizamos React , uma bibliotecas JavaScript que possibilita a criação dos componentes necessários na construção

do site. Para a interface possuir mapa e gráficos, foram escolhidas duas extensões do React: Leaflet-react e Recharts, respectivamente.

A interação com a interface é feita através do mapa, onde no primeiro momento estão disponíveis todas as cidades que podem servir como origem para uma viagem. A escolha de uma cidade no mapa traz mais informações sobre as promoções, como por exemplo o preço médio das promoções em cada destino. Com a escolha da origem e destino da viagem, o usuário pode solicitar informações mais completas sobre essa viagem, tendo acesso a gráficos que detalham a distribuição das promoções em intervalos de dia da semana, meses do ano, dias antes da viagem e duração da viagem.

3.4.3.1 Mapa

No primeiro momento, o mapa mostra todas as localizações das cidades que podem tornar-se a origem de uma viagem. Essas cidades estão destacadas com os marcadores em verde (Figura 3.6a) tornando assim disponível o nome da cidade naquela localização ao se interagir com os marcadores. A ação de clicar em um marcador verde resulta na escolha da cidade origem de uma viagem e o marcador muda para a cor vermelha (Figura 3.6b).

Com a escolha da origem da viagem, o mapa mostra marcadores azul (Figura 3.6c) representando as possíveis cidades destino. A interação com os marcadores azuis permite a visualização do nome da cidade e o preço médio das promoções, como previsto na interação (I1). Apesar de mostrar a palavra "preço médio"na visualização, para encontrar esse valor utilizamos a mediana dos preços de promoções entre a origem e o destino, já que o valor da mediana sofre menos influência dos valores extremos dos preços promocionais que estão em menor quantidade nos dados. Já para a função média, esses valores extremos teriam grande influência no resultado (Seo (2002)). Mesmo havendo essa diferença, optou-se por usar o termo "preço médio"por ser uma expressão mais acessível para as pessoas no geral e sabendo que geralmente a mediana e a média de um conjunto numérico têm valores muito próximos (Muniz (2018))

Com a visualização das possíveis cidades destino, o usuário pode clicar em qualquer marcador azul escolhendo assim o destino da viagem, alterando sua cor para amarela (Figura 3.6d) e adicionando no mapa um caminho verde que representa a orientação da viagem (Figura 3.7). Todos os outros possíveis destinos têm sua opacidade diminuída (Figura 3.6e), deixado em destaque as cidades escolhidas para a viagem e seu percurso. Ao passar o mouse por cima do marcador amarelo (cidade destino) uma nova caixa aparece com as informações resumidas dos melhores momentos para encontrar as promoções de passagem mais interessantes (Figura 3.8), realizando assim a interação (I2). Essa nova caixa, além de continuar mostrando o nome da cidade e o preço médio das promoções, também apresenta o melhor dia da semana para comprar a passagem para a viagem. Outra informação é o melhor período antes da viagem para comprar a passagem aérea. Esses períodos têm intervalos de 10 dias e possuem o valor máximo de até 350 dias (exemplo: 10 a 20 dias antes da viagem). A última informação a ser mostrada é o melhor mês para se comprar a passagem. Essas informações baseiam-se no menor valor da mediana do preço das promoções para cada uma das informações da viagem escolhida.

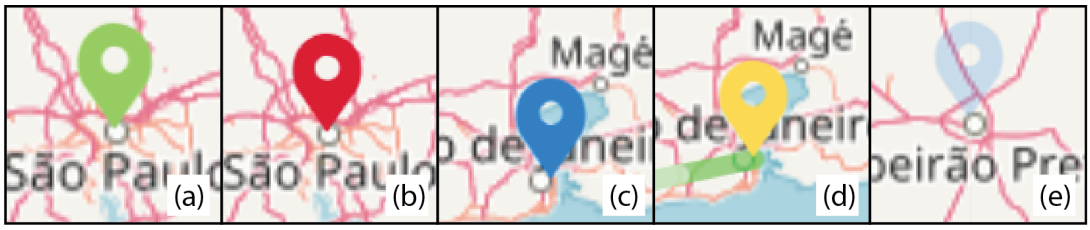


Figura 3.6: Marcadores que representam a localização das cidades. (a) Marcador verde representando possíveis cidades origem da viagem. (b) Marcador vermelho representando a origem da viagem. (c) Marcador azul representando possíveis cidades destino da viagem. (d) Marcador amarelo representando o destino da viagem. (e) Marcador azul com baixa opacidade representando possíveis cidades destino da viagem quando uma cidade destino já foi escolhida.

Figura 3.7: Caminho entre as cidades origem e o destino da viagem que auxilia na orientação do percurso

3.4.3.2 Gráficos

Na mesma caixa das informações resumidas, é possível obter informações mais detalhadas sobre a variação das quantidades e preços das promoções interagindo com o botão "Informações Completas". Uma nova área com gráficos de barra e linha surge no lado direito do mapa.

