

PREDICTING SEVERE RAINFALL IN SÃO PAULO WITH MACHINE LEARNING MODELS

Alex Fernandes de Souza

Federal University of Itajubá

BPS Avenue, 1303 - Pinheirinho, Itajubá - Minas Gerais

alex_fernandes1989@live.com

Paulo Henrique da Silva Campos

Federal University of Itajubá

BPS Avenue, 1303 - Pinheirinho, Itajubá - Minas Gerais

paulohcampos@unifei.edu.br

Pedro Paulo Balestrassi

Federal University of Itajubá

BPS Avenue, 1303 - Pinheirinho, Itajubá - Minas Gerais

pedro@unifei.edu.br

Filipe Alves Neto Verri

Aeronautics Institute of Technology

Marshal Eduardo Gomes Square, 50 - Vila das Acacias, São José dos Campos - São Paulo

filipe.verri@gp.ita.br

ABSTRACT

Climate change has made the prediction of intense rainfall increasingly challenging, impacting urban management. The objective of this study was to evaluate the application of machine learning methods in forecasting intense rainfall in São Paulo. The methodology included the application of PCA combined with K-means for data segmentation and the use of models such as Logistic Regression, Random Forest, SVM, and Gradient Boosting. The results showed that Random Forest stood out with high values of precision, recall, F1-Score, and accuracy, indicating its robustness and superior generalization capability. In particular, Random Forest achieved a recall and F1-Score of 0.87 and an AUC of 0.81, demonstrating its effectiveness in discriminating between rainfall classes. The integration of machine learning techniques proved effective in forecasting intense rainfall, which can contribute to risk mitigation strategies.

KEYWORDS. Machine Learning. Rainfall Forecasting. Climate Change.

Paper topics (Introduction. Methodology. Results and Discussion. Conclusion. Literature Review)

1. Introduction

Climate change has become an increasing concern worldwide, directly impacting the environment and society. Rising global temperatures, alterations in precipitation patterns, and the increased frequency of extreme weather events are some of the most evident consequences of this phenomenon. Heavy rains and severe storms, in particular, have been occurring with greater intensity and frequency, causing floods, landslides, and other natural disasters. These extreme events pose a significant challenge to communities and infrastructure, requiring an effective response in both prevention and mitigation of their effects.

Meteorological forecasting plays a crucial role in preparing for and responding to severe weather events. With accurate and timely forecasts, it is possible to take preventive measures, such as evacuating risk areas and reinforcing vulnerable structures, thereby minimizing material damage and saving human lives. However, the complexity and variability of climate systems make forecasting a significant challenge.

In recent years, machine learning models have emerged as a powerful tool to enhance the accuracy of meteorological forecasts. Unlike traditional methods, which rely primarily on physical equations and historical data, machine learning models can process large volumes of data in real-time and identify complex patterns that may not be evident to conventional methods Slater et al. [2023].

According to Rasp and Lerch [2018], this hybrid approach utilizes neural networks to refine predictions, resulting in greater accuracy. Additionally, Weyn et al. [2021] highlight the potential of deep neural networks to predict complex climate patterns over various timescales. Furthermore, Tang et al. [2022] discuss that the challenges and opportunities of Big Data in weather forecasting are substantial, offering new insights and capabilities. This allows for a more detailed and precise analysis of weather conditions, contributing to more reliable forecasts.

The application of machine learning techniques extends to a wide range of climatic phenomena. Advanced algorithms, such as neural networks and random forests, have been used to predict the trajectory of hurricanes, the intensity of heatwaves, and the occurrence of droughts. These models are trained with historical data and continuously adjusted as new data becomes available, increasing their accuracy and reliability over time.

The prediction of intense rainfall, which can cause floods and inundations, is of particular importance due to the devastating impacts these events can have on urban and rural areas. Flash floods can destroy properties, disrupt essential services, and result in significant human losses. Machine learning models can provide more accurate and timely forecasts, enabling authorities and populations to take effective preventive measures. The ability to anticipate these events with greater precision can help mitigate their effects, saving lives and reducing economic and social damage.

The objective of this study was to evaluate the application of machine learning methods for classifying intense rainfall in São Paulo based on various climatic variables, such as wind and temperature. Additionally, the study aimed to emphasize the value of these methods for meteorologists and decision-makers, contributing to the development of more effective risk mitigation strategies.

