XIV ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO
ANAIS DO XIV ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO - EATI
| ANO 12, N. 1 | NOVEMBRO / 2023

Abordagem com dados de distribuição pluviométrica na cidade de São Carlos - SP

Lucas de Barros Teixeira[1], Marilde Terezinha Prado Santos[1]

[1]LabDES – Laboratório de Banco de Dados e Engenharia de Software
Universidade Federal de São Carlos (UFSCar)
13.565-905– São Carlos – SP – Brazil

lucas.barros@estudante.ufscar.br,marilde.santos@ufscar.br

Abstract. *The advancement of artificial intelligence in the prediction of natural disasters, such as heavy rainfall, is addressed in this work. A quantile regression approach was used to analyze rainfall data in São Carlos - SP and estimate future rainfall data from 2016 to 2023. The obtained results showed a good approximation in relation to the actual values disclosed, with Pinball Loss indexes of 0.011 and 0.018 for the 5% and 95% confidence intervals, respectively. With a view to future development, the method is expected to be applied in different contexts.*

Resumo. *O avanço da inteligência artificial na previsão de desastres naturais, como chuvas intensas, é abordado neste trabalho. Foi utilizada uma abordagem de regressão quantílica para analisar os dados de chuvas em São Carlos - SP e estimar os dados pluviométricos futuros no período de 2016 a 2023. Os resultados obtidos apresentaram uma boa aproximação em relação aos valores reais divulgados, com índices de Perda de Pinball de 0,011 e 0,018 para os intervalos de confiança de 5% e 95%, respectivamente. Com vistas ao desenvolvimento futuro, espera-se aplicar o método em diferentes contextos.*

1. Introdução e Objetivos

O uso da inteligência artificial para prever chuvas tem crescido nos últimos anos. Com eventos pluviais mais frequentes e graves devido às mudanças climáticas, previsões precisas são essenciais. Através de algoritmos de aprendizado de máquina e análise de dados históricos, pesquisadores e órgãos públicos podem fazer previsões futuras de chuva.

Para começar a cidade de São Carlos está localizada no centro geográfico do estado de São Paulo, com uma área total de 1.136 km^2, com uma população de 254.822 habitantes segundo IBGE Cidades (2022), e com os limites de coordenadas geográficas de 47º30′ e 48º30′ Longitude Oeste e 21º30′ e 22º30′ Latitude Sul (Prefeitura de São Carlos, 2020).

Os impactos das chuvas em São Carlos são uma preocupação constante para a população e as autoridades locais. As chuvas intensas frequentemente causam alagamentos, deslizamentos de terra e danos à infraestrutura da cidade. As áreas mais afetadas são as encostas e as proximidades dos córregos. A previsão de chuvas é crucial para a pesquisa climática preventiva, pois desastres naturais têm impactos graves no ambiente, na vida humana, na economia e nas atividades sociais. É difícil se recuperar de desastres com recursos limitados. Previsões precisas ajudam a detectar chuvas ou

tempestades futuras, facilitando a gestão de desastres. A previsão é a principal preocupação técnica e científica das pesquisas.

O trabalho atual de pesquisa propõe a abordagem e uso da regressão quantílica (ou regressão quantílica condicional) para prever chuvas e tempestades com base em vários dados climáticos. Nesse método, a regressão é construída a partir dos dados de treinamento, dividindo o conjunto de dados com base em determinados critérios nos nós internos da árvore, até chegar às folhas que contêm as previsões, de forma que as árvores cresçam de forma vertical conforme Shehadeh et al. (2021).

Esta pesquisa procura preencher a lacuna encontrando soluções tecnológicas sem grandes investimentos financeiros em infraestrutura e tecnologia, de maneira criativa, incluindo os atores envolvidos no processo da prevenção a resposta a emergências. Para tanto, será utilizada a base de dados do INMET (2023) da cidade de São Carlos-SP. A seguir, a seção 2 descreve a fundamentação teórica, seção 3 menciona a abordagem proposta, as seções 4 e 5 descrevem, respectivamente, os resultados e discussões, finalizando assim com a seção 6 com a conclusão do artigo.