A utilização de gráficos de barras na vertical, tem como objetivo facilitar a interpretação dos dados, já que esse é um dos gráficos mais comuns para a maioria das pessoas e representa de forma coerente dados do tipo ordinal (Chartblocks, 2018a). Todos os gráficos em barra usados na visualização têm como o eixo y a quantidade de promoções para cada categoria presente no eixo x. No primeiro gráfico de barra, tem-se a informação do "Preço Médio"que mostra a relação da quantidade de promoções com faixas de preços das passagens promocionais (Figura 3.9a). No segundo gráfico de barra, "Dias antes da viagem", mostra-se a variação da quantidade de promoções para cada intervalo de 10 dias antes da data da viagem (Figura 3.9f). O terceiro e quarto gráficos de barra mostram a variação da oferta de promoções no decorrer dos dias da semana, tendo como diferença entre os gráficos o momento desses dias, sendo o primeiro em relação ao dia da semana da compra da passagem (Figura 3.9d) e o segundo com relação ao dia da semana da viagem (Figura 3.9b). Existe também uma relação entre os dois últimos gráficos de barra, mostrando a variação da quantidade de promoções nos meses do ano, sendo que o primeiro mostra a variação nos meses da compra (Figura 3.9e) e o segundo tem relação com os meses do início da viagem (Figura 3.9c). Ao interagir com as barras

Figura 3.8: Informações resumidas dos melhores momentos para encontrar promoções das passagens aéreas

dos gráficos, todos eles apresentam também a média de preços em cada um dos momentos; então além de saber a quantidade de promoções em um determinado dia da semana da viagem (Figura 3.9g), o usuário também pode saber o valor médio das promoções em cada um desses momentos.

Gráficos de linha se mostram interessantes para dados em série com valores contínuos, ajudando na visualização das variações em longos períodos (Chartblocks, 2018b). O último gráfico da caixa de informações completa da viagem escolhida, "Duração da Viagem", foi construído utilizando um gráfico de linha. Nele, temos a variação da quantidade de promoções em relação à duração da viagem (Figura 3.10).

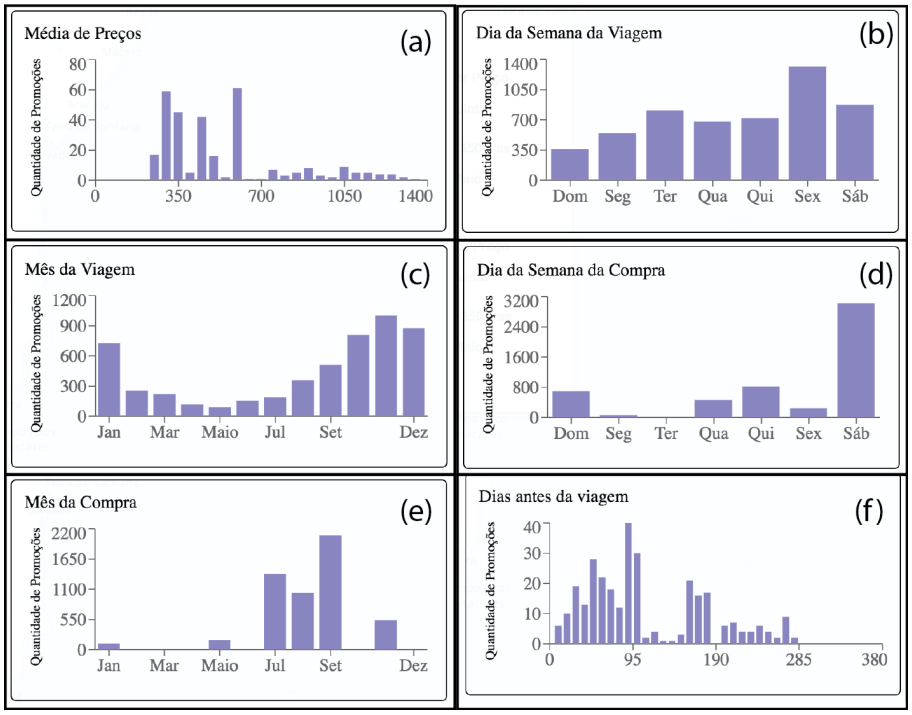


Figura 3.9: Gráficos de barra encontrados na ferramenta de visualização dos dados. a) Gráfico da quantidade de promoções em relação a intervalos de preço; b) Gráfico da quantidade de promoções em relação aos dias da semana do início da viagem; c) Gráfico da quantidade de promoções em relação a possíveis meses da viagem; d) Gráfico da quantidade de promoções em relação aos dias da semana da compra; e) Gráfico da quantidade de promoções em relação aos meses da compra; f) Gráfico da quantidade de promoções em relação a intervalos de duração da viagem;

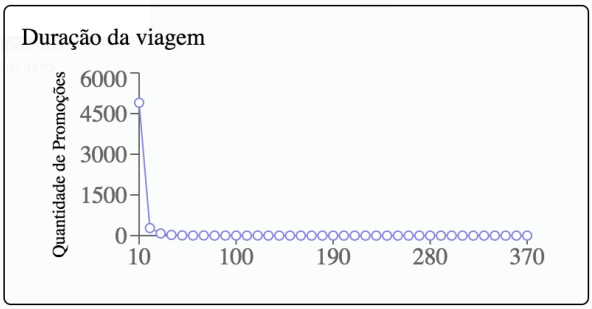


Figura 3.10: Gráfico de linha encontrado na ferramenta de visualização dos dados representando à quantidade de promoções em relação à duração da viagem.

