Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

André Vitor Kuduavski

Explorando as dificuldades da aplicação de dados do mundo real na previsão de volume de chuva com dados de radares meteorológicos

Curitiba
2023

André Vitor Kuduavski

Explorando as dificuldades da aplicação de dados do mundo real na previsão de volume de chuva com dados de radares meteorológicos

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Pr. Dr. Luiz Eduardo S. Oliveira

Curitiba
2023

Explorando as dificuldades da aplicação de dados do mundo real na previsão de volume de chuva com dados de radares meteorológicos

Exploring the difficulties of applying real-world data in predicting rainfall with weather radar data

André Vitor Kuduavski

Este artigo descreve o processo de análise de regressões utilizando dados reais e aborda com ênfase os desafios e soluções encontrados ao lidar com observações reais. O estudo utiliza um conjunto de dados coletados por radares meteorológicos no estado do Paraná fornecidos pelo SIMEPAR, com o objetivo de prever o volume de chuva. Foram realizadas análises exploratórias, identificação e tratamento de dados ausentes, remoção de *outliers*, além do balanceamento dos dados por meio de um algoritmo de *undersampling*. Em seguida, foram aplicados seis modelos de regressão para estimar o volume de chuva. Os resultados obtidos foram comparados e avaliados utilizando a métrica RMSE (Root Mean Squared Error).
Palavras-chave: Dados reais, regressões, imputação de dados, chuva

This article describes the process of regression analysis using realworld data and emphasizes the challenges and solutions encountered when dealing with real observations. The study utilizes a dataset collected by meteorological radars in the state of Paraná, provided by SIMEPAR, with the objective of predicting rainfall volume. Exploratory analyses, identification and handling of missing data, outlier removal, and data balancing through an undersampling algorithm, were performed. Subsequently, six regression models were applied to estimate the rainfall volume. The obtained results were compared and evaluated using the RMSE (Root Mean Squared Error) metric.
Keywords: Realworld data, regressions, data imputation, rainfall

1. Introdução

Para realização de estudos com objetivo de estimar variáveis através de regressões existem dois caminhos possíveis relativo ao tipo de dados a serem utilizados, esses podem ser sintéticos ou coletados diretamente do mundo real.

Nas duas situações existem desafios a serem superados para que a análise seja realizada. Dados sintéticos podem não retratar a realidade, pois são gerados a partir de algoritmos [1] baseados em hipóteses e distribuições especificas ou ainda serem criados por inteligências artificiais. Já os dados coletados diretamente do mundo real refletem a complexidade e a variabilidade inerentes às situações concretas.

O objetivo principal desse trabalho é descrever o caminho e os desafios ao utilizar dados reais para realização de análises de regressões. Serão apresentados os problemas encontrados nesse tipo de dado, as possíveis soluções que podem ser aplicadas e no final medir os resultados após as tratativas implementadas.

O tema escolhido para realização desse trabalho foi a previsão de volume de chuva utilizando medições de radares meteorológicos. Essas observações representam um desafio significativo na aplicação de técnicas de regressão devido à sua natureza complexa além de serem caracterizados por sua heterogeneidade, variabilidade temporal e espacial.

Ao longo deste artigo serão descritas as analises exploratórias realizadas nos dados, os tratamentos de problemas encontrados e os resultados obtidos nas regressões implementadas.

2. Conjunto de dados

O conjunto de dados utilizado neste trabalho consiste em medições coletadas por radares meteorológicos[2], juntamente com dados de pluviômetros, em 26 estações meteorológicas localizadas no estado do Paraná. Esses dados foram fornecidos pelo SIMEPAR (Sistema de Tecnologia e Monitoramento Ambiental do Paraná) [3], uma instituição responsável por prover à sociedade informações relacionadas ao clima, hidrologia e meio ambiente.

O *dataset* em questão é composto por 19 colunas, das quais 18 são variáveis explicativas e uma repre-

senta a variável de interesse, as quais são detalhadas na Tabela 1:

Tabela 1: Variáveis do conjunto de dados

Variável	Definição
TP_EST	Volume de chuva
EST	Estação meteorológica
TIME	Data da medição.
AZIMUTH	Ângulo de azimute.
RANGE	Alcance da medição.
UH	Componente horizontal do vento.
UV	Componente vertical do vento.
DBZH	Refletividade horizontal.
DBZV	Refletividade vertical.
KDP	Diferença de fase específica.
ZDR	Taxa diferencial de refletividade.
RHOHV	Correlação diferencial.
X	Coordenada X.
Y	Coordenada Y.
Z	Coordenada Z.
LAT	Latitude.
LON	Longitude.
ALT	Altitude.
DISTANCIA	Distância da estação de coleta.

