Universidade de São Paulo

Instituto De Ciências Matemáticas e de Computação

Trabalho 2 - Análise e Modelagem Preditiva de Dados

Estudo sobre enchentes em áreas urbanas

Carlos Henrique Hannas de Carvalho	nº USP: 11965988
Carlos Nery Ribeiro	nº USP: 12547698
Erik Melges	nº USP: 12547399
Gabriel Ribeiro Rodrigues Dessotti	nº USP: 12547228
Gustavo Barbosa Sanchez	nº USP: 11802440
Pedro Antonio Bruno Grando	nº USP: 12547166
Pedro Manicardi Soares	nº USP: 12547621

Prof. Dra. Solange Oliveira Rezende

SCC0630 - Inteligência Artificial

Sumário

Lista de Figuras

1 Introdução

O período da Guerra Fria (1947-1991) foi marcado por uma ascensão tecnológica devido aos seguintes fatores: corridas armamentistas e espaciais, ascensão econômica e, principalmente, no âmbito de computação e telecomunicações. Esse progresso causou certa globalização tecnológica, o que está ligado ao aumento da quantidade de dados nos dias atuais. Essa crescente exponencial de dados requer análises computacionais para extrair informações e conhecimentos sobre um determinado assunto, uma vez que a análise manual é inviável.

A faculdade de extrair informações e conhecimentos, citada acima, no âmbito de inteligência artificial, é denominada "mineração de dados". Ela auxilia o ser humano a fazer escolhas mediante a um problema identificado - no caso desse relatório, faz-se a análise e modelagem preditiva de dados, para prever a probabilidade de ocorrência e poder auxiliar na tomada decisões em relação às enchentes em áreas urbanas. O contexto de análise de dados em relação aos alagamentos em cidades influencia políticas públicas, para minimizar os danos sobre a população que sofre esse desastre.

Os tópicos a seguir discutem sobre a identificação do problema (enchentes em áreas urbanas) e quais fatores motivaram essa análise, bem como a implementação de algoritmos e análise de resultados, oriundos de *datasets* e técnincas de *machine learning*, para, possivelmente, determinar a probabilidade de ocorrência de enchente nas regiões e, assim, minimizar, através de políticas públicas, os efeitos colaterais.

2 Problemática e Dados

2.1 Problemática

As últimas décadas foram marcadas por inúmeros desastres ambientais no Brasil, ocasionando prejuízos financeiros e mortes para o Estado. Nos últimos 10 anos, mais especificamente, registrou-se alguns desastres em relação às enchentes em áreas urbanas, sobretudo no período de verão no Brasil - durante essa época, a precipitação é mais intensa e muitas cidades brasileiras não suportam a quantidade de chuva.

Um evento recente e significativo, que motivou o estudo de dados em relação às enchentes, foi o transbordamento do rio Guaíba, no final de abril de 2024, na capital Porto Alegre, localizada no Rio Grande do Sul. A inundação, na verdade, afetou grande parte do estado, e não apenas a capital, e causou óbitos e desabrigo sobre uma parcela significativa da população gaúcha. Um exemplar de notícia, pós-desastre no estado, pode ser visualizada em [3].

2.2 Dados

Apesar de uma motivação brasileira, não é apenas o Brasil que sofre com enchentes em áreas urbanas - isso é uma problemática mundial. A fim de abranger o máximo de casos possíveis, analisou-se, através de algoritmos, um *dataset* global, presente em [4], para extração de conhecimento sobre o tema em questão.

O *dataset*, *Regression with a Flood Prediction Dataset*, consta com 1.117.957 exemplares e algumas *features*, como “Intensidade das Monções”, “Drenagem da Topografia”, “Gestão dos Rios”, “Desmatamento”, “Urbanização”, “Mudança Climática”, “Qualidade das Barragens”, “Assoreamento”, “Práticas Agrícolas”, “Invasões”, “Preparação Ineficaz para Desastres”, “Sistemas de Drenagem”, “Vulnerabilidade Costeira”, “Deslizamentos de Terra”, “Bacias Hidrográficas”, “Infraestrutura Deteriorada”, “Índice de População”, “Perda de Áreas Úmidas”, “Planejamento Inadequado”, “Fatores Políticos”e “Probabilidade de Inundação”.