CAPÍTULO 4

Avaliação

Neste capitulo, serão comentados os resultados obtidos após a análise e limpeza dos dados, assim como os resultados obtidos após serem aplicados os métodos de classificação de aprendizado de máquina utilizados para predição dos preços promocionais de passagens aéreas para as viagem escolhidas.

4.1 Dados

Durante o período de 01 de agosto de 2017 a 27 de janeiro de 2018, foram coletadas 11.039.783 promoções de passagens aéreas no site Melhores Destinos. Contudo, foram observadas algumas anomalias nesse conjunto de dados e dessa forma foram aplicadas técnicas de limpeza de dados como explicado na Seção 3.2. Após observar uma variação de preço anormal para os conjunto de dados e terem sido removidos esses valores anormais, o conjunto de dados teve uma diminuição na sua quantidade para 9.440.603 promoções.

Com a criação de novas variáveis no conjunto de dados, foi possível observar novos valores estranhos. A primeira nova variável que apresentou valores estranhos foi a de durações das viagens onde foram observados valores negativos. Já que não existem durações negativas de tempo, as promoções que continham valores negativos na variável de duração foram retiradas do conjunto de dados, deixando esse com um tamanho de 9.432.618 promoções restantes. A segunda nova variável que apresentou valores anormais foi a de dias antes da viagem, representando o tempo entre a postagem da promoção e o início da viagem. Novamente, foi possível observar nessa variável valores negativos para intervalo de tempo. Dessa forma, as promoções que tinham valores negativos nessa variável foram retiradas do conjunto de dados que apresentou, ao final da remoção, uma quantidade de 9.426.850 promoções.

4.2 Predição

Com os dados da viagem separados em dois formatos diferentes, onde no primeiro formato os dados só estão separados pelo tipo da viagem, sendo eles viagens nacionais e internacionais, sendo esses considerados um conjunto único dados. No segundo formato, os dados foram separados em pares de cidades, onde para cada par origem e destino foi criado um conjunto de dados, sendo esse considerado um conjunto separado por viagens. Foram escolhidos cinco pares de cidades para viagens nacionais e cinco para viagens internacionais (Tabela 3.2).

Após terem sido aplicados todos os métodos de aprendizagem de máquina nos dados selecionados, foram obtidos os resultados e verificados os erros a partir da aplicação de fórmulas

que verificam a diferença entre o resultado esperado e o obtido.

No conjunto único de dados, foram aplicados os métodos *Random Forest* e Regressão Linear para a predição do preço de passagens aéreas. O tamanho do conjunto de dados de treino e teste para viagens do tipo internacionais foi de 74.499 promoções e 29.956 promoções respectivamente. Já para o conjunto de dados das viagens do tipo nacionais, os dados de treino foram com tamanho de 106.386 e os dados de teste com tamanho de 40.412 promoções. Aos resultados obtidos, aplicaram-se as funções de MRE e a MSE. Os seus valores encontram-se nas Tabelas 4.1 e 4.2

Tabela 4.1: Erros obtidos com os método de MRE e MSE onde foram utilizados os resultados gerados pelos modelos do *Random Forest* e de Regressão Linear, aplicando-se o conjunto de dados único com viagens do tipo internacionais

Método	MRE	MSE
Random Forest	0,1451	168.505,06
Regressão Linear	0,1141	86.615,56

Tabela 4.2: Erros obtidos com os método de MRE e MSE onde foram utilizados os resultados gerados pelos modelos do *Random Forest* e de Regressão Linear aplicando-se o conjunto de dados único com viagens do tipo nacionais

Método	MRE	MSE
Random Forest	0,2287	35.824,10
Regressão Linear	0,3000	33.334,74

Pode-se observar que, para o conjunto único de dados com viagens internacionais, obteve-se os menores erros quando foi aplicado o método de Regressão Linear. O oposto aconteceu para o conjunto único de dados de viagens nacionais onde os menores erros foram obtidos com os resultados do método de predição *Random Forest*.

Para os conjuntos de dados separados por viagens, aplicaram-se os modelos de predição obtidos com os métodos *Random Forest*, Regressão Linear e SVR. Os resultados dos preços das promoções gerados foram comparados com os resultados esperados no conjunto de dados do teste e calcularam-se os valores do MRE e MSE. Os resultados dos erros estão nas Tabelas 4.3 e 4.4

Observados os valores do MRE, pode-se afirmar que os modelos produzidos pelo método SVR foram os que se mantiveram mais estáveis para todas as viagens. Quando se observa o gráfico do MSE, os valores dos erros estão mais próximos nas viagens nacionais e, para as viagens internacionais, os erros produzidos pelo modelo obtido utilizando o método SVR mostram-se mais próximo de zero.