2. Literature Review

Machine learning is transforming weather forecasting by combining traditional numerical models with modern ML techniques. This integration allows for the enhancement of predictive capabilities, leveraging the strengths of both approaches. According to Rasp and Lerch [2018], this hybrid approach uses neural networks to refine the predictions of numerical models, resulting in greater accuracy, especially for extreme weather events such as hurricanes and tornadoes, which are notoriously difficult to predict accurately.

Deep neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are being applied to generate weather forecasts with high spatial and temporal resolution. These sophisticated models can capture intricate patterns in atmospheric data that simpler models might miss. As highlighted by Weyn et al. [2021], these networks demonstrate the potential to predict complex climate patterns on timescales ranging from weeks to months, providing valuable insights for both short-term weather events and long-term climate trends.

Rainfall forecasting is crucial for various activities, from agriculture to water resource management, as well as disaster preparedness and mitigation. ML models have proven effective in short-term rainfall forecasting (nowcasting), using radar data and real-time meteorological information. As explored by Ravuri et al. [2021], these models can quickly process vast amounts of data to deliver timely and accurate predictions, which are essential for preventing flood damage and optimizing water usage in drought-prone areas.

The interpretability of ML models is essential for meteorologists and decision-makers to trust their predictions. Unlike traditional models, which often act as a “black box”, ML models can be opaque in their decision-making processes. According to Slater et al. [2023], Explainable Artificial Intelligence (XAI) techniques help understand the factors influencing predictions, making them more transparent and reliable. This transparency is crucial for gaining the confidence of stakeholders who rely on these forecasts for critical decision-making.

The availability of large volumes of meteorological data from satellites, radars, weather stations, and other sources drives the development of increasingly sophisticated ML models. These data sources provide the raw information needed to train and refine machine learning algorithms. Tang et al. [2022] discuss that the challenges and opportunities of Big Data in weather forecasting are substantial, offering new insights and capabilities. The ability to process and analyze this data efficiently can lead to significant improvements in forecast accuracy and reliability.

In addition to improving short-term forecasts, AI models are being applied to long-term climate predictions. These models help in understanding the broader impacts of climate change over decades. According to Bochenek and Ustrnul [2022], these models can analyze trends and patterns to provide insights into future climate scenarios, aiding in planning and mitigation efforts. This long-term perspective is crucial for developing strategies to cope with potential changes in climate patterns.

Machine learning models play a crucial role in enhancing resilience to adverse weather conditions. They enable communities to better prepare for and respond to severe weather events. Jaseena and Koor [2022] highlight that by accurately predicting severe weather events, these models enable better preparation and response, reducing the impact on communities and infrastructure. This capability is essential for minimizing the economic and human costs associated with extreme weather phenomena.

The integration of AI into operational weather forecasting systems is the final step to revolutionizing the field. This involves continuous collaboration between human meteorologists and AI methods to deliver accurate, timely, and actionable weather forecasts.

3. Methodology

Initially, meteorological data from 2021 to April 2024 were collected from the city of São Paulo¹. The variables collected include hourly total precipitation, atmospheric pressure at the station level, maximum and minimum atmospheric pressure in the previous hour, global radiation, air temperature, dew point temperature, maximum and minimum temperature in the previous hour, minimum dew point temperature in the previous hour, relative humidity, wind direction, maximum wind gust, and hourly wind speed.

¹<https://portal.inmet.gov.br/dadoshistoricos>

Next, a categorical variable for classifying rainfall was created based on precipitation intensity. This classification allows for categorizing rainfall into different levels of severity, facilitating the interpretation and use of the data in the machine learning model, as shown in Table 1.

Intensity (mm/h)	Classification
Less than 2.5	Light rain
Between 2.5 and 10 (inclusive)	Moderate rain
Between 10 and 50 (inclusive)	Heavy rain
Greater than 50	Violent rain

Table 1: Rainfall intensity classification.

The collected data underwent a cleaning and preparation process. This included handling missing values and normalization to ensure they are on the same scale, using the z-score standardization technique.

To avoid model bias, it was necessary to balance the classes. We used the SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples of the minority classes, ensuring the training dataset is balanced. This ensures that the model is not biased towards the majority classes, improving its ability to predict rare events.

