# Temperature and Humidity Prediction Based on Machine Learning

*Yanqi* Xiong

School of Software, Jiangxi Normal University, Nanchang city, Jiangxi Province, 330000, China

**Abstract.** The growing impact of global climate change, emphasizing the critical importance of accurately predicting weather conditions, particularly temperature and humidity. These predictions are crucial for key sectors such as agriculture, energy management, and public safety. This paper employs various machine learning models, including Linear Regression(LR), Support Vector Machine(SVM), Neural Network(NN), and Random Forest(RF), to analyze their accuracy in predicting temperature and humidity. The results indicate that the NN model outperforms the others, showing excellent performance in the dataset. In addition to the outstanding performance of the neural NN, the RF and SVM also demonstrated strong performance, particularly in handling specific features within the dataset. the model's performance can be further optimized by adjusting the NN's hyperparameters or introducing more feature engineering, which could lead to even better results in future data analyses. This research highlights the significant potential of machine learning techniques in enhancing meteorological forecasting, providing valuable insights and tools for improving decision-making in industries heavily influenced by weather conditions.

## 1 Introduction

As global climate change intensifies, the accuracy of predicting future weather conditions becomes increasingly important. Weather forecasting relies on the accurate measurement and analysis of meteorological parameters such as temperature and humidity [1]. Changes in temperature and humidity directly affect atmospheric circulation patterns, which in turn determine the development and movement of weather systems. Sudden extreme weather events, such as unexpected high or low temperatures, can reduce crop yields and negatively impact farmers' incomes. High humidity or heavy rain can increase the demand for electricity, placing stress on the power grid [2]. Additionally, humidity can impact public transportation safety; slippery roads may lead to an increase in traffic accidents. Therefore, accurately predicting future temperature and humidity is of great significance to society and the economy [3].

Traditional meteorological forecasting methods primarily rely on physical models and statistical techniques. These methods typically require large amounts of historical data and complex mathematical calculations. Challenges such as sensitivity to initial conditions,

Corresponding author: sweetumz2011@email. phoenix. edu

high computational demands, and insufficient capacity to handle complex nonlinear relationships often result in less than satisfactory prediction accuracy. with the rapid development of big data and computational power, the application of machine learning techniques in weather forecasting has garnered increasing attention. Machine learning algorithms can analyze vast amounts of historical meteorological data, capturing underlying patterns and trends, thereby enhancing predictions' accuracy and efficiency. Numerous studies have demonstrated the advantages of machine learning in meteorological forecasting. For example, R. Kimura and colleagues employed existing data to predict temperature changes. Similarly, Peter Bauer, Alan Thorpe, and Gilbert Brunet utilized deep learning methods to predict weather, and their results showed significantly improved accuracy compared to traditional methods achieving favorable outcomes. These studies indicate that machine learning methods have promising applications in weather forecasting.

The study aims to explore the application of machine learning techniques in predicting temperature and humidity. The primary objective is to develop a robust prediction model by training historical meteorological data to forecast future changes in temperature and humidity. To achieve this goal, the following research tasks will be conducted: First, data collection and preprocessing: historical meteorological datasets, including parameters such as temperature, humidity, atmospheric pressure, and wind speed, will be collected from Kaggle [4]. Subsequently, the data will undergo cleaning and processing. Next, feature selection and model construction: key features will be extracted to construct various machine learning models, including LR, SVM, RF and NN. Following this step is model training and evaluation using metrics such as confusion matrix analysis, Mean Squared Error, mean MAE, and cross-validation methods for assessing model accuracy before selecting the best-performing model. Finally, the best model will be applied to test data for temperature and humidity predictions with comparative analysis against traditional methods [5].

## 2 Data and Methods

### 2.1 Dateset

The dataset used in the paper is sourced from the Kaggle platform, which specializes in the fields of data science and machine learning and typically provides accurate and authoritative data. Containing 96, 454 entries, the dataset offers several advantages. The dataset includes various meteorological parameters, not only temperature and humidity but also wind speed, atmospheric pressure, and other data. It is arranged in chronological order, making it suitable for constructing time series analyses and predictive models. Although most of the data is relatively complete, there may still be missing or outlier values that necessitate preprocessing.