Foi feita uma comparação entre os valores dos MRE obtidos modelos dos métodos *Random Forest* e Regressão Linear para o conjunto único de dados separado por viagens em nacionais e internacionais, e para os conjuntos de dados separados por viagens com pares de cidade origem e destino. Já que o modelo gerado a partir do SVR necessita de um grande processamento, não

Tabela 4.3: Erros obtidos ao se utilizar o método MRE nos resultados gerados pelos modelos dos métodos *Random Forest*, Regressão Linear e SVR ao se utilizarem os conjuntos de dados separados por viagens

Viagem (Origem - Destino)	Random Florest	Regressão Linear	SVR
Campo Grande - São Paulo	0,2799	0,4432	0,3462
Cuiabá - São Paulo	0,2379	0,2690	0,2585
Florianópolis - Rio de Janeiro	0,2345	0,4184	0,2873
Goiânia - Rio de Janeiro	0,2462	0,2840	0,2363
Salvador-São Paulo	0,1634	0,1542	0,1303
Rio de Janeiro - Miami	0,2566	0,1801	0,1715
Rio de Janeiro - Orlando	0,992	0,1068	0,1128
São Paulo - Madri	0,1135	0,2101	0,1003
São Paulo - Miami	0,1230	0,1097	0,1018
São Paulo - Nova York	0,1748	0,1641	0,1772

foi possível obter as predições desse modelo para o conjunto único de dados e dessa forma não há uma comparação para os resultados desse modelo. As comparações entre os valores dos MRE encontram-se nas Tabelas 4.5 e 4.6.

Para os dois modelos, *Random Forest* e Regressão Linear, houve uma melhora nos erros quando aplicados para os conjuntos de dados gerais, havendo apenas a divisão entre viagens nacionais e internacionais. Podem ser vistos erros ainda menores quando utilizado o modelo obtido a partir do método de Regressão Linear.

É importante ressaltar que, apesar de os erros terem valores muito acima do esperado, que seriam erros mais próximos do valor zero, eles apresentam valores aceitáveis. Para diminuir os erros, seria necessário possuir dados que não estão disponíveis na internet, já que os preços das passagens aéreas variam com relação à quantidade de assentos disponíveis na aeronave, o valor do combustível, dentre outros fatores que só a companhia aérea tem a informação (Freitas, 2012).

Tabela 4.4: Erros obtidos ao se utilizar o método MSE nos resultados gerados pelos modelos dos métodos *Random Forest*, Regressão Linear e SVR ao se utilizarem os conjuntos de dados separados por viagens

Viagem (Origem - Destino)	Random Florest	Regressão Linear	SVR
Campo Grande - São Paulo	67.417,80	59.463,08	57.648,37
Cuiabá - São Paulo	36.412,73	25.044,84	29.537,51
Florianópolis - Rio de Janeiro	40.233,49	48.491,80	42.384,94
Goiânia - Rio de Janeiro	38.810,50	231.713,12	33.971,44
Salvador-São Paulo	9.382,67	8.216,48	5.366,46
Rio de Janeiro - Miami	559.507,42	254.144,41	230.252,81
Rio de Janeiro - Orlando	67.321,37	60.751,57	68.283,84
São Paulo - Madri	131.801,19	259.491,17	73.298,17
São Paulo - Miami	114.998,09	82.940,81	77.215,20
São Paulo - Nova York	152.589,95	138.235,02	143.775,36

Tabela 4.5: Erros obtidos a partir da aplicação do modelo do *Random Forest* para os conjuntos de dados separados por viagem e para o conjunto único de dados

Viagem (Origem - Destino)	Erros Dados de Pares de Cidade	Erros Dados Gerais
Campo Grande - São Paulo	0,2799	0,2596
Cuiabá - São Paulo	0,2379	0,2412
Florianópolis - Rio de Janeiro	0,2345	0,2296
Goiânia - Rio de Janeiro	0,2462	0,2553
Salvador-São Paulo	0,1634	0,1866
Rio de Janeiro - Miami	0,2566	0,2497
Rio de Janeiro - Orlando	0,992	0,1277
São Paulo - Madri	0,1135	0,1012
São Paulo - Miami	0,1230	0,1140
São Paulo - Nova York	0,1748	0,1720

Tabela 4.6: Erros obtidos a partir da aplicação do modelo da Regressão Linear para os conjuntos de dados separados por viagem e para o conjunto único de dados

Viagem (Origem - Destino)	Erros Dados de Pares de Cidade	Erros Dados Gerais
Campo Grande - São Paulo	0,4432	0,4318
Cuiabá - São Paulo	0,2690	0,3330
Florianópolis - Rio de Janeiro	0,4184	0,3618
Goiânia - Rio de Janeiro	0,2840	0,2407
Salvador-São Paulo	0,1542	0,1457
Rio de Janeiro - Miami	0,1801	0,1227
Rio de Janeiro - Orlando	0,1068	0,872
São Paulo - Madri	0,2101	0,767
São Paulo - Miami	0,1097	0,900
São Paulo - Nova York	0,1641	0,1942

CAPÍTULO 5

Estudo de Caso do Sistema

Este Capítulo tem como objetivo analisar os preços promocionais para viagens através das técnicas de visualização de dados existentes no sistema.