2. Fundamentação Teórica

Usman et al. (2023), compararam algoritmos de aprendizado de máquina, incluindo regressão linear múltipla, regressão de floresta aleatória e redes neurais. Eles usaram dados de precipitação da cidade Semarang, Indonésia, e avaliaram os modelos utilizando Erro Absoluto Médio (MAE) e Erro Médio Quadrático (RMSE). O modelo de rede neural teve o melhor desempenho para calcular o volume diário de chuva.

Em seu trabalho, Yovan Felix et al. (2019) o algoritmo de aprendizado de máquina K-means foi utilizado. A abordagem foi descrita em etapas sequenciais: coleta de dados, pré-processamento, filtragem, seleção de recursos e clusterização K-means. Por fim, é feita uma comparação entre os algoritmos K-means, Fuzzy Logic e Neuro-fuzzy Genetic. Em resumo, considerando a acurácia e o tempo de processamento, K-means obteve melhores resultados.

Dananjali et al. (2020) compararam 3 técnicas de mineração de dados para prever chuvas semanais em Badulla, Sri Lanka. Os modelos usados foram regressão linear, árvore modelo M5P e otimização mínima sequencial (SMO). O artigo apresentou várias métricas para avaliar os algoritmos, como MAE, RMSE, RRSE, RAE e DA. A árvore modelo M5P obteve os melhores resultados e uma correlação mais alta entre os valores de precipitação real e prevista.

Do mesmo modo, Li et al. (2023) realiza um estudo de caso com dados de duas cidades da China, Zhengzhou e Jiangxi. O modelo WaterLogging da RainStorm avalia a capacidade da rede em lidar com desastres, mas apenas em uma área específica. Isso dificulta a avaliação da resiliência das redes em outras regiões. É necessário considerar o impacto de tempestades com diferentes durações de chuva na rede de distribuição. Além disso, é importante estudar um método aprimorado do algoritmo de simulação de Monte Carlo para melhorar a eficiência do cálculo.

3. Apresentação da abordagem

Com respeito ao uso da inteligência artificial (IA) e à descoberta do conhecimento, essas são áreas essenciais para a extração de informações valiosas a partir de grandes conjuntos

de dados. Resumidamente, a IA envolve a identificação de padrões, correlações e tendências nos dados, enquanto a descoberta do conhecimento refere-se à aplicação desses padrões para a tomada de decisões e a criação de soluções inovadoras.

Tanto a análise de dados quanto a descoberta de conhecimento e padrões são essenciais em áreas como negócios, ciência, saúde e tecnologia. Resumidamente, esse processo envolve a seleção e preparação dos dados, aplicação de algoritmos, avaliação e interpretação dos resultados, e implementação de soluções com base neles.

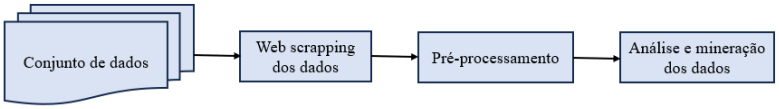

Figura 1 - Visão Geral do Processo

Em relação a linguagem de programação utilizada no estudo, decidiu-se usar o Python, por ser tratar de uma linguagem de alto nível, interpretada por script, de tipagem forte e dinâmica, segundo python Software Foundation (2023). Em segundo lugar, optamos pela plataforma Google Colab para a implementação e visualização dos dados e análise. A figura 1, demonstra a visão do geral do processo.

3.1 Coleção de conjuntos de dados

Com referência ao conjunto de dados, serão utilizados os dados do INMET (2023), como fonte primária dos dados. Em seguida, será descrito o período temporal selecionado para análise, de 2016 até 05/2023.

3.2 Web *scraping* dos dados

O web scraping é o processo de extrair dados de sites usando ferramentas automatizadas. Essa técnica é comumente usada por empresas e pesquisadores para coletar informações sobre o clima e analisar tendências. Além disso, a raspagem da web também auxilia na análise e visualização de dados, ajudando pesquisadores e órgãos públicos a tomar decisões mais assertivas (Thapelo et al., 2021).

3.3 Pré-processamento dos dados

No que diz respeito ao pré-processamento de dados climáticos, é uma etapa fundamental na análise preditiva. Os dados brutos coletados passam por técnicas de limpeza, transformação e normalização para garantir sua adequação à análise. Os dados climáticos são complexos e têm várias fontes de ruído, como falhas nos sensores, erros de medição e fenômenos climáticos extremos. Portanto, o pré-processamento dos dados é essencial para garantir análises precisas e confiáveis (Juneja & Das, 2019).