Before model training, a series of operations must be performed, with data preprocessing being a crucial step prior to model construction. The main purpose of data preprocessing is to ensure the quality of the data used for model training and to minimize the impact of missing or anomalous values on the model's predictive results [6]. The data preprocessing steps in the paper include handling missing values by addressing gaps in the data through methods such as interpolation, mean imputation, or deletion; detecting and processing outliers in the data to avoid compromising the model's accuracy; and standardizing the data, as temperature and humidity have different units of measurement, making the data more suitable for processing by machine learning algorithms.

## 2.2 Models

This study employs various predictive models with the goal of selecting the optimal one for forecasting. The following is an explanation and introduction to each model.

### *2.2.1 Linear Regression Model(LR)*

Linear regression predicts the dependent variable by finding the best-fitting line that minimizes the sum of squared differences between data points and the line. Advantages for predicting temperature and humidity include the following. LR is computationally efficient, particularly when dealing with large-scale temperature and humidity datasets, enabling fast training and prediction. This makes it well-suited for real-time or near-real-time temperature and humidity forecasting. Additionally, when the relationship between temperature, humidity, and the input features is approximately linear, LR can effectively fit the data and provide stable predictions.

### *2.2.2 Support Vector Machine Model(SVM)*

SVM is a supervised learning model that seeks to maximize the margin between classes to find an optimal decision boundary for improved generalization in classification and regression tasks. SVM identifies the optimal classification hyperplane by maximizing the minimum distance between the data points and the decision boundary. In regression tasks, SVM seeks to find a hyperplane that ensures most data points fall within a specified threshold while minimizing the violation of this threshold.

Advantages for predicting temperature and humidity include the following. By using kernel functions such as the Radial Basis Function (RBF) kernel or polynomial kernel, SVM can construct complex nonlinear models in high-dimensional spaces. This gives SVM an advantage in handling the nonlinear relationships between temperature, humidity, and the input features, especially when these relationships are complex and nonlinear. Additionally, SVM is less sensitive to outliers, as it primarily relies on support vectors—the data points closest to the decision boundary—to construct the model. As a result, extreme temperature or humidity values are less likely to significantly impact the overall model performance.

### *2.2.3 Neural Network(NN) Model*

The NN model is inspired by biological neural systems and consists of multiple layers of neurons[7]. These layers include the input layer, hidden layers, and output layer. NN learn the complex mapping between input data and output targets through the connections and weight adjustments between layers [8].

The fundamental working principle of NN involves processing input data through a hierarchical structure, ultimately generating an output. This process can be broken down into three key steps. First, in forward propagation, data begins at the input layer and passes through each neuron in the hidden layers, eventually reaching the output layer. Each neuron receives input from the previous layer, multiplies it by corresponding weights, adds a bias term, and then applies an activation function to produce the output. This process is known as forward propagation. Next, the activation function is crucial for introducing nonlinearity into the NN. Common activation functions include ReLU, Sigmoid, and Tanh. The activation function transforms the linear combination of inputs into a nonlinear output, enabling the NN to handle complex nonlinear relationships. Finally, backpropagation is the core training algorithm of NN. In each training iteration, the difference (loss) between the

predicted output and the actual result is calculated, and the error is propagated back through the network using the chain rule. Gradients are computed for each weight, and optimization algorithms are used to adjust the weights, minimizing the loss function and improving the model's predictive accuracy [9].

Advantages for predicting temperature and humidity include the following. NN can automatically extract useful features from the input data without requiring manual feature design. For temperature and humidity prediction, NN can learn complex patterns and high-level features related to the prediction task, enhancing prediction accuracy. This makes NN adaptable to various tasks and data types, as they can automatically extract features and learn underlying rules and patterns in the data. Additionally, the structure of a NN, such as the number of layers and the number of neurons per layer, can be flexibly adjusted based on the task requirements. This flexibility allows NN to accommodate different scales and complexities of temperature and humidity prediction tasks [10].