Os registros são datados de janeiro de 2018 até dezembro de 2022 contendo um total de 2.856.379 observações que foram disponibilizados inicialmente em um arquivo em formato CSV.

3. Exploração dos dados

Nesta etapa, foi realizada uma análise inicial dos dados para identificar e tratar possíveis valores ausentes, inconsistentes ou *outliers*.

Para iniciar as análises exploratórias os dados foram importados em um banco de dados local relacional (*DuckDB*) [4] e as consultas foram feitas através de um programa desenvolvido em *Python* [5] através das bibliotecas Pandas [6], *NumPy* [7] e *Plotly* [8].

Durante a análise exploratória dos dados, foi observado que a presença de valores ausentes nas colunas coletadas pelo radar era uma característica bastante prevalente no *dataset*.

Com o intuito de compreender melhor a distribuição desses dados, foram realizadas consultas filtrando os registros que possuíam todas as variáveis preenchidas. Os resultados revelaram que cerca de 87% dos registros apresentavam pelo menos uma coluna com dados faltantes, conforme ilustrado na Figura 1.

Para prosseguir com o objetivo do trabalho, foi necessário adotar estratégias para tratar esse problema, e assim foram utilizadas técnicas de imputação de dados.

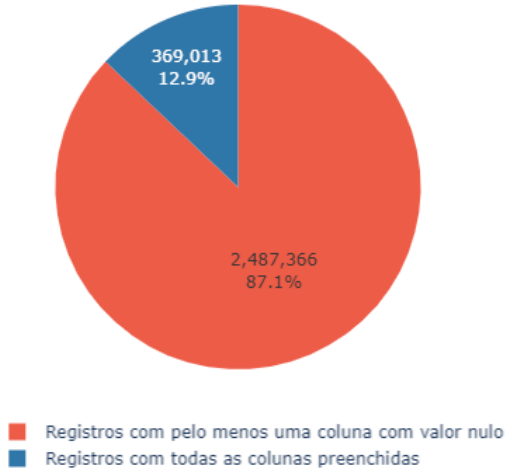

Figura 1: Distribuição dos registros antes da imputação.

Esses 22% dos dados, representado na Figura 1, seriam os registros elegíveis para realização das regressões por possuírem todas as colunas explicativas preenchidas. Ao continuar a exploração, focando nesse percentual dos dados foi possível também, encontrar problemas de balanceamento e *outliers* que eventualmente podem enviesar e prejudicar futuros resultados.

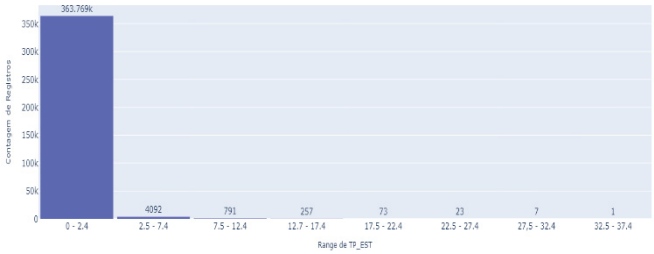

Figura 2: Distribuição da variável TP_EST

Como é apresentado na Figura 2 temos uma distribuição desproporcional da nossa variável de interesse, com grande maioria dos registros alocados no range de 0 a 2,4.

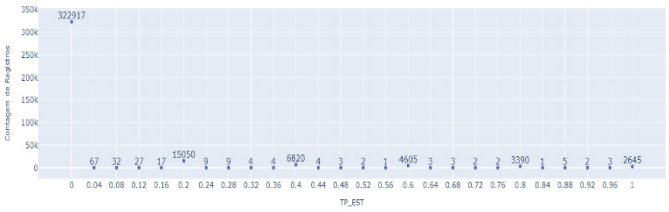

Figura 3: Range TP_EST de 0 a 1

Quando ampliado, podemos observar na Figura 3 que 87% dos registros possui valor de TP_EST igual a zero demonstrando um forte desbalanceamento na variável de interesse.

Figura 4: Distribuição dos registros depois da imputação.

4. Tratamento dos dados

Neste capitulo, serão apresentados os tratamentos realizados nos dados referentes aos pontos identificados no capítulo anterior.