The dataset was then divided into two parts: training (70%) and testing (30%). Four models were used: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Classifier. Each model was trained using the balanced training set and evaluated on the test set. The training process involved using 10-fold cross-validation to assess the models' performance on the training data. This cross-validation process helps in understanding the model's performance stability and robustness. After training, the models were evaluated on the test set based on performance metrics, including accuracy, precision, recall, F1-Score, and area under the ROC curve (AUC-ROC). The Python code can be seen in the online GitHub repository.²

4. Results and Discussion

The first point analyzed was the identification of extreme rainfall events in São Paulo between January 2021 and January 2024. Figure 1 illustrates the daily precipitation, segmented into intensity categories: light rain (< 2.5 mm), moderate rain (2.5 - 10 mm), heavy rain (10 - 50 mm), and violent rain (> 50 mm).

It is observed that intense precipitation events occur mainly at the beginning and end of each year. Significant peaks of violent rain, with precipitation above 50 mm, are recorded in February and March 2021, January, March, and December 2022, and January, February, and April 2023. These events highlight the need for effective strategies to mitigate the impacts of extreme rainfall, such as floods and landslides, and emphasize the importance of early warning systems and resilient urban planning.

Next, we performed a feature selection considering only the use of numerical variables. The selected variables and their respective units of measurement are presented in Table 2

The type of rain was classified using the parameters from Table 1, resulting in Table 3, which presents the distribution of rainfall classifications.

Analyzing Table 3, between 2021 and 2024, it is observed that most days recorded light rain, totaling 28,073 occurrences. Days without data sum up to 662, while moderate rain was observed on 352 days and heavy rain on 97 days.

²https://github.com/alexsouza1989/RainfallPrediction-SaoPaulo/blob/main/sbpo_chuva_sp_git.ipynb

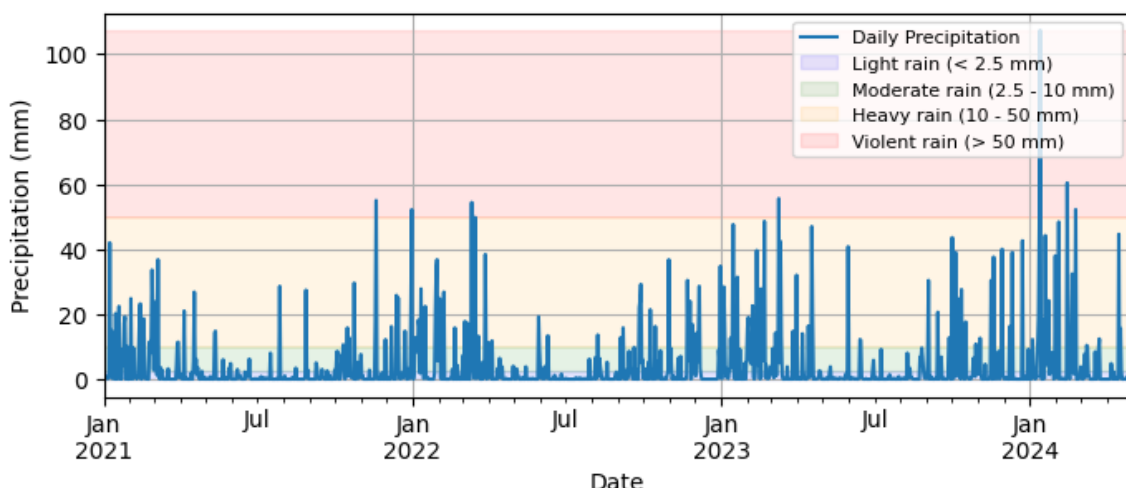


Figure 1: Daily Precipitation with Rainfall Intensity Classification.

Variable	Unit of Measurement
TOTAL PRECIPITATION, HOURLY	mm
ATMOSPHERIC PRESSURE AT STATION LEVEL, HOURLY	mB
ATMOSPHERIC PRESSURE MAX. IN THE PREVIOUS HOUR (AUT)	mB
ATMOSPHERIC PRESSURE MIN. IN THE PREVIOUS HOUR (AUT)	mB
GLOBAL RADIATION	Kj/m ²
AIR TEMPERATURE - DRY BULB, HOURLY	°C
DEW POINT TEMPERATURE	°C
MAXIMUM TEMPERATURE IN THE PREVIOUS HOUR (AUT)	°C
MINIMUM TEMPERATURE IN THE PREVIOUS HOUR (AUT)	°C
DEW POINT MIN. TEMPERATURE IN THE PREVIOUS HOUR (AUT)	°C
RELATIVE AIR HUMIDITY, HOURLY	%
WIND, MAX GUST	m/s
WIND, HOURLY SPEED	m/s

Table 2: Variables and Their Units of Measurement

This distribution indicates that light rain is the most common in São Paulo, which is consistent with the region's climate. However, the presence of moderate and heavy rain events, although less frequent, is significant and warrants attention due to their potential impact in terms of floods and landslides. Therefore, this data allows for the development of risk management strategies and the implementation of warning systems that can mitigate the adverse effects of these extreme weather events.

