ISSN 2959-6157

Weather Prediction with Feature Selection and Random Forest

Zhihao Dai[1, *]

[1] Zhixin high school, Guangzhou, China
*Corresponding author: murderlone@gmail.com

Abstract:

Accurate weather prediction is crucial for various sectors, including agriculture, disaster management, and aviation. Traditional weather prediction relies heavily on numerical weather prediction (NWP) models that simulate atmospheric processes using mathematical equations. While these methods have been the backbone of forecasting for decades, they often require significant computational resources and may need help capturing localized weather events. In contrast, machine learning models can quickly analyze large datasets to identify patterns and relationships that traditional methods might miss. This paper employs a Random Forest model to investigate the impact of feature selection on weather prediction accuracy. This paper identifies those critical to model performance by systematically excluding individual features, offering valuable insights into how ML can enhance traditional forecasting techniques. The results underscore the importance of features such as "Present_Tmin" and "LDAPS_LH", which significantly influence model performance as measured by Root Mean Square Error (RMSE). This study contributes to a better understanding of feature selection in developing accurate weather prediction models. It lays the groundwork for future research to improve prediction accuracy through advanced techniques and expanded datasets.

Keywords: Weather Prediction; Feature Selection; Machine Learning; Random Forest.

1. Introduction

Weather prediction is a complex task involving numerous environmental and atmospheric factors. Accurate forecasting is vital for numerous sectors, including agriculture, disaster management, and aviation, enabling better decision-making and risk mitigation. Weather prediction has always been a complex task due to the environmental and atmospheric factors involved. Historically, this task has been addressed using Numerical Weather Prediction (NWP) models, which simulate atmospheric processes through mathematical equations. These models, based on the principles of physics and dynamics, have been the foundation of weather forecasting for decades. However, NWP models have their limitations. They depend heavily on the quality of initial conditions and require substantial computational power to process vast data. Their resolution needs to be revised to accurately capture micro-scale weather phenomena, such as localized thunderstorms or fog, leading to less accurate short-term forecasts. Despite significant improvements over the years, these models still face challenges, mainly when predicting weather events that require real-time data processing and high-resolution outputs [1-3].

The advent of machine learning has introduced new possibilities in weather prediction. Unlike NWP models, machine learning algorithms excel at identifying complex, non-linear relationships within large datasets. This capability allows them to process and analyze vast amounts of historical weather data, uncovering patterns that traditional models might miss. Machine learning techniques, such as deep learning and ensemble methods, have already shown promise in various domains, including finance, healthcare, and marketing. These techniques can improve weather forecast accuracy, particularly in short-term predictions and localized events. Recent studies have demonstrated that machine learning models can complement traditional methods by providing more accurate predictions when the two approaches are combined.

Machine learning models, such as Random Forests, Support Vector Machines (SVM), and Neural Networks, have been employed in weather prediction with varying degrees of success. However, the performance of these models heavily depends on selecting relevant features. The right combination of features can significantly enhance model accuracy, while including irrelevant or redundant features can lead to overfitting, resulting in poor generalization of new data. Therefore, understanding the importance of individual features in a weather prediction model is crucial for developing efficient and accurate forecasting systems [4, 5].

Feature selection is a critical step in the machine learning

pipeline, especially in the context of weather prediction. In many cases, the raw data contains many variables, not all of which contribute equally to the model's predictive power. Some features may be highly correlated with others, leading to redundancy, while others might introduce noise, thereby degrading the model's performance. By carefully selecting the most relevant features, the researcher can reduce the complexity of the model, improve its interpretability, and enhance its predictive accuracy. Previous research has explored various methods for feature selection, including statistical techniques, such as correlation analysis, and model-based approaches, such as feature importance scores from ensemble models. In weather prediction, selecting features like temperature, humidity, and solar radiation is significant, as these variables are directly related to weather dynamics.