#### *2.2.4 Random Forest Model(RF)*

(RF is an ensemble learning technique that enhances model accuracy and robustness by generating multiple decision trees and synthesizing their predictions into a cohesive output. RF constructs the model through the following steps. First, Bootstrap sampling is performed, where multiple subsets of the original dataset are drawn with replacement. Second, random feature selection is applied during the construction of each decision tree, where only a subset of features is chosen to split nodes. Third, decision trees are constructed independently based on the sampled data and randomly selected features without pruning. Finally, the predictions are aggregated; for classification problems, the final prediction is determined by majority voting, and for regression problems, the average of all tree predictions is taken as the final prediction. RF, composed of multiple decision trees, effectively handles complex nonlinear relationships between input features and temperature or humidity. Each decision tree captures different patterns and feature combinations, allowing RF to combine these results and provide more accurate predictions.

Advantages for predicting temperature and humidity include the following. RF, composed of multiple decision trees, effectively handles complex nonlinear relationships between input features and temperature or humidity. Each decision tree captures different patterns and feature combinations, allowing RF to combine these results and provide more accurate predictions. Additionally, each decision tree in the RF can be constructed independently, making the algorithm well-suited for parallel computation. This enables efficient use of multi-core processors for large-scale datasets, significantly reducing training time.

## 3 Results Analysis

This paper employs the aforementioned four models for training and prediction. However, it is also necessary to determine which of these models is the most optimal to facilitate better and more accurate predictions. Based on this, the paper utilizes several methods to evaluate and explain the models. Below are the four evaluation methods we used.

### 3.1 Confusion Matrix

A confusion matrix is a tabular tool for evaluating a classification model's performance. as shown in Fig. 1, Fig. 2, Fig. 3, and Fig. 4:

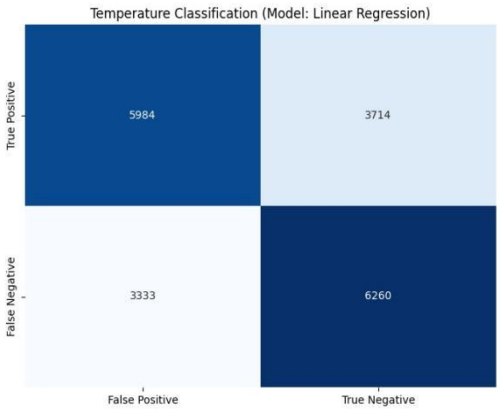


**Fig. 1** Confusion Matrix of the LR Model in temperature

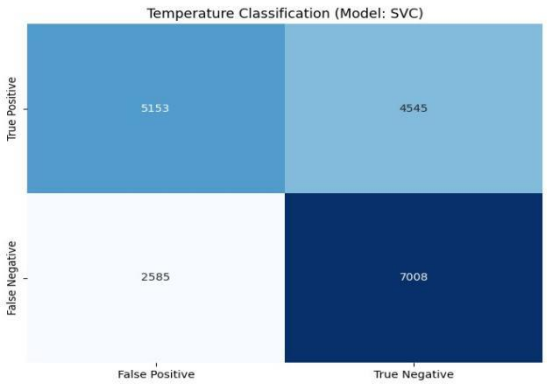


**Fig. 2** Confusion Matrix of the SVM Model in temperature

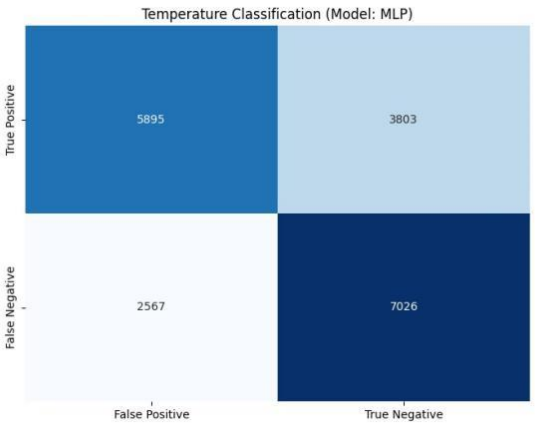


**Fig. 3** Confusion Matrix of the NN Model in temperature

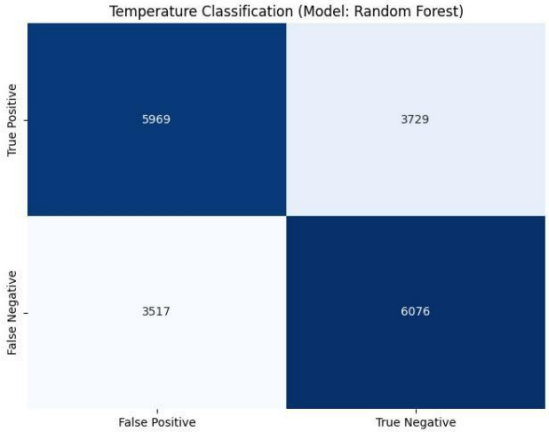


**Fig. 4** Confusion Matrix of the RF Model in temperature

Based on the confusion matrices provided in Fig. 1, Fig. 2, Fig. 3 and Fig. 4, it can be observed that the LR model in Fig. 1 shows a moderate balance between True Positives with 5984 and True Negatives with 6260, but the number of False Positives, 3714, is slightly higher than that of False Negatives, 3333, indicating that the model tends to over-predict the positive class. The Support Vector Classifier model in Fig. 2 demonstrates a high True Negative count of 7008 and reduces False Negatives to 2585, though it has a relatively high False Positive rate of 4545, suggesting that the model is more cautious in identifying the negative class. The MLP model in Fig. 3 performs best in reducing False Negatives, with the lowest count of 2567, indicating a high capability in identifying the positive class while maintaining a high True Negative count of 7026. The RF model in Fig. 4 shows a balanced performance between True Positives with 5969 and True Negatives with 6076, but it has the highest False Negative count of 3517, indicating some difficulties in correctly identifying the positive class. In conclusion, the MLP model exhibits the most balanced performance overall, while the Support Vector Classifier and RF models have deficiencies in terms of False Positives and False Negatives, respectively.

### 3.2 MAE and MSE

MSE and MAE are metrics for evaluating regression model accuracy, with smaller values indicating lower prediction errors and higher accuracy. As shown in Table 1and 1 and Table 2:

**Table 1** The four prediction models for MAE and MSE for temperature prediction

| Model | MSE | MAE |
|---|---|---|
| LR Model | 81. 6240 | 7. 4221 |
| SVM Model | 82. 6678 | 7. 4935 |
| NN Model | 69. 7983 | 6. 7367 |
| RF Model | 82. 0609 | 7. 2129 |

**Table 2** The four prediction models for MAE and MSE for humidity prediction

| Model | MSE | MAE |
|---|---|---|
| LR Model | 0. 0359 | 0. 1567 |
| SVM Model | 0. 0358 | 0. 1600 |
| NN Model | 0. 0337 | 0. 1515 |
| RF Model | 0. 0394 | 0. 1588 |

### 3.3 Cross-Validation Method

Cross-validation is a technique used to evaluate a model's generalization by dividing the dataset into subsets and repeatedly training and validating on these subsets. It helps prevent overfitting and offers a reliable estimate of the model's performance on new data. Below are the MAE and MSE for temperature and humidity predictions by each model, As shown in Table 3:

**Table 3** The four prediction models for Cross-Validation for temperature and humidity prediction

| Model | temperature | humidity |
|---|---|---|
| LR Model | 7. 4237 | 0. 1569 |
| SVM Model | 7. 5313 | 0. 1601 |
| NN Model | 6. 7250 | 0. 1504 |
| RF Model | 7. 2007 | 0. 1585 |

From the data in Table 1, Table 2 and Table 3, it is evident that the NN model performs the best in predicting both temperature and humidity. Specifically, the NN model has a MSE of 69. 7983 and a MAE of 6. 7367 in temperature prediction, while in humidity prediction, it has a MSE of 0. 0337 and a MAE of 0. 1515. These values are lower than those of the other models, indicating higher predictive accuracy. In comparison, the SVM and RF models perform slightly worse across all metrics, for instance, the SVM model has a MAE of 7. 4935 in temperature prediction and 0. 1600 in humidity prediction, both higher than those of the NN model.