4.1. Imputação de dados

Na implementação do código de imputação de dados faltantes foram levadas em consideração algumas características do *dataset*. A base é uma união de registros coletados por radares alocados em várias cidades do Paraná com aspectos geográficas diferentes que podem influenciar os valores medidos, por essa razão, os dados foram separados por estação meteorológica e ordenados pela coluna TIME.

Todo o processo foi realizado via instruções no banco de dados executadas em um algoritmo implementado em Python. Para poder aplicar a imputação em todas as colunas, a execução foi realizada em *loop* verificando nos registros quais variáveis estavam com valores ausentes.

A imputação de dados teve duas etapas e a primeira foi baseada em calcular a média dos valores das linhas anterior e posterior de cada registro que apresentou alguma coluna nula.

Para a segunda etapa, os registros que ainda permaneceram com variáveis nulas o preenchimento foi baseado na média dos valores de cada coluna, mas se restringindo a registros que estiverem dentro da mesma hora de coleta. Assim foi possível passar de 12% para 22% de dados que podem ser aproveitados nesse estudo como visto na figura 4:

4.2. Outliers

Para o tratamento de *outliers* foi aplicado um filtro na consulta dos dados para limitar os resultados, removendo os registros com baixa frequência de alguns valores de TP_EST.

Para não restringir demais os dados foi definido uma linha de corte para a remoção de *outliers*. Ao consultar os dados de uma estação são removidos do resultado registros com TP_EST com uma frequência menor ou igual a 10, dessa forma, foram removidos os *outliers* mais expressivos.

4.3. Balanceamento

Em relação ao desbalanceamento encontrado nos dados foi implementado um algoritmo de *undersampling* que tem como objetivo reduzir o número de observações da classe majoritária [9].

Na aplicação dessa técnica foi feita a divisão dos dados em dois grupos: registros com valor de TP_EST igual a zero e registros com TP_EST maior que zero. Nesse caso, o objetivo final da aplicação do *undersampling* é manter a mesma proporção de registros entre os dois grupos.

O que o algoritmo faz durante a execução é segregar os dados em dois *dataframes*, e em seguida é feito o cálculo de diferença da quantidade de registros entre eles, e por fim, são selecionadas linhas de forma aleatória para serem retiradas do grupo majoritário, deixando assim, o *dataset* numa proporção 50/50 de registros com TP_EST igual a zero e maiores que zero.

Essa função foi implementada das seguintes formas: A primeira faz o balanceamento no *dataset* de uma só vez com todos os dados, já a segunda leva em consi-

deração as estações meteorológicas, nessa forma, as execuções são realizadas separadamente para cada estação, gerando 26 *dataframes* balanceados (Tabela 2).

Tabela 2: Resultados do balanceamento em quantidade de registros

Estação	Antes	Depois
AguasVere	17836	7312
Altonia	22061	4884
AssisChateaubriand	79852	7512
BaixoIguacu	41236	7702
BelaVistaJusante	1840	810
BoaVistaAparecida	35869	7482
CampoMourao	23580	6540
CoronelDomingosSoares	13952	7186
DerivacaoRioJordao	1892	774
FozIguacuItaipu	30677	6556
Guaira	5216	898
LaranjeirasSul	21395	6878
Loanda	12189	4798
Palotina	13404	978
Paranavai	13016	4976
PatoBranco	1578	592
PortoFormosa	72257	7578
PortoSantoAntonio	23857	7512
ReservatorioSaltoCaxias	5213	758
SaltoCaxias	39220	8010
SantaHelena	62347	7802
Segredo	15499	6692
SolaisNovo	12876	6092
Toledo	9184	1008
Ubirata	50110	7036
Umuarama	2762	632

5. Aplicação de Regressões

Neste capitulo, serão descritos os processos implementados para execução dos modelos de regressão utilizados no estudo. Para isso, foram selecionados seis modelos de regressão para tentar estimar o volume de chuva utilizando as medições do radar. Vale ressaltar que a abordagem aqui escolhida não visa realizar análises temporais, mas sim, utilizar as outras variáveis das medições colhidas pelos radares para gerar a predição.

Das regressões utilizadas, cinco são do pacote *sklearn* [10] (*LinearRegression, Ridge, Lasso, GradientBoostingRegressor* e *RandomForestRegressor*) e uma do pacote *statsmodels* [11] (*Tweedie*)

Após a imputação de dados, tratamento de *outliers* e balanceamento, restaram 65.077 registros aptos para serem utilizados nas regressões que foram realizadas de duas formas descritas a seguir.