This study adopts a straightforward approach to feature selection by systematically excluding each feature and measuring the impact on model performance. This method provides a transparent and interpretable measure of feature importance, allowing for a direct comparison of the contributions of individual features. By identifying the key features that drive model accuracy, we aim to improve the efficiency of weather prediction models, making them more robust and reliable.

2. Methodology

2.1 Machine Learning Models

This paper uses the Random Forest (RF) model, a robust ensemble learning method known for its effectiveness in handling non-linear relationships and reducing overfitting. The Random Forest algorithm operates by constructing many decision trees during training time and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. This approach is particularly advantageous for weather prediction tasks due to its ability to manage high-dimensional data and its resilience to overfitting, a common issue in models with many features. The RF model's use of bagging, where each tree is trained on a random subset of the data, helps capture diverse aspects of the data, leading to more robust predictions.

RF offers a good balance between interpretability and predictive power compared to other models like Gradient Boosting Machines (GBM) or Support Vector Machines (SVM). While GBMs often provide better predictive accuracy through boosting, they are more prone to overfitting, especially in noisy data environments like weather prediction. On the other hand, SVMs, though powerful, typically require careful tuning of hyperparameters and are less intuitive to interpret. The choice of Random Forest in this study is guided by its proven effectiveness in previous research on weather prediction and its ability to handle the complex, non-linear relationships inherent in weather data [6].

2.2 Data Preprocessing and Feature Selection

Before training the model, the dataset undergoes several preprocessing steps to ensure the data is clean and suitable for analysis. The dataset used in this study is derived from a Kaggle competition and includes various features related to temperature, humidity, solar radiation, and topographical data. The first step in data preprocessing involves handling missing data. Missing values can lead to biases in model training if not appropriately addressed. In this study, we employ imputation techniques where missing values are replaced based on the mean or median of the available data. This method is particularly effective when the missing data is random, and the proportion of missing values is relatively small.

Next, the features are normalized to ensure that they are on a comparable scale. This step is crucial because many machine learning models, including Random Forests, can be sensitive to the scale of the input features. Normalization is achieved by scaling the features to a range between 0 and 1, which helps accelerate the convergence of the learning algorithm and improve model performance.

Feature selection is an integral part of the preprocessing process. This study utilises domain knowledge and statistical methods to identify the most relevant features for weather prediction. Domain knowledge helps in the initial identification of critical features such as temperature, humidity, and solar radiation, which are known to have a direct impact on weather patterns. Statistical methods, such as correlation analysis and mutual information, are then applied to refine the selection by identifying features with solid relationships with the target variable and low redundancy with other features.

2.3 Model Evaluation

The performance of the Random Forest model is evaluated using the Root Mean Square Error (RMSE), a widely used metric in regression tasks. RMSE measures the average magnitude of the error between the predicted and actual values, with lower RMSE values indicating better model performance. RMSE is chosen over metrics like Mean Absolute Error (MAE) or R-squared due to its sensitivity to significant errors. This is particularly important in weather prediction, where extreme values can have significant implications.

In addition to RMSE, cross-validation is performed to assess the robustness of the model. Cross-validation involves dividing the dataset into multiple folds and training

the model on different subsets of the data, ensuring that the model's performance is consistent across various data splits. This approach helps mitigate the risk of overfitting and provides a more reliable estimate of the model's generalization ability.

3. Experimental Results and Analysis

3.1 Dataset

The dataset utilized in this study originates from a Kaggle competition and contains a rich set of features related to weather conditions. The data spans a specific period, providing a comprehensive view of weather patterns. Key features include temperature-related variables such as Present_Tmin (minimum temperature at present), LDAPS_Tmax_lapse (temperature lapse rate for the maximum temperature), and LDAPS_Tmin_lapse (temperature lapse rate for the minimum temperature). Humidity-related features include LDAPS_RHmin (minimum relative humidity) and LDAPS_RHmax (maximum relative humidity), while solar radiation and topographical features like latitude, longitude, and slope are also included. Each feature in the dataset plays a critical role in capturing the complex dynamics of weather patterns. Temperature and humidity are directly linked to the formation of weather phenomena, while solar radiation influences temperature fluctuations and evaporation rates. Topographical features, such as elevation and slope, can affect local climate conditions by influencing wind patterns and precipitation.