Para a primeira análise, ao escolher a origem da viagem em uma cidade brasileira, pode-se visualizar no mapa, uma grande diferença na variação de preços entre viagens com destinos nacionais (cidades brasileiras) e viagem com destinos internacionais (cidades em outros países). Para exemplificar, escolheu-se São Paulo como origem da viagem e observamos uma variação de preços entre R$ 287,62 (Rio de Janeiro, Brasil) e R$ 1.866,60 (Carajás, Brasil) nas viagens nacionais. Para as viagens internacionais a variação é de R$ 747,47 (Santiago, Chile) a R$ 4.400,21 (Krabi, Tailândia). Esses valores podem ser vistos na Figura 5.1.

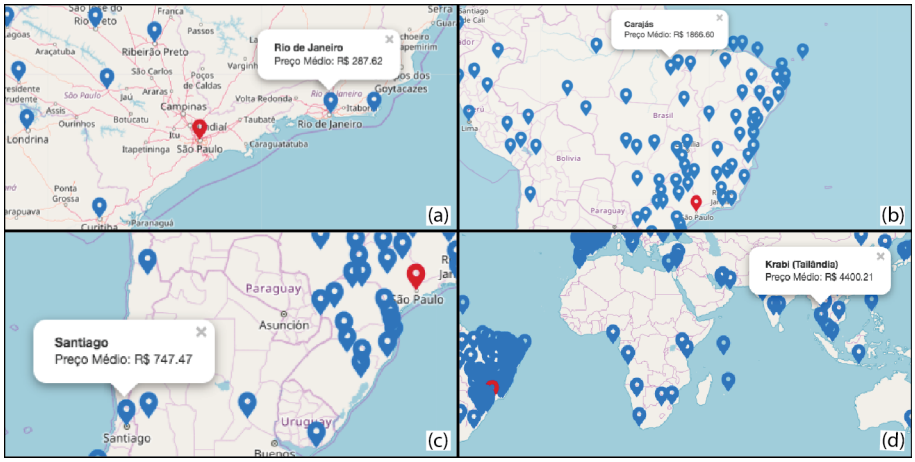


Figura 5.1: Preço médio das promoções. Em todas as imagens, o ponto destacado em vermelho marca a localização de São Paulo, a cidade de origem. (a) Preço médio da promoção para o Rio de Janeiro, como cidade destino. (b) Preço médio da promoção para Carajás, como cidade destino. (c) Preço médio da promoção para Santiago, Chile, como cidade destino, (d) Preço médio da promoção para Krabi, Tailândia, como cidade destino

Em uma segunda análise, quando foram escolhidas origens distintas, percebemos uma variação de quantidade de destinos sugeridos. Ao selecionar São Paulo como cidade de origem, observa-se uma grande quantidade de sugestões de destinos (Figura 5.2). Já quando foi selecionada Petrolina, a quantidade de destinos sugeridos que aparece no mapa é muito menor (Figura 5.3). Essa diferença da quantidade de sugestões de destinos tem duas possíveis explicações. Uma das explicações é o tamanho do aeroporto (quantidade de voos suportados durante um dia de funcionamento). Apesar do aeroporto de Petrolina ser muito importante para

o Nordeste, o seu tamanho não permite um fluxo de aeronaves igual aos aeroportos de São Paulo (ANAC (2015)). Porém, ao ser escolhido Londres como origem da viagem, observam-se somente duas cidades sugeridas como destino no mapa (Figura 5.4), mesmo sabendo que os aeroportos de Londres comportam um fluxo maior de aeronaves do que os aeroportos de São Paulo (CAA-UK (2015)). Essa baixa quantidade de sugestões de destinos dá-se pela segunda explicação, pois a quantidade de dados coletados para cidades de origem fora do território brasileiro foi muito baixa, uma vez que o site Melhores Destinos (onde os dados foram coletados, conforme explicado na Seção 3.1) é brasileiro e está focado em oferecer promoções para viagens que se originam no Brasil.

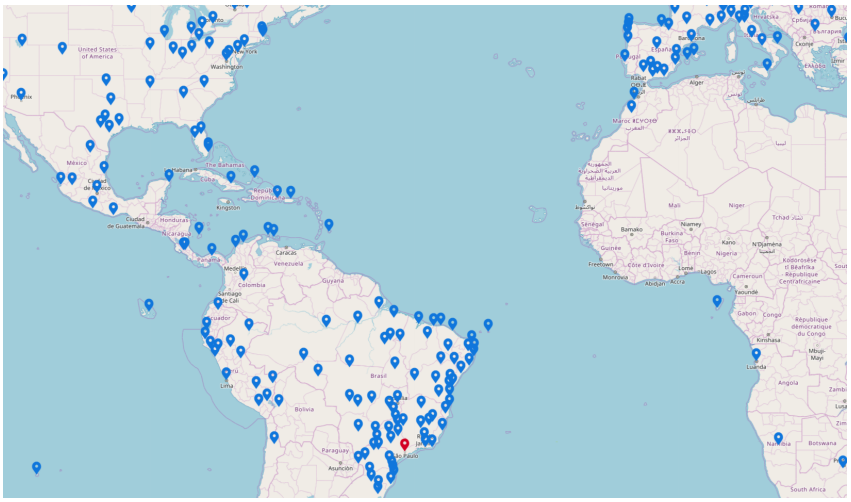


Figura 5.2: Possíveis destinos (pontos destacados em azul) para viagens com origem em São Paulo (ponto destacado em vermelho)