5.1. Base completa

O primeiro experimento foi realizado diretamente nos 65 mil registros restantes. Para a seleção de variáveis foram analisadas previamente as correlações entre as colunas do *dataframe* através de um gráfico de calor.

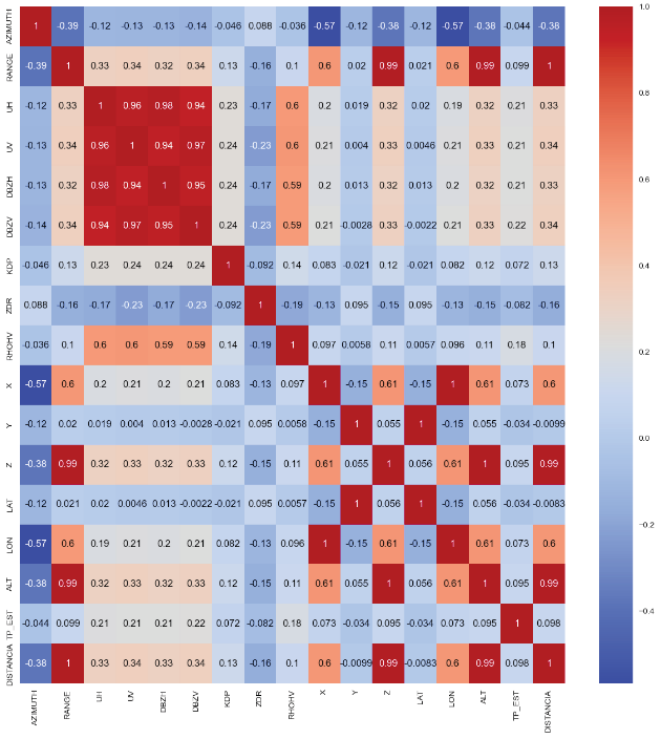

Figura 5: Matriz de correlação.

Analisando o gráfico da Figura 5 é possível entender o comportamento de correlação entre as variáveis e determinar quais podem ser removidas.

Pôdese concluir que algumas colunas teriam potencial para serem removidas do *dataframe*, pois representam a mesma informação de forma diferente, como é o caso das variáveis X, Y, Z e RANGE que acabam tendo o mesmo grau de correlação em TP_EST que os campos LON, LAT, ALT e DISTANCIA. Com isso aplicado, restaram 12 variáveis para serem utilizadas nas regressões.

Em seguida, os dados foram divididos em treino e teste com uma proporção 80/20 e submetidos aos treinamentos dos 6 modelos.

5.2. Execução por estação meteorológica

A segunda forma de aplicação das regressões foi executada separadamente para cada estação de coleta

das observações, e com isso, foram necessários alguns ajustes no *dataset*.

Para a escolha das variáveis também foi levado em consideração a correlação entre elas, mas nesse caso, como as execuções foram realizadas separadamente para cada estação meteorológica, as colunas que contém informações sobre a localização dessas estações não teriam influência sobre a variável de interesse, já que durante as execuções teriam o mesmo valor repedido em todos os registros.

Considerando o que foi dito anteriormente e após remover esses campos, o *dataset* ficou com 8 colunas: UH, UV, DBZH, DBZV, KDP, ZDR, RHOHV e a variável de interesse TP_EST.

6. Resultados e método de avaliação

Este capítulo irá apresentar os resultados obtidos nas duas formas de experimento realizadas e o método de avaliação utilizado para medir a efetividade de cada execução.

6.1. Método de avaliação

Como métrica de avaliação do melhor modelo para cada cenário foi utilizado o RMSE (*Root Mean Squared Error*) que é comumente utilizado quando se pretende penalizar erros maiores de forma mais significativa.

Para ser mais assertivo na definição e garantir um melhor resultado, foram realizadas 30 execuções completas (*split* dos dados, treino e teste) para cada cenário de experimento.

6.2. Resultados primeiro cenário

Durante o *loop* de execuções foram armazenados em listas os resultados obtidos nos testes das 6 regressões e o valor de RMSE para cada iteração. Para avaliação final foi realizado um *rank* com a média dos resultados obtidos, representados pela Figura 6, que apontou *Gradint Boosting* como melhor modelo.

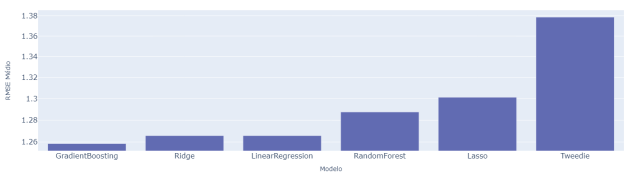

Figura 6: RMSE médio de cada modelo.

6.3. Resultados segundo cenário

Para o segundo cenário a estratégia de execução foi a mesma, os treinamentos foram executados em *loops* e seus resultados armazenados em listas para comparação posterior.

Nesse caso foram obtidos 26 resultados, um para cada estação meteorológica, descritos na Tabela 3:

Tabela 3: Resultados do balanceamento em quantidade de registros

Estação	Regressão e RMSE
AguasVere	Ridge - 1.17
Altonia	Ridge - 1.10
AssisChateaubriand	Ridge - 0.93
BaixoIguacu	GradientBoosting - 1.15
BelaVistaJusante	GradientBoosting - 0.50
BoaVistaAparecida	Ridge - 1.03
CampoMourao	Ridge - 1.28
CoronelDomingosSoares	Ridge - 1.13
DerivacaoRioJordao	RandomForest - 0.49
FozIguacuItaipu	Ridge - 1.12
Guaira	Ridge - 0.85
LaranjeirasSul	Ridge - 1.01
Loanda	Ridge - 1.22
Palotina	Ridge - 0.95
Paranavai	Ridge - 1.12
PatoBranco	RandomForest - 0.82
PortoFormosa	Ridge - 1.23
PortoSantoAntonio	GradientBoosting - 1.43
ReservatorioSaltoCaxias	Ridge - 0.78
SaltoCaxias	Ridge - 1.27
SantaHelena	Ridge - 1.13
Segredo	Ridge - 1.10
SolaisNovo	Ridge - 1.08
Toledo	Ridge - 0.95
Ubirata	Ridge - 1.23
Umuarama	Ridge - 0.71

6.4. Análise dos Resultados

Ao comparar os resultados obtidos é possível afirmar que existem modelos que melhor se ajustam aos dados quando realizado execuções separadas para cada estação meteorológica.

Devido a diferença na quantidade de registros para cada estação, casos com mais observações podem enviesar o resultado e mascarar as peculiaridades de cada região quando somente uma execução é realizada utilizando todos os dados de uma só vez.

Realizar as regressões separadamente, na maioria dos casos, também resultou em um RMSE médio menor que quando utilizado a base completa.

7. Conclusão

Ao utilizar dados provenientes do ambiente real, inúmeros aspectos podem impactar na sua composição e confiabilidade, nesse trabalho, foi demonstrado que a análise de regressões utilizando dados reais apresenta desafios, como dados ausentes, *outliers* e desbalanceamento. No entanto, foram aplicadas técnicas de imputação de dados, filtragem e *undersampling* para superar esses desafios. Os resultados destacam a importância de considerar os problemas específicos dos dados reais ao realizar análises de regressões e a relevância do uso de técnicas adequadas para lidar com essas dificuldades. Este estudo contribui para a compreensão e aplicação de análises de regressões em cenários com da-dos reais, fornecendo *insights* e orientações para futuras pesquisas nessa área.

Referências

[1] Dilmegani, C. (2022, December 10). Synthetic Data vs Real Data: Benefits, Challenges in 2023. Recuperado de https://research.aimultiple.com/synthetic-data-vs-real-data/.

[2] Barton, D. (1988). Modern Radar System Analysis. Artech Print.

[3] Sistema Meteorológico do Paraná. Recuperado de http://www.simepar.br.

[4] DuckDB. SQL OLAP database management system. Disponível em: https://duckdb.org.

[5] Python Software Foundation. Python Programming Language. Disponível em: https://www.python.org.

[6] Pandas. Biblioteca Python para manipulação e análise de dados. Disponível em: https://pandas.pydata.org.

[7] NumPy. Biblioteca Python para manipulação de arrays multidimensionais e cálculos matemáticos. Disponível em: https://numpy.org.

[8] Plotly. Biblioteca Python interativa para visualização de dados. Disponível em: https://plotly.com.

[9] Scikitlearn. Biblioteca Python de aprendizado de máquina com ferramentas para classificação, regressão, agrupamento e pré-processamento de dados. Disponível em: https://scikitlearn.org.

[10] Mohammed, R., Abdullah, M. A.,& Rawashdeh, J. (April, 2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results.

[11] Biblioteca Python para estimativa estatística e modelagem de dados. Disponível em: https://www.statsmodels.org.