To explore the data structure and identify natural groupings, we performed a Principal Component Analysis (PCA) combined with the K-means algorithm. PCA was used to reduce the dimensionality of the data, facilitating visualization in a two-dimensional space. Figure 2 presents the result of this analysis, where the points were grouped into four distinct clusters.

Rain Classification	Count
Light rain	28073
Moderate rain	352
Heavy rain	97

Table 3: Distribution of Rainfall Classification.

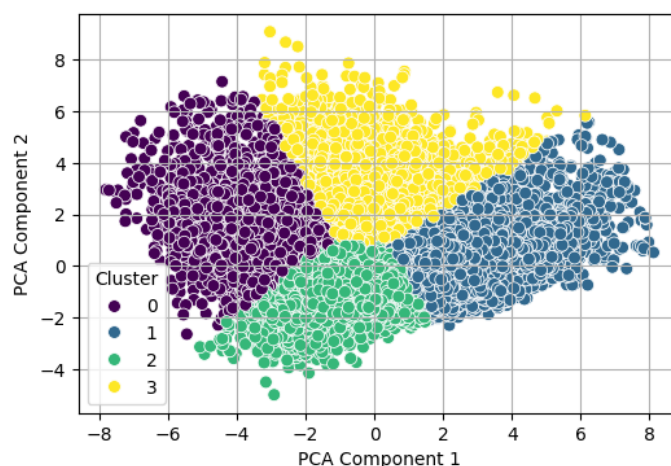


Figure 2: KMeans Clusters on PCA-Reduced Data.

PCA is a dimensionality reduction technique that allows visualizing data variability in fewer dimensions, retaining most of the original data's information [Hasan and Abdulazeez, 2021]. In this study, PCA reduced the data to two principal components. The K-means algorithm was then applied to these reduced data, defining four clusters with the parameter `n_clusters=4` and `random_state=42` to ensure reproducibility. The resulting clusters were added to the original DataFrame.

The PCA analysis combined with K-means identified four distinct clusters in the data, indicating different patterns or categories. The clear separation between clusters in the two-dimensional space suggests a well-defined internal structure in the original data.

The clear separation of clusters reinforces the effectiveness of K-means in finding significant groupings in the data reduced by PCA. This analysis highlights the importance of combining dimensionality reduction techniques with clustering algorithms to gain deeper and more actionable insights from the data.

Next, we applied different machine learning models to the balanced and PCA-reduced data, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Classifier. Each model was trained and evaluated using 10-fold cross-validation to ensure the robustness of the results.

The models were tuned and evaluated as necessary. Logistic Regression was evaluated using a grid search to optimize the regularization hyperparameter. Random Forest was tuned by varying the number of estimators and the maximum tree depth. For SVM, different kernel functions and C values were tested to find the best combination. The Gradient Boosting Classifier was tuned by adjusting the number of estimators and the learning rate.

The models were tuned and evaluated as necessary. Logistic Regression was evaluated using a grid search to optimize the regularization hyperparameter. Random Forest was tuned by varying the number of estimators and the maximum tree depth. For SVM, different kernel functions

and C values were tested to find the best combination. The Gradient Boosting Classifier was tuned by adjusting the number of estimators and the learning rate. Thus, the hyperparameters automatically selected by the Grid Search can be seen in Table 4, providing a detailed view of the optimal configurations used for each model.

Model	Best Hyperparameters
Logistic Regression	{ 'C': 0.1 }
Random Forest Classifier	{ 'max_depth': None, 'n_estimators': 200 }
Support Vector Machine (SVM)	{ 'C': 100, 'kernel': 'rbf' }
Gradient Boosting Classifier	{ 'learning_rate': 0.2, 'n_estimators': 300 }

Table 4: Best hyperparameters and test accuracy for each model.

Once the hyperparameters were determined, we plotted the learning curves for each of the models. These curves provide a clear view of the model's performance in terms of its accuracy on the training set and in cross-validation, as the amount of training data increases. The learning curves for each model can be seen in Figure 3, showing how each model fits the data and its ability to generalize.

