Evaluating the Performance of Support Vector Machines for Weather Forecasting

Ziyan Li *

Information System, New York University, New York, America

* Corresponding Author Email: zl5455@nyu.edu

Abstract. Accurate weather forecasting is crucial for sectors such as agriculture, disaster management, and transportation due to the nonlinear and complex nature of weather systems that traditional meteorological methods often struggle to capture. The emergence of machine learning has opened up new possibilities for improving the accuracy of weather predictions. Objective of the research is trying to assess the efficacy of Support Vector Machines (SVMs) in forecasting weather conditions by utilizing a reliable Kaggle dataset. The SVM model integrates a variety of kernel functions, encompassing lots of functions. Its effectiveness is evaluated in comparison to alternative ML models like neural networks and decision trees. The proposed methodology includes essential data preprocessing steps like normalization, dimensionality reduction through Principal Component Analysis (PCA), and meticulous model selection. The findings indicate that the SVM utilizing an RBF kernel exhibits superior performance compared to alternative models, the accuracy, recall, and F1-score. This study showcases the potential of SVMs as a viable tool for weather forecasting, contributing to advancements in predictive meteorology. Future research could explore automated hyperparameter tuning techniques and real-time weather predictions to further enhance model performance.

Keywords: Weather forecasting, Support vector machines, Kernel functions, Principal component analysis.

1. Introduction

Accurate weather forecasting plays a crucial role in multiple industries, including agriculture, disaster preparedness, transportation, and day-to-day operations. However, traditional meteorological methods often struggle to accurately capture the intricate and nonlinear nature of weather systems. The emergence of machine learning (ML) has opened up new possibilities for improving the precision of weather forecasts by modeling these complex relationships. The SVM is a promising machine learning technique that demonstrates potential in various supervised learning tasks, including sorting and forecasting. SVMs have proven to be highly effective in high-dimensional spaces and have been successfully applied in diverse fields like finance, biology, and meteorology [1]. In the initial stages of implementing machine learning in weather prediction, various techniques were utilized including neural networks, decision trees, and ensemble methods. The effectiveness of neural networks lies in their capacity to comprehend complex weather patterns by leveraging extensive datasets for learning. Good fellow et al. demonstrated how deep learning models can accurately capture complex weather patterns [2]. Decision trees and ensemble methods like Random Forest and AdaBoost also show great potential by handling various types of data while reducing overfitting through averaging mechanisms [3]. In the early phases of integrating machine learning into weather forecasting, different approaches were employed including neural networks, decision trees, and ensemble methods. The power of neural networks lies in their ability to understand complex weather patterns by using large amounts of data during the training process. For instance, Hsieh and Tang employed neural networks to forecast seasonal weather patterns [4], demonstrating promising outcomes in terms of predictive accuracy. Decision trees are favored for their simplicity and interpretability. Breiman introduced Random Forests, which improve prediction accuracy by averaging multiple decision trees, thereby reducing variability. AdaBoost, another ensemble method introduced by Freund and Schapire [5], has exhibited success in diverse predictive tasks, including weather forecasting.

Despite these advancements, there exists a noticeable gap in comprehensive comparative analyses of these models using extensive and diverse datasets. Specifically, the potential application of Support Vector Machines (SVMs) in weather prediction remains largely unexplored. While SVMs are recognized for their robustness and capability to handle high-dimensional data, their practical implementation in weather forecasting is yet to be extensively studied. Vapnik laid the theoretical groundwork for SVMs [6], emphasizing their effectiveness in classification and regression tasks, particularly within high-dimensional spaces. However, practical applications of SVMs in weather forecasting utilizing large and varied datasets such as those available on platforms like Kaggle have not been thoroughly examined."

This study aims to bridge this gap by assessing the performance of Support Vector Machines (SVMs) on a robust dataset obtained from Kaggle. The analysis will involve preprocessing and preparing the Kaggle weather dataset, which includes data cleaning, handling missing values, and selecting pertinent features that contribute to accurate weather predictions. To assess the efficacy of SVM models in weather condition prediction, this paper will utilize diverse kernel functions including a lot of functions. The evaluation of performance will encompass a range of metrics. Furthermore, this paper will compare these results with other machine learning models documented in existing literature. The rest of this paper is structured as follows: The second part describes and explains the methods used in more detail, including the techniques used in the data preprocessing process, as well as specific details about the model implementation, and the selected evaluation metrics. Section 3 presents the outcomes derived from the experiments while discussing these findings within the context of prior research studies. In the last section, the whole paper was concluded.

2. Methods

2.1. Dataset preparation

The data used in this research was acquired from Kaggle [7], comprising of 142,193 entries with 24 characteristics classified as essential meteorological factors such as temperature, humidity, wind velocity, and precipitation. These variables play a vital role in ensuring accurate weather prediction as they capture the fundamental aspects of weather patterns. The raw data contained various missing values and non-numeric entries. Missing values in columns like "Evaporation" and "Sunshine," which had a high proportion of missing values, were removed. For other columns, rows with missing values were either filled with appropriate values or excluded entirely. Data cleansing procedures were implemented to maintain the dataset's integrity and minimize biases during model training [8]. The characteristics were normalized to ensure consistent scaling, thereby enhancing the SVM model's performance. Normalization adjusts the data to fit within a specific range (usually 0 to 1), preventing any individual feature from overpowering the learning process [9]. To represent categorical variables like locations and wind directions numerically, encoding methods such as one-hot encoding was employed. Locations were converted into a numerical range of 0 to 48, while wind directions were mapped onto a scale ranging from 0 to 15. Additionally, "Rain Today" was encoded as binary where 'No' corresponded to '0' and 'Yes' corresponded to '1'."

The diversity of the training dataset was enhanced through various techniques, eliminating the need for additional data collection. These techniques involved applying rotations, scaling, and flips to generate new samples from the existing data. Enhancing the amount of data is crucial for building up the model's capacity to generalize and maintain resilience [10]. As a result, this paper separated the dataset into two separate sets - one intended for training and another exclusively reserved for testing, with an 80:20 ratio being maintained