Além disso, o pré-processamento dos dados também pode incluir a seleção de recursos relevantes, a identificação de padrões e a redução da dimensionalidade dos dados. Essas técnicas ajudam a melhorar a eficiência do modelo preditivo e reduzir o tempo de processamento necessário para gerar resultados precisos. Inicialmente, foram implementadas variáveis sazonais conforme mostrado na Figura 2, e a decomposição da série temporal. O uso de técnicas de análise de séries temporais tem se tornado cada vez mais popular na área de previsões de dados climáticos. Esses métodos envolvem a análise de pontos de dados sequenciais ao longo de um período para identificar padrões e tendências nos dados.

```
# variáveis sazonais
df['hora'] = df.index.hour
df['mes'] = df.index.month
df['trimestre'] = df.index.quarter

from statsmodels.tsa.seasonal import seasonal_decompose
# salvar os componentes, decomposição da série temporal
result = seasonal_decompose(df.precipitacao_total_h,period=8760)
```

Figura 2 - Séries temporais

Em seguida, visualizou-se a volumetria de precipitação da média mensal por anos (em mm). Por meio da Figura 3, constatamos que a precipitação média se encontra concentrada nos três últimos meses e nos três primeiros meses de cada ano.

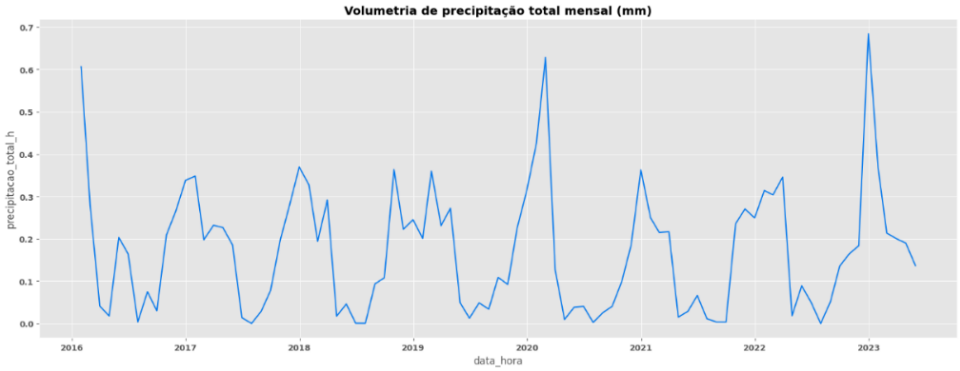

Figura 3 - Volumetria de precipitação da média mensal

Após isso, verificamos e avaliamos a demonstração da volumetria de precipitação total mensal (em mm), a Figura 4 demonstra que os anos de 2020 e 2023 apresentaram uma volumetria acentuada em comparação aos outros anos do período selecionado.

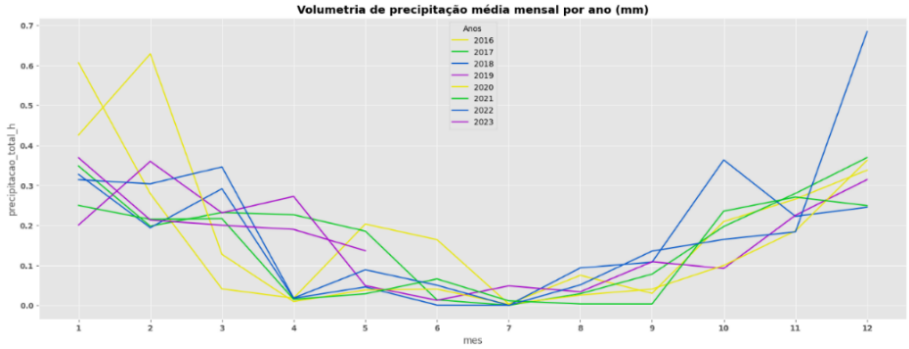

Figura 4 - Volumetria de precipitação total mensal

Não menos importante, na Figura 5 evidenciou-se o mapa de calor entre as variáveis do conjunto de dados. O coeficiente de correlação de Pearson indica a associação entre duas variáveis, conforme Karatayev et al. (2022). Na análise realizada, encontra-se uma correlação entre a precipitação e variáveis de temperatura de orvalho e dados extraídos do vento. Por último, variáveis de umidade e temperatura apresentam uma alta correlação entre si.