3.2 Feature Exclusion Results

The feature exclusion experiments reveal significant insights into the importance of individual features in weather prediction. When critical features such as "Present_Tmin" and "LDAPS_LH" are excluded, there is a notable increase in RMSE, indicating their substantial impact on model performance and specifically, excluding "Present_Tmin" results in an RMSE increase from 1.533 to 1.564, highlighting its importance in temperature-related predictions. Similarly, excluding "LDAPS_LH" leads to an RMSE rise from 1.533 to 1.557, underscoring the significance of humidity-related features in accurate weather prediction.

The increase in RMSE when excluding these features can be attributed to their strong correlation with the target variable. "Present_Tmin" is closely linked to daily temperature variations, critical in predicting weather conditions such as frost or heatwaves. "LDAPS_LH" (Liquid et al.) is essential for understanding atmospheric moisture availability, directly influencing precipitation patterns and cloud formation.

3.3 Results and Analysis

The results of the feature exclusion experiments underscore the importance of accurate feature selection in machine learning models for weather prediction. The significant RMSE increases observed when key features are excluded demonstrate that these features are essential for capturing the underlying weather dynamics. This finding aligns with the physical understanding of weather systems, where temperature and humidity dominate in shaping weather events.

In addition to confirming the importance of specific features, the experiments suggest that there is potential for further improvement in model performance through advanced feature engineering. By creating interaction terms or transforming existing features, capturing additional aspects of the data not fully represented by the current feature set may be possible. Furthermore, combining multiple machine learning models, such as a Random Forest model with a Gradient Boosting Machine, could potentially enhance predictive accuracy by leveraging the strengths of different algorithms.

Table 1. Model performance without different features

Feature Excluded	RMSE
Baseline (All Features)	1.533
w/o Present_Tmin	1.564
w/o LDAPS_RHmin	1.527
w/o LDAPS_RHmax	1.553
w/o LDAPS_Tmax_lapse	1.545
w/o LDAPS_Tmin_lapse	1.14
w/o LDAPS_WS	1.518
w/o LDAPS_LH	1.557
w/o LDAPS_CC1	1.528

w/o LDAPS_CC2	1.550
w/o LDAPS_CC3	1.544
w/o LDAPS_CC4	1.541
w/o LDAPS_PPT1	1.535
w/o LDAPS_PPT2	1.527
w/o LDAPS_PPT3	1.530
w/o LDAPS_PPT4	1.528
w/o lat	1.529
w/o lon	1.530
w/o DEM	1.536
w/o Slope	1.524
w/o Solar radiation	1.539

The results of the feature exclusion experiments are shown in Table 1. Excluding the "Present_Tmin" feature results in the highest RMSE increase (from 1.533 to 1.564), indicating its critical importance for accurate temperature prediction. Excluding the "LDAPS_LH" feature also leads to a notable increase in RMSE (from 1.533 to 1.557), highlighting its significant contribution to the model's performance.The impact of excluding other features varies, with some features showing minimal effect on RMSE, while others exhibit moderate changes. This suggests that while certain features are crucial, the contribution of other features depends on the specific context and the relationship between features.

4. Discussion

This paper highlights the importance of feature selection in weather prediction models. Identifying and incorporating critical features like "Present_Tmin" and "LDAPS_LH" can significantly improve model accuracy. This finding aligns with our understanding of the importance of temperature and humidity in weather prediction.

The systematic approach of excluding each feature provides a precise measure of feature importance. It allows us to identify essential features for accurate predictions and those that can be excluded without significantly affecting performance. This information can guide the development of more efficient models with fewer input features, reducing computational complexity and improving interpretability.