3 Algoritmos e Implementação

3.1 Algoritmos

Nesse trabalho, utilizou-se, principalmente, o algoritmo de árvores de decisão para a elaboração do preditor principal, além de treinar alguns outros modelos que serão apresentados ao longo dessa seção.

Primeiramente, falando do algoritmo principal, para melhorar a capacidade de previsão das árvores de decisão, utilizou-se uma Floresta Aleatória (*Random Forest*): método de aprendizado por conjuntos, que consiste em produzir múltiplas árvores de decisão e extrair a média da regressão em cada uma delas para compor o resultado final. Esse método utiliza *Bootstrap Aggregating* para compor diferentes conjuntos de treinamento, a partir de um mesmo dataset e, assim, poder gerar diferentes árvores de decisão. Por utilizá-lo, as florestas aleatórias tendem a ser muito mais estáveis e precisas, removendo o comum característica de *overfitting* nas árvores de decisão. Para além, é interessante notar que no caso deste último algoritmo, a seleção das características na árvore é inclinada àquelas que “mais fortemente descrevem o conjunto”, em teoria, mas no caso das florestas, a seleção é aleatória, o que contribui também para uma melhoria na diversidade da árvore e uma distribuição menos enviesada nos resultados da regressão.

Em se tratando de outras formas de regressão, utilizou-se algoritmos para estabelecer um critério de comparação e estudo dos resultados. O primeiro foi a **Regressão Linear**: um método estatístico que utiliza um conjunto de um ou mais variáveis explanatórias, que explicam um comportamento alvo através de uma função de predição linear.

Em sequência, utilizou-se o método de **K-Nearest-Neighbors**, que busca associar a um novo dado inserido a média dos K vizinhos mais próximos deles. Nesse caso, fala-se de média, pois pensamos no processo de um regressão, em um domínio contínuo.

Por fim, tem-se também outros dois tipos de regressores mais complexos: a **Regressão de Ridge** e a **Support Vector Regression (SVR)**. Ambos os modelos usam relações algébricas violentas para extrair propriedades muito úteis à regressão. Acredita-se que foge ao escopo de projeto descrevê-las, então restringir-se-á a apenas mencioná-las.

3.2 Implementação

O primeiro passo na implementação de qualquer modelo de aprendizado é o tratamento dos dados. Para isso, utilizou-se a biblioteca *Pandas* em *Python*, através do comando *DataFrame.describe()*, que extrai diversas operações, como por exemplo, o número de amostras em cada *feature*, suas médias, desvio padrão, máximo, mínimo e percentis.

Em sequência, renomeia-se as características e remove-se o identificador (id) dos dados. Após isso, calcula-se a correlação entre os dados, usando o coeficiente correlação de Pearson, que é a razão entre a covariância de duas variáveis dividido pelo produto de seus desvios padrões. A figura 1 mostra a matriz de correlação utilizada:

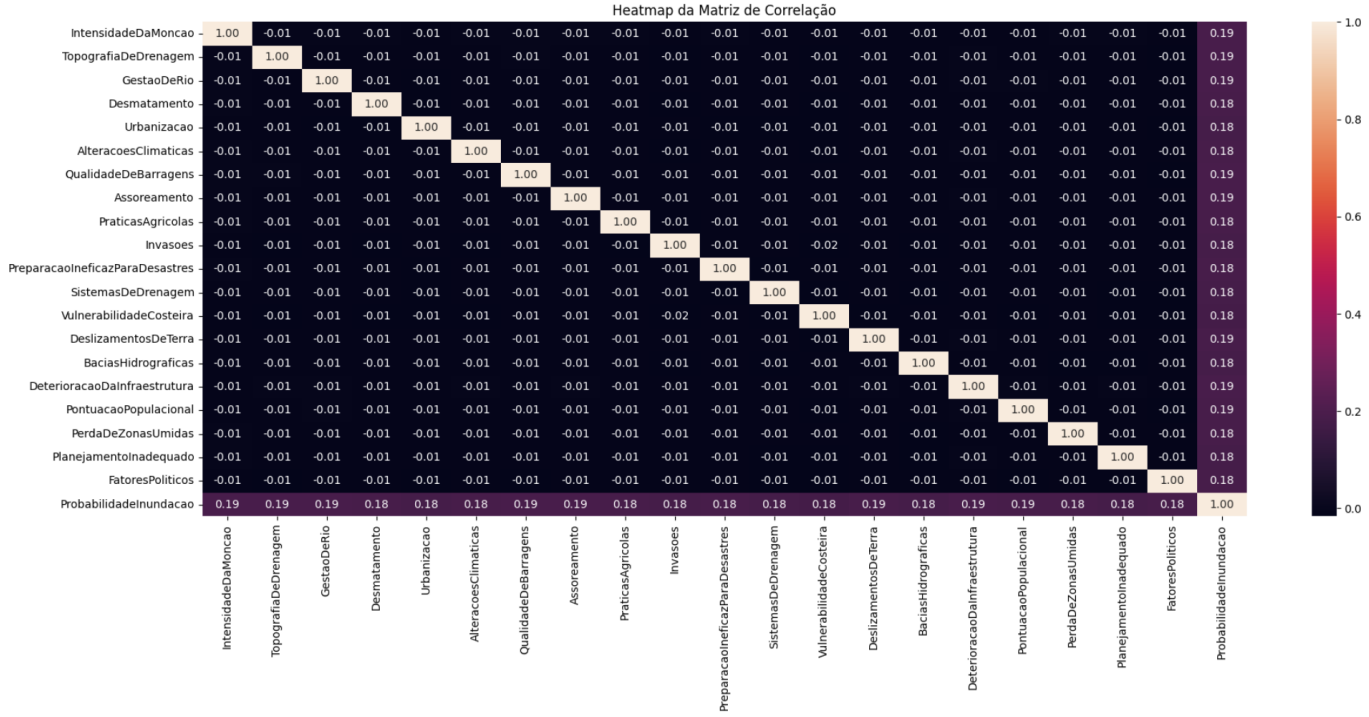

Figura 1: Matriz de correlção.

Finalizada a parte de preparação dos dados, parte-se para o treinamento da floresta aleatória, utilizando a biblioteca *SKLearn*. Primeiro, defini-se uma amostragem de 10.000. Amostra-se aleatoriamente o conjunto total de dados, extraindo o *dataset* que será usado para treinamento e teste das árvores. Extrai-se a probabilidade de inundação (uma das colunas do *DataFrame*), para definir o conjunto das dados não classificados e o conjunto das classificação. Dando sequência, faz-se a separação dos *datasets* de treinamento e teste, através da função *train_test_split*, com uma relação 0.8/0.2, respectivamente. Por fim, treina-se o modelo de *Random Forest* com 100 árvores.

É interessante que, apesar disso, para melhorar o resultado de florestas de árvores de decisão em dados com muitas características, referenciado em [5], costuma-se fazer uma seleção daqueles que podem representar ruído para os dados. Essa seleção é usada com o método *SelectKBest*, com erro qui-quadrado.

Separados os dados, faz-se um novo *dataframe*, mas agora apenas com as características mais relevantes, e é treinado um novo modelo de floresta de 100 árvores. Essa análise é feita para os K melhores, com K variando de 1 a 20

características. Por fim, são treinados os outros modelos com o conjunto de dados original.

4 Resultados e Conclusão

4.1 Resultados

A figura 2 mostra um gráfico de barras, com o erro quadrático médio de cada um dos modelos:

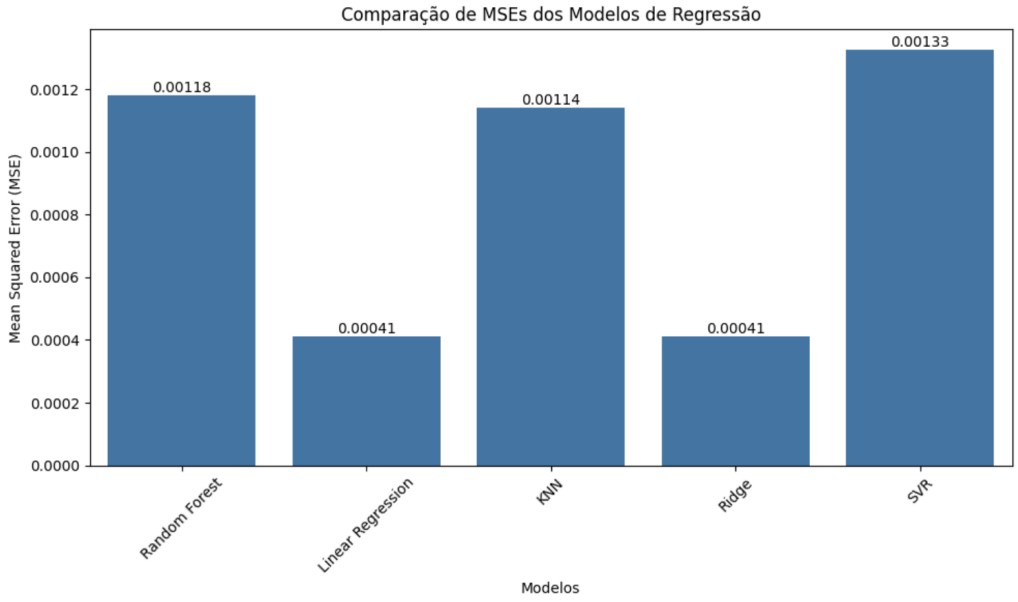

Figura 2: Comparação entre cada regressor utilizado na previsão da probabilidade de enchente.

É interessante notar que o *Random Forest*, algoritmo de interesse, performou relativamente bem, sobrepassando o KNN e o SVR (esse último muito provavelmente pela implementação que é disponibilizada pela biblioteca). A regressão linear e a regressão Ridge performaram significamente melhor, atingindo um valor absurdamente preciso. Esse resultado é muito importante, pois mostra que pequenos tratamentos na execução de um algoritmo simples, como a árvore de decisão, possuem resultados satisfatórios e competitivos com métodos muito mais complexos.

Outro ponto de intersse diz respeito ao fato de que o processamento de dados feito para melhorar a execução da floresta, na verdade, não produziu melhorias, muito pelo contrário. Na prática, o melhor conjunto foi aquele que utilizou todas as características, o que pode indicar que todas as características eram pertinentes à construção do modelo (o que era esperado, dado que o *dataset* provavelmente já havia sido pré-processado).

4.2 Conclusão

Tendo em vista tudo o que foi apresentado ao longo desse trabalho, foi possível adquirir familiaridade com o processo de desenvolvimento de uma solução em inteligência artificial utilizando conhecimentos adquiridos em sala de

aula, como o processo de mineração de dados, tratamento e finalmente a construção e treinamento de modelos.

Aliado a isso, foi exemplificado o poder e alcance de algoritmos de aprendizado de máquinas, que além de diversos, em príncipios e aplicações, tratam diferentemente um mesmo conjunto de dados e podem ser associados com o uso de meta-algoritmos, como o *Bootstrap Aggregating*. Os resultados da execução da Floresta também foram extremamente satisfatórios, constituindo um modelo muito bem comportado para o *dataset* em questão.

Dessa forma, foi possível verificar na prática o poder de soluções em inteligência artificial, sua diversidade e sua aplicabilidade na resolução de problemas concretos.

Referências

[1] S.Russel; P. Norvig. "Inteligência Artificial - Uma Abordagem Moderna". 4ª ed., 2022.

[2] Rezende, Solange. "Visão geral de Mineração de Dados - slide de aula". 2024.

[3] Notícia RS - G1. Disponível em: https://g1.globo.com/rs/rio-grande-do-sul/noticia/2024/06/23/3-a-cada-10-moradores-do-rs-pensam-em-mudar-de-casa-em-razao-de-eventos-climaticos-diz-pesquisa.ghtml

[4] *Regression with a Flood Prediction Dataset*. Disponível em: https://www.kaggle.com/competitions/playground-series-s4e5

[5] Dessi, N. & Milia, G. & Pes, B. (2013). Enhancing random forests performance in microarray data classification. Conference paper, 99-103. 10.1007/978-3-642-38326-7_15.