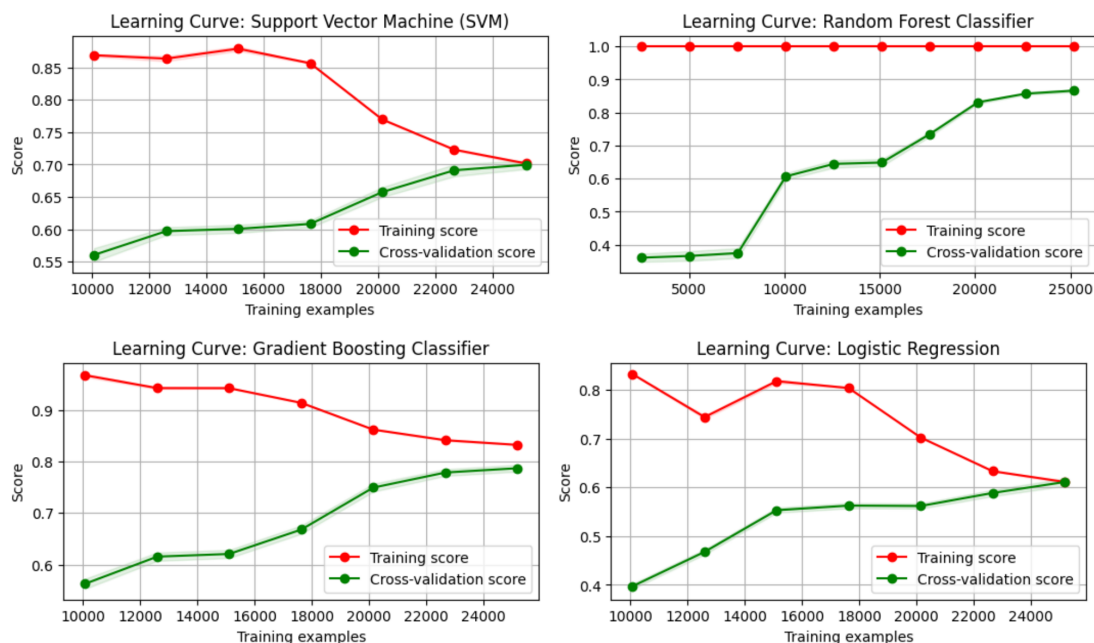


Figure 3: Learning Curves of Machine Learning Models for Classifying Intense Rainfall in São Paulo

For the SVM, we observe high performance on the training set but a significant drop in cross-validation as the number of training examples increases, indicating overfitting. The Random Forest shows consistently high performance in training and continuous improvement in cross-validation, suggesting good generalization as more data is added. The Gradient Boosting Classifier also demonstrates high performance in training but with a smaller gap in cross-validation, indicating a good fit with slight overfitting. The Logistic Regression, on the other hand, shows convergence

of the training and validation curves, suggesting a well-balanced model with good generalization capability.

Subsequently, the ROC curves provide an additional evaluation of the models' performance, complementing the learning curves. Figure 4 presents the ROC curves for each model, highlighting their ability to discriminate between classes.

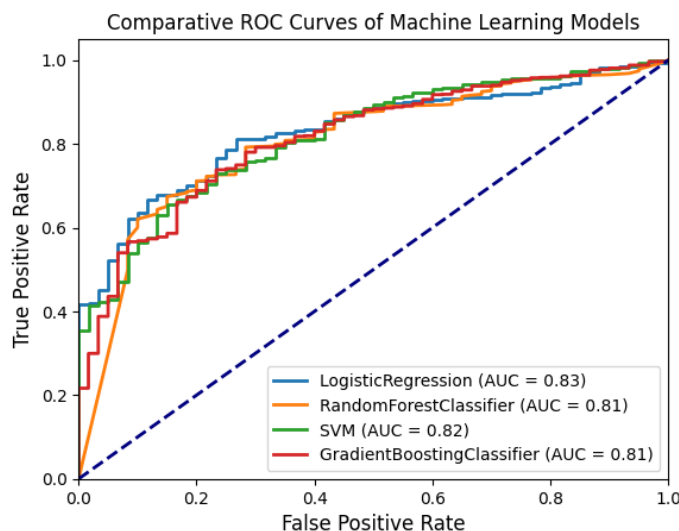


Figure 4: Comparative ROC Curves for Machine Learning Models in Rainfall Prediction.