## 4 Discussion

Based on the evaluation of the four models mentioned above, we can conclude that the NN Model is the optimal choice. Therefore, let's analyze the reasons for the prediction inaccuracies in the other three models.

From the data above, it is evident that the LR model performs poorly in prediction, and the reasons are as follows. LR is highly sensitive to outliers in the data. Extreme temperature or humidity values may disproportionately influence the regression coefficients, thereby affecting the overall predictive performance of the model. Additionally, due to the simplicity of the LR model, it may fail to capture the complex interactions among meteorological variables, leading to underfitting, where the model performs poorly on both training and test data. This limitation is particularly evident when multiple complex meteorological factors interact.

At the same time, the prediction results of the SVM model are also unsatisfactory, possibly because SVM's performance is highly dependent on the selection of hyperparameters, such as the type of kernel function, regularization parameter CCC, and kernel parameters. These hyperparameters require tuning through cross-validation, which increases the complexity of model building.

RF models consist of multiple decision trees, which leads to high model complexity, especially when the number of trees is large. While the construction of individual trees is relatively simple, training and predicting with a large forest can require substantial computational resources.

From the data presented, it can be observed that the RF model exhibits a relatively high bias. This may be attributed to the fact that the RF model is composed of multiple decision trees, resulting in increased model complexity, particularly when the number of trees is large. Although the construction of individual trees is relatively straightforward, training and making predictions with a large forest can require substantial computational resources.

## 5 Conclusion

This paper comprehensively studies the application of four machine learning models, LR, SVM, NN, and RF, in predicting temperature and humidity, which are crucial for various sectors. The primary focus is utilizing these models to analyze and compare their accuracy and performance in forecasting temperature and humidity. The research process involves extensive data collection, meticulous preprocessing, model training, and thorough evaluation to determine which model can deliver the most accurate predictions. The study's results clearly indicate that the NN model significantly outperforms the others in capturing the complex nonlinear relationships present within the data, leading to more precise and reliable predictions. In contrast, while still effective, the other models are slightly less capable of handling these intricate relationships. Future research could further enhance model accuracy by incorporating more diverse and comprehensive datasets, such as wind speed and atmospheric pressure, and by exploring more advanced machine learning algorithms or developing sophisticated ensemble methods that combine the strengths of multiple models. This approach could eventually enable the precise and reliable prediction of complete weather systems. Moreover, this research provides crucial and more reliable decision-making support for weather-dependent industries, including agriculture, energy management, and public safety. It plays a significant role in addressing the multifaceted challenges posed by climate change and offers numerous possibilities for effectively tackling the complex issues that arise from it.

## References

1. Lorenc A C. Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society, 1986, 112(474): 1177-1194.
2. Alley R B, Emanuel K A, Zhang F. Advances in weather prediction. Science, 2019, 363(6425): 342-344.
3. Bauer P, Thorpe A, Brunet G. The quiet revolution of numerical weather prediction. Nature, 2015, 525(7567): 47-55.
4. Kimura R. Numerical weather prediction. Journal of Wind Engineering and Industrial Aerodynamics, 2002, 90(12-15): 1403-1414.
5. Radhika Y, Shashi M. Atmospheric temperature prediction using support vector machines. International journal of computer theory and engineering, 2009, 1(1): 55.
6. Chen J, Wang H, Xie P. Pavement temperature prediction: Theoretical models and critical affecting factors. Applied thermal engineering, 2019, 158: 113755.
7. Shank D B, Hoogenboom G, McClendon R W. Dewpoint temperature prediction using artificial neural networks. Journal of applied meteorology and climatology, 2008, 47(6): 1757-1769.

8. Dutta, Bimal, and Susanta Mitra. "Better prediction of humidity using artificial neural network. " Fourth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2011). IEEE, 2011.
9. Lu T, Viljanen M. Prediction of indoor temperature and relative humidity using neural network models: model comparison. Neural Computing and Applications, 2009, 18: 345-357.
10. Lorenz E N. Energy and numerical weather prediction. Tellus, 1960, 12(4): 364-373.