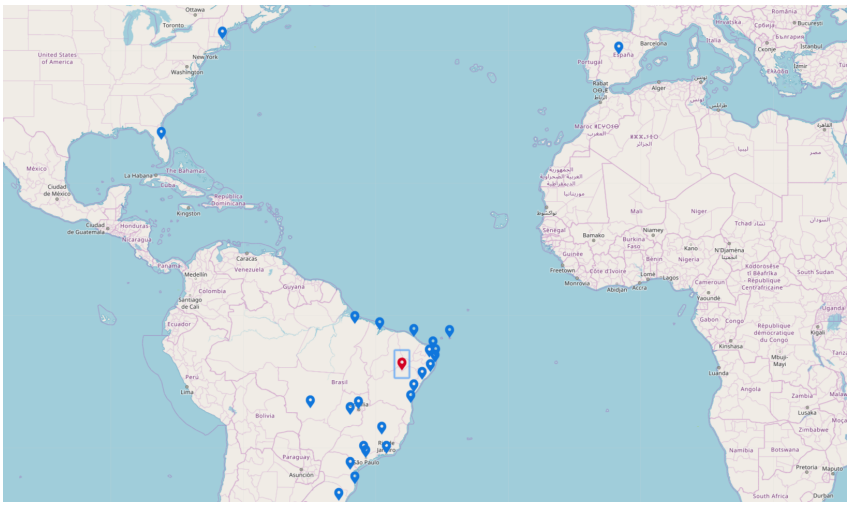


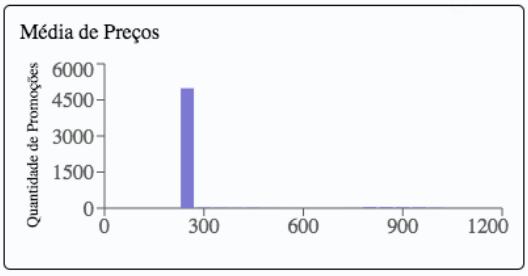
Figura 5.3: Possíveis destinos (pontos destacados em azul) para viagens com origem em Petrolina (ponto destacado em vermelho)

Figura 5.4: O ponto destacado em vermelho marca a localização de Londres, a cidade de origem. Os pontos destacados em azul são as possíveis cidades destino

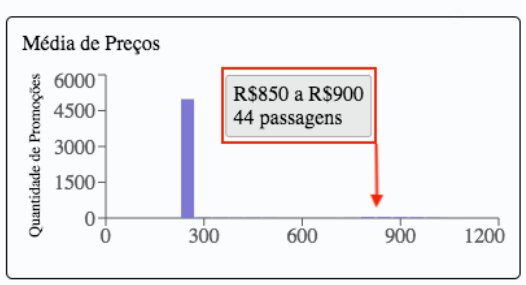
Para terceira análise, foram observados os gráficos obtidos quando escolheu-se a viagem entre São Paulo, como cidade de origem, e Rio de Janeiro, como cidade destino e requisitamos as informações completas da viagem (conforme explicado na Subsubseção 3.4.3.2). O gráfico Média de Preços mostra apenas uma barra (Figura 5.5a), pois quase todas as promoções (4893 ofertas de passagens promocionais) estão no intervalo de preço entre R$ 250,00 e R$ 300,00. É possível observar que há algumas passagens promocionais em outros intervalos de preços como mostrados na Figura 5.5b, entretanto essa quantidade é muito pequena comparada ao intervalo de maior quantidade. Para mostrar que o gráfico Média de Preços tem um formato particular na viagem entre São Paulo e Rio de Janeiro, foram selecionadas as cidades de São Paulo e Recife, origem e destino da viagem, respectivamente. Para essa nova viagem, o gráfico Média de Preço mostra uma maior variação na distribuição das promoções de passagem aérea, como é possível observar na Figura 5.5c.

Continuando a análise das informações obtidas nos gráficos da viagem entre São Paulo e Rio de Janeiro, pode-se observar que os gráficos de Dia da Semana da Compra (Figura 5.6) e Mês da Compra (Figura 5.7) apresentam o valor zero para a quantidade de promoções em alguns dias da semana e meses (destacados com as setas vermelhas). Essa falta de informação ocorre porque o site Melhores Destinos não forneceu nenhuma promoção nos períodos de tempo onde a quantidade tem o valor igual a zero, impossibilitando assim a coleta dos dados.