The learning curves and ROC curves provide a comprehensive view of the performance of the machine learning models. The Support Vector Machine (SVM) showed high performance on the training set but a significant drop in cross-validation as the number of training examples increased, indicating overfitting. However, the AUC of 0.82 suggests that the SVM can still be useful in specific contexts. The Random Forest Classifier exhibited consistently high performance in training and continuous improvement in cross-validation, suggesting good generalization with an AUC of 0.81. The Gradient Boosting Classifier also demonstrated high performance in training, with a smaller gap in cross-validation, indicating a good fit with slight overfitting and an AUC of 0.81. The Logistic Regression showed convergence of the training and validation curves, suggesting a well-balanced model with good generalization capacity, reflected in its AUC of 0.83, the highest among the models.

The combined analysis of the learning curves and AUC values in the ROC curves reveals that Logistic Regression stands out as the most balanced model, with excellent generalization and discrimination capabilities. The Random Forest and Gradient Boosting also proved robust, with good generalization and discrimination capabilities, despite slight overfitting in Gradient Boosting. The SVM, despite its overfitting, maintained a competitive AUC, indicating potential utility with additional adjustments.

Complementing the analyses, four additional metrics were calculated: precision, recall, F1-Score, and accuracy, as shown in Table 5.

Model	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.614 ± 0.007	0.612 ± 0.007	0.613 ± 0.007	0.612 ± 0.007
Random Forest Classifier	0.870 ± 0.005	0.867 ± 0.005	0.866 ± 0.005	0.867 ± 0.005
SVM	0.734 ± 0.009	0.699 ± 0.010	0.705 ± 0.010	0.699 ± 0.010
Gradient Boosting Classifier	0.799 ± 0.007	0.787 ± 0.007	0.788 ± 0.007	0.787 ± 0.007

Table 5: Performance metrics (Mean ± Std) for each model.

Based on these metrics, we observe that the model with the best performance, associated with AUC, was the Random Forest Classifier, with high average precision, recall, F1-Score, and accuracy, all around 0.87. Conversely, the model with the lowest performance was Logistic Regression, with average precision, recall, F1-Score, and accuracy around 0.61. These results indicate that Random Forest not only performed well in terms of AUC but also excelled in all performance metrics, standing out as the most robust model for classifying intense rainfall in São Paulo.

5. Conclusion

The analyses conducted in this study demonstrate that machine learning models are effective tools for predicting intense rainfall in São Paulo. The learning curves and performance metrics highlight the robustness of the models, especially the Random Forest Classifier, which stood out with high precision, recall, F1-Score, and accuracy values, as well as a high AUC. This model not only exhibited excellent performance in terms of class discrimination but also showed superior generalization ability compared to the other models analyzed.

With the growing concern about climate change and its impacts, it is becoming increasingly crucial to develop methods capable of identifying and predicting extreme climate variations. This study confirms that the use of advanced machine learning techniques can significantly contribute to mitigating the risks associated with severe weather events. Logistic Regression, although having the lowest relative performance, still demonstrated an acceptable predictive capability, reinforcing that all models have potential for practical applications in classifying intense rainfall.

These results not only highlight the effectiveness of the models in the task but also underline the importance of innovative technological approaches to address the challenges posed by climate change and protect urban populations. The combination of learning and ROC curve analyses provided a comprehensive understanding of the models' effectiveness, highlighting the Random Forest as the best choice for this specific application.

6. Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

- Bochenek, B. and Ustrnul, Z. (2022). Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere*, 13:180. ISSN 2073-4433.
- Hasan, B. M. S. and Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, 02. ISSN 2716621X.
- Jaseena, K. and Koor, B. C. (2022). Deterministic weather forecasting models based on intelligent predictors: A survey. *Journal of King Saud University - Computer and Information Sciences*, 34: 3393–3412. ISSN 13191578.

- Rasp, S. and Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146:3885–3900. ISSN 0027-0644.
- Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672–677. ISSN 0028-0836.
- Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R. L., Wood, A., and Zappa, M. (2023). Hybrid forecasting: blending climate predictions with ai models. *Hydrology and Earth System Sciences*, 27:1865–1889. ISSN 1607-7938.
- Tang, L., Li, J., Du, H., Li, L., Wu, J., and Wang, S. (2022). Big data in forecasting research: A literature review. *Big Data Research*, 27:100289. ISSN 22145796.
- Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, 13. ISSN 1942-2466.