Our study has some limitations. The dataset used in this study is limited in scope and may only encompass some relevant weather variables. Additionally, we focus on a specific machine learning model (Random Forest) and a single evaluation metric (RMSE). Exploring different models and metrics could provide a more comprehensive understanding of the importance of features.

Future research could investigate the following directions:

Exploring Feature Combinations: Investigating the impact of combining multiple features could further reveal synergistic effects and improve model performance.

Advanced Ensemble Methods: Employing advanced ensemble methods like Random Forest or Gradient Boosting could enhance prediction accuracy by leveraging the strengths of multiple models.

Feature Engineering: Exploring feature engineering techniques, such as creating interaction terms or transforming existing features, could uncover additional predictive power.

Model Interpretability: Developing methods to interpret the model's decision-making process, especially for critical features like "Present_Tmin" and "LDAPS_LH", could provide valuable insights into weather prediction mechanisms.

Dataset Expansion: Expanding the dataset to include additional features like wind direction, soil moisture, or atmospheric pressure could improve model performance and provide a more comprehensive understanding of weather patterns.

Feature Exclusion Results: The results from the feature exclusion experiments are summarized in Table 1. Excluding key features such as "Present_Tmin" and "LDAPS_LH" significantly increased the RMSE, confirming their critical role in accurate weather prediction. The exclusion of "Present_Tmin" led to an RMSE increase from 1.533 to 1.564, highlighting its importance in temperature-related predictions. Similarly, "LDAPS_LH" showed a significant impact, with RMSE rising from 1.533 to 1.557 upon its exclusion.

Interpretation of Results: The increase in RMSE when excluding specific features underscores the importance

of accurate feature selection in ML models for weather prediction. Features related to temperature and humidity were particularly crucial, aligning with the physical understanding of weather dynamics, where these variables play a dominant role.

Comparative Analysis: The Random Forest model's ability to handle non-linear relationships is evident from the relatively low RMSE values achieved with the complete feature set. However, the results also suggest that further improvements might be possible by exploring feature engineering techniques or combining multiple ML models to capture different aspects of the data.

5. Conclusion

This study investigates the impact of excluding individual features on the performance of a weather prediction model. This paper identifies critical features that significantly influence prediction accuracy by systematically excluding each feature and measuring the change in RMSE. The findings highlight the importance of feature selection in developing accurate weather prediction models and provide valuable insights for future research. Focusing on critical features and exploring advanced techniques can lead to more accurate and efficient weather prediction systems, benefiting various sectors and improving decision-making and risk mitigation.

References

[1] Sara Pereira, Paulo Canhoto, Rui Salgado. Development and assessment of artificial neural network models for direct normal solar irradiance forecasting using operational numerical weather prediction data. Energy and AI, 2024, 15: 100314.

[2] Weijing Dou, Kai Wang, Shou Shan, et al. Day-ahead Numerical Weather Prediction solar irradiance correction using a clustering method based on weather conditions. Applied Energy, 2024, 365: 123239.

[3] Methodjia M Shapkalijevski. Perspectives toward Stochastic and Learned-by-Data Turbulence in Numerical Weather Prediction. Weather and Forecasting, 2024, 39(2): 261-274.

[4] Ali Ayoub, Haruko M. Wainwright, Giovanni Sansavini. Machine learning-enabled weather forecasting for real-time radioactive transport and contamination prediction. Progress in Nuclear Energy, 2024, 173: 105255.

[5] Changqing Meng, Zhihan Hu, Yuankun Wang, et al. A forecasting method for corrected numerical weather prediction precipitation based on modal decomposition and coupling of multiple intelligent algorithms. Meteorology and Atmospheric Physics, 2024, 136(5): 1-14.

[6] Zied Ben Bouallègue, Mariana C A Clare, Linus Magnusson, et al. The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. Bulletin of the American Meteorological Society, 2024, 105(6): E864-E883.