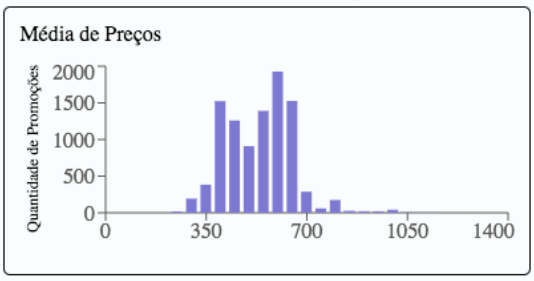
No gráfico de Mês da Compra da viagem entre São Paulo e Rio de Janeiro, observa-se uma grande quantidade de ofertas de promoções entre os meses de julho e setembro (Figura 5.7). Juntamos essa informação com a do gráfico de Dias Antes da Viagem, que mostra que a maior quantidade de ofertas de promoções ocorre entre 70 a 80 dias antes da viagem (equivalente a dois a três meses) como mostrado na Figura 5.8. Dessa forma, o gráfico de Mês da Viagem mostra as maiores quantidade de promoções entre os meses de setembro e dezembro e também o mês de janeiro (Figura 5.9).



(a) Gráfico com Rio de Janeiro sendo a cidade destino da viagem



(b) Destaque ao segundo intervalo com maior quantidade de promoções



(c) Gráfico com Recife sendo a cidade destino da viagem

Figura 5.5: Gráficos Média de Preço para viagens com origem em São Paulo

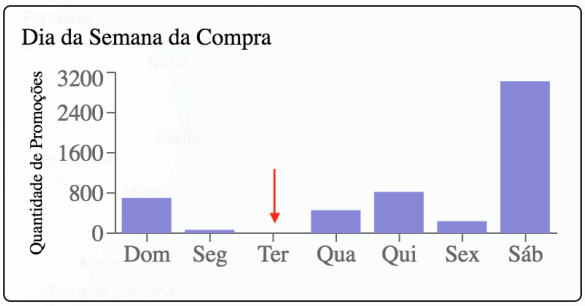


Figura 5.6: Gráfico Dia da Semana da Compra para viagem com origem em São Paulo e destino Rio de Janeiro. Setas vermelhas destacam dia da semana com quantidade de promoções igual a zero

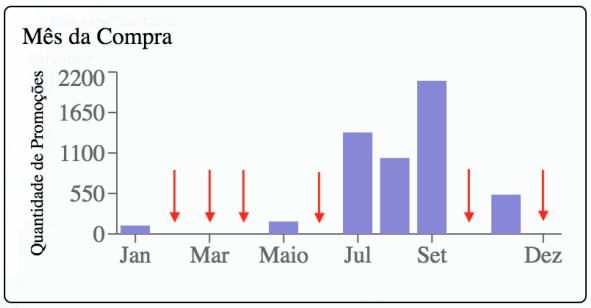


Figura 5.7: Gráfico Mês da Compra para viagem com origem em São Paulo e destino Rio de Janeiro. Setas vermelhas destacam meses com quantidade de promoções igual a zero

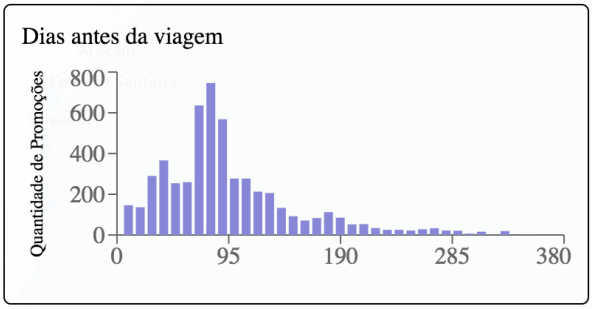


Figura 5.8: Gráfico Dias Antes da Viagem com origem em São Paulo e destino Rio de Janeiro.

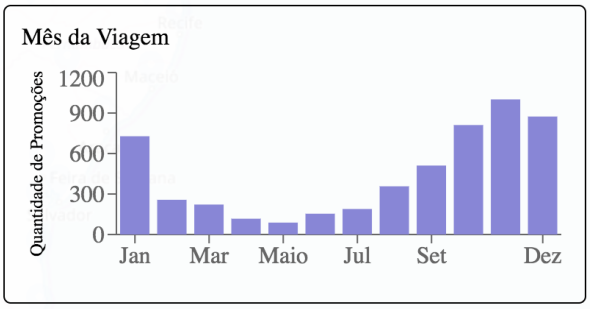


Figura 5.9: Gráfico Mês da Viagem com origem em São Paulo e destino Rio de Janeiro.

CAPÍTULO 6

Conclusão

Neste trabalho, foi apresentado um sistema de coleta de dados consolidado e uma aplicação eficiente de técnicas de limpeza de dados na retirada de *outliers* como os preços de passagens aéreas desproporcionais à realidade, com valores acima de R$ 3.000.000,00 ou abaixo de R$ 200,00 ou retirando valores negativos para as variáveis de duração da viagem e quantidade de dias antes da viagem. As escolhas da organização dos dados para aplicar os métodos de predição, separando os dados para cada dupla de viagens, mostrou-se eficiente no *Mean Absolute Error* e o *Mean Squared Error*. Apesar dos erros apresentarem valores elevados, a sua melhora depende de dados que estão sobre o domínio das companhias aéreas, não sendo disponibilizados na internet.