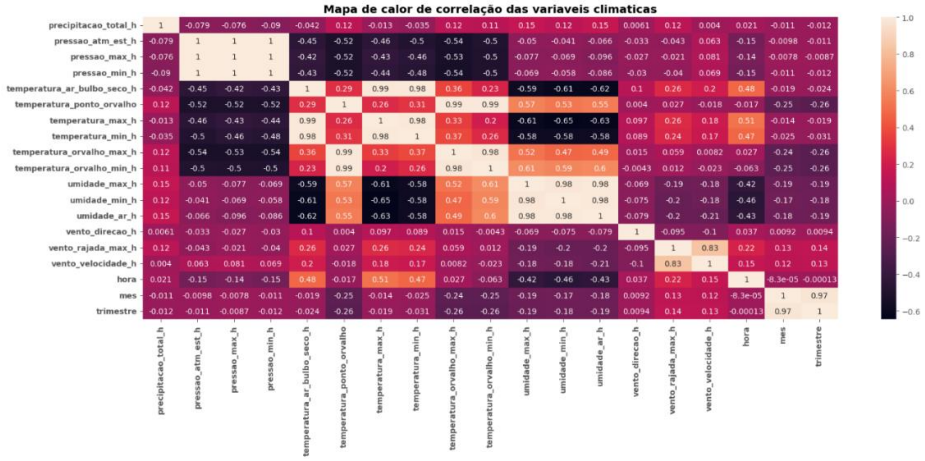

Figura 5 - Correlação de Pearson das variáveis

3.4 Métricas utilizadas no Modelo

Segundo Yang et al. (2023), a função de perda de pinball, também chamada de perda de quantil, é usada para avaliar a precisão de uma previsão de quantil. Ao contrário das previsões clássicas, onde o objetivo é ter uma previsão o mais próxima possível dos valores observados, as previsões quantílicas são intencionalmente tendenciosas. Portanto, a comparação direta entre o observado e as previsões não é satisfatória. A função de perda retorna um valor que pode ser interpretado como a precisão de um modelo de previsão quantílica. Em resumo, quanto menor a perda de pinball, mais precisa é a previsão do quantil.

3.5 Aprendizado de máquina: Regressão Quantílica

Conforme Vantas et al. (2020), a regressão logística é amplamente usada para prever chuvas em diferentes regiões. Essa técnica identifica as variáveis mais relevantes e seus impactos nas condições climáticas. É importante entender as informações meteorológicas e escolher as variáveis corretas para o modelo. A precisão da previsão pode ser melhorada com análise de dados históricos e algoritmos de aprendizado de máquina. O LightGBM foi usado para prever os dados. Para ilustrar, a Figura 6 demonstra o código criado.

```
# Models
params = {"objective" : 'quantile', "alpha" : 0.5,'force_col_wise':True} #median
mid = lgbm.train(params,train_data,num_boost_round=999)
params = {"objective" : 'quantile', "alpha" : 0.05,'force_col_wise':True} #5%
lower = lgbm.train(params,train_data,num_boost_round=999)
params = {"objective" : 'quantile', "alpha" : 0.95,'force_col_wise':True} #95%
upper = lgbm.train(params,train_data,num_boost_round=999)

# Prediction
y_pred_lower = lower.predict(x_test2)
y_pred_upper = upper.predict(x_test2)
y_pred_mid = mid.predict(x_test2)

# Results
fig,ax = plt.subplots(figsize=(20,7))
ax.fill_between(x_test2.index,y_pred_lower,y_pred_upper,
alpha=0.85,color='#FF6347',label='Intervalo de confiaça')
ax.plot(x_test2.index,y_pred_mid,linestyle='--',linewidth=3,label='Mediana',color='#D30000')
ax.scatter(x_test2.index,y_test2,label='Real',color='#00708A')
plt.title('Regressão quantílica de volumetria de precipitação (mm)',fontweight='bold')
ax.legend()
fig.tight_layout()
plt.savefig('Regressão quantílica de volumetria de precipitação (mm).png', format='png')
```

Figura 6 – código do LightGBM

4. Resultados

Com relação aos resultados obtidos, pode se observar a relação calculada na Tabela 1.

Tabela 1 – Valores

Métrica *Pinball Loss*	Valores encontrados
Quantil de 5%	0,011
Quantil de 95%	0,018

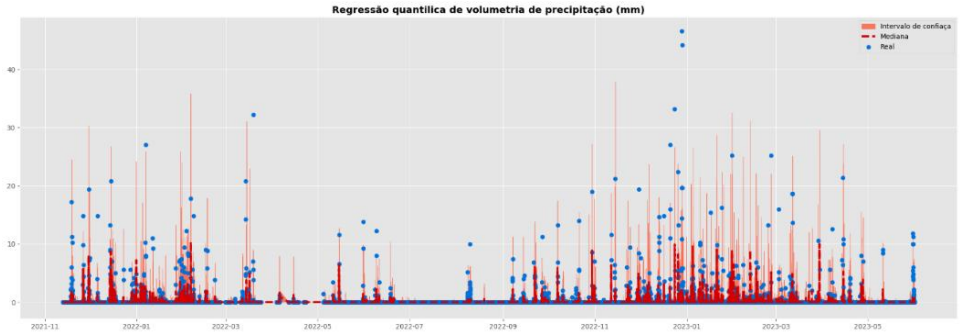

Figura 6 – Modelo de regressão de volumetria

Com referência ao modelo de regressão de volumetria, Figura 6, foram utilizados os valores de 5% e 95% como intervalos de confiança, a fim de detectar anomalias de acordo com as variáveis climáticas.

5. Discussão

Em primeiro lugar, o modelo apresentou um desempenho consistente em relação às métricas de Pinball Loos para os diferentes quantis. Para o caso, do quantil de 5%, o resultado obtido foi de 0,011, para o quantil de 95% encontrou-se um resultado igualmente satisfatório de 0,018. Logo, quanto menor o valor da métrica de *Pinball loss*, melhor a previsão se encontrará dos valores reais, de acordo com Gneiting et al. (2023).

Para finalizar, com base na métrica utilizada, pode-se concluir que o modelo de regressão quantílica *LightGBM* apresenta um desempenho razoável na previsão da volumetria de precipitação. No entanto, é importante considerar o contexto específico da aplicação e avaliar as métricas em relação às necessidades e requisitos do problema em questão.

6. Conclusão

A IA tem sido amplamente utilizada na pesquisa climática e na previsão de chuvas. Com avanços tecnológicos, os modelos de previsão climática se tornaram mais precisos, permitindo aos pesquisadores identificar padrões e prever eventos extremos com antecedência. Além disso, a análise de grandes conjuntos de dados com IA proporciona uma compreensão mais profunda dos fenômenos e uma melhor previsão das mudanças climáticas futuras.

Em relação à fundamentação teórica, foram citados trabalhos relacionados ao uso da IA para prever a quantidade de chuva. Em resumo, as lacunas encontradas, os resultados e as métricas permitem o estudo e desenvolvimento de novas técnicas.

No artigo, foi desenvolvido um processo que envolve a coleta, pré-processamento e análise dos dados de chuva. A abordagem proposta apresentou melhores resultados em relação às métricas comumente usadas para avaliar modelos. Além disso, é fácil de usar e requer poucos recursos financeiros e tecnológicos.

A regressão de árvore com LightGBM é útil para prever a quantidade de chuva, capturando padrões relevantes e melhorando a precisão da previsão. Considerando fatores como topografia, uso do solo e características climáticas regionais, esse modelo fornece informações importantes para identificar áreas de risco e definir medidas de prevenção e mitigação de desastres naturais. A regressão quantílica é uma abordagem estatística eficaz na análise de dados volumétricos, permitindo uma compreensão melhor das relações e uma tomada de decisão informada. Portanto, a utilização da regressão quantílica é uma ferramenta valiosa para aprimorar a análise e previsão de profissionais.

A regressão quantílica é vantajosa para lidar com os desafios dos dados de chuvas. Essa técnica estatística permite uma análise robusta e abrangente, considerando diferentes percentis dos dados. Além disso, ela oferece insights valiosos sobre padrões e variações das chuvas, sendo útil em áreas como agricultura, gerenciamento de recursos hídricos e previsão de riscos climáticos. Portanto, a regressão quantílica possibilita explorar de forma completa as características e benefícios dos dados de chuvas, contribuindo para uma análise precisa.

O modelo apresentou um Pinball loss de 0,011 para o quantil de 5% e 0,018 para o quantil de 95%, indicando precisão nas previsões em comparação com os valores reais. Em resumo, o modelo de regressão LightGBM tem um desempenho razoável na previsão da volumetria de precipitação, porém é importante avaliar as métricas conforme as necessidades do problema.

Quanto às pesquisas futuras, é previsto o estudo de algoritmos de aprendizado de máquina para prever o volume de chuva de forma mais precisa. Isso visa melhorar os resultados em diferentes situações, considerando especificidades de outros cenários e cidades. Assim, será possível aprimorar um modelo de predição que possa ser aplicado em uma escala maior, abrangendo mais dados e informações.

Referências

Dananjali, T., Wijesinghe, S., & Ekanayake, J. (2020). Forecasting weekly rainfall using data mining technologies. *2020 From Innovation to Impact, FITI 2020*. https://doi.org/10.1109/FITI52050.2020.9424877

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., Dimitriadis, T., Hagenmeyer, V., Jordan, A. I., Lerch, S., Phipps, K., & Schienle, M. (2023). Model Diagnostics and Forecast Evaluation for Quantiles. *Annual Review of Statistics and Its Application*, *10*(1), 597–621. https://doi.org/10.1146/annurev-statistics-032921-020240

IBGE Cidades. (2022). *Portal IBGE - Instituto Brasileiro de Geoagrafia e Estatística*. IBGE. https://cidades.ibge.gov.br/brasil/sp/sao-carlos/panorama

INMET. (2023, August 26). *Portal INMET - Instituto Nacional de Meteorologia*. Portal INMET. https://portal.inmet.gov.br/dadoshistoricos

Juneja, A., & Das, N. N. (2019, February). Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 559–563. https://doi.org/10.1109/COMITCon.2019.8862267

Karatayev, M., Clarke, M., Salnikov, V., Bekseitova, R., & Nizamova, M. (2022). Monitoring climate change, drought conditions and wheat production in Eurasia: the case study of Kazakhstan. *Heliyon*, *8*(1), e08660. https://doi.org/10.1016/j.heliyon.2021.e08660

Li, K., Ma, J., Gao, J., Xu, C., Li, W., Mao, Y., & Jiang, S. (2023). Resilience Assessment of Urban Distribution Network Under Heavy Rain: A Knowledge-Informed Data-Driven Approach (April 2023). *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3288341

Prefeitura de São Carlos. (2020). *DADOS DA CIDADE - Prefeitura de São Carlos*. DADOS DA CIDADES. http://www.saocarlos.sp.gov.br/index.php/conheca-sao-carlos/115442-dados-da-cidade-geografico-e-demografico.html

python Software Foundation. (2023). *python SOFTWARE FOUNDATION*. PYTHON.ORG. https://www.python.org/psf-landing/

Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. *Automation in Construction*, *129*, 103827. https://doi.org/https://doi.org/10.1016/j.autcon.2021.103827

Thapelo, T. S., Namoshe, M., Matsebe, O., Motshegwa, T., & Bopape, M.-J. M. (2021). SASSCAL WebSAPI: A Web Scraping Application Programming Interface to Support Access to SASSCAL's Weather Data. *Data Science Journal*, *20*. https://doi.org/10.5334/dsj-2021-024

Usman, C. D., Widodo, A. P., Adi, K., & Gernowo, R. (2023). Rainfall prediction model in Semarang City using machine learning. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(2), 1224. https://doi.org/10.11591/ijeecs.v30.i2.pp1224-1231

Vantas, K., Sidiropoulos, E., & Loukas, A. (2020). Estimating current and future rainfall erosivity in greece using regional climate models and spatial quantile regression forests. *Water (Switzerland)*, *12*(3), 1–20. https://doi.org/10.3390/w12030687

Yang, D., Yang, G., & Liu, B. (2023). Combining quantiles of calibrated solar forecasts from ensemble numerical weather prediction. *Renewable Energy*, *215*. https://doi.org/10.1016/j.renene.2023.118993

Yovan Felix, A., Vinay, G. S. S., & Akhik, G. (2019). K-Means Cluster Using Rainfall and Storm Prediction in Machine Learning Technique. *Journal of Computational and Theoretical Nanoscience*, *16*(8), 3265–3269. https://doi.org/10.1166/jctn.2019.8174