Comparando a ferramenta de visualização proposta neste trabalho com as já existentes no mercado, pode-se observar que o mapa é um diferencial, comparado com a maioria dos sites que oferecem promoções de passagens aéreas e que foi apresentada grande quantidade de informação através de gráficos de barra e linha para auxiliar na decisão do usuário na hora da compra de uma passagem aérea. Para trabalhos futuros com relação à ferramenta de visualização, é interessante fazer teste com usuários para validar a interação e o design da ferramenta.

É possível identificar pontos de melhoras no projeto que podem ser feitas em trabalhos futuros. A melhora dos dados é sempre importante e ela pode acontecer tanto na parte da coleta, adicionando as informações do horário da viagem ou o horário em que a promoção foi posta, tanto na limpeza, buscando observar outros *outliers* que prejudiquem as analises feitas nos dados. Na predição dos preços das promoções, podem-se adicionar novas variáveis para auxiliar na criação dos modelos.

Referências Bibliográficas

Adaniya, M. H. A. C., & Proença, M. L. (2009). Extração de informações na web. In *Cadernos de informática* (Vol. 4).

ANAC. (2015, October). *Aeroportos.* (Disponível em: <http://www.aviacao.gov.br/assuntos/aeroportos>. Acesso em: 3 jun. 2018)

CAA-UK. (2015, October). *Data and analysis.* (Disponível em: <http://www.caa.co.uk/data-and-analysis/>. Acesso em: 3 jun. 2018)

Camilo, C. O., & da Silva, J. C. (2009, August). *Mineração de dados: Conceitos, tarefas, métodos e ferramentas* (Tech. Rep.). Instituto de Informática Universidade Federal de Goiás.

Chartblocks. (2018a, June). *When to use a bar chart.* (Disponível em: <https://www.chartblocks.com/en/support/faqs/faq/when-to-use-a-bar-chart>. Acesso em: 6 jun. 2018)

Chartblocks. (2018b, June). *When to use a line chart.* (Disponível em: <https://www.chartblocks.com/en/support/faqs/faq/when-to-use-a-bar-chart>. Acesso em: 6 jun. 2018)

de Jonge, E., & van der Loo, M. (2013). An introduction to data cleaning with r. In (p. 33). Statistics Netherlands.

Etzioni, O., Tuchinda, R., Knoblock, C. A., & Yates, A. (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price. In *Kdd.*

Fabian Pedregosa, A. G. V. M. B. T. O. G. M. B. P. P. R. W. V. D. J. V. A. P. D. C. M. B. M. P. É. D., Gaël Varoquaux. (2011, October). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Freitas, A. (2012). *Preço de passagem de avião muda a cada minuto; veja 10 dicas para economizar.* (Disponível em: <https://economia.uol.com.br/noticias/redacao/2012/09/27/preco-de-passagem-de-aviao-e-calculado-minuto-a-minuto-veja-10-dicas-para-economizar.htm>. Acesso em: 17 jun. 2018)

Hijmans, R. J., Williams, E., & Vennes, C. (2017). Package 'geosphere' [Computer software manual].

Maps, G. (2018, March). *Primeiros passos.* (Disponível em: <https://developers.google.com/maps/documentation/geocoding/start?hl=pt-br>. Acesso em: 18 jun. 2018)

Marques, Y. B. (2015). *Mirnacle: Aprendizagem de máquina utilizando smote e random forest para prover aumento da seletividade na predição ab initio de pre-mirnas* (Unpublished master's thesis). Universidade Federal de Viçosa.

Medlock, J. (2017, March). *An overview of mongodb & mongoose.* (Disponível em: <https://medium.com/chingu/an-overview-of-mongodb-mongoose-b980858a8994>. Acesso em: 6 jun. 2018)

Muniz, S. R. (2018). Fundamentos da matemática ii. In (p. 264 - 285).

N. Marcelo, P. D. N. I., S. Amarindo. (2003, December). Banco de dados e sistema de apoio à decisão para as culturas de milho e soja. *Embrapa Informática Agropecuária*.

Pandas. (2018, June). *Pandas relase note.* (Disponível em: <https://pandas.pydata.org/pandas-docs/stable/release.html>. Acesso em: 17 jun. 2018)

Peternelli, L. A. (2004, March). *Regressão linear e correlação.*

Santos, G. (2016, June). *Node.js — o que é, por que usar e primeiros passos.* (Disponível em: <https://medium.com/thdesenvolvedores/node-js-o-que-%C3%A9-por-que-usar-e-primeiros-passos-1118f771b889>. Acesso em: 6 jun. 2018)

Seo, S. (2002). *A review and comparison of methods for detecting outliers in univariate data sets* (Unpublished master's thesis). Kyunghee University.

Todesco, F., Lovadine, D., de Andrade Januário Bettini, H. F., & Vassallo, M. D. (2008). Web pricing de companhias aéreas durante uma guerra de preços: onde estão os descontos? *Journal of Transport Literature*, *2*(1).

Xplenty. (2017). *The sql vs nosql difference: Mysql vs mongodb.* (Disponível em: <https://medium.com/xplenty-blog/the-sql-vs-nosql-difference-mysql-vs-mongodb-32c9980e67b2>. Acesso em: 6 jun. 2018)

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis) pelo autor (?@cin.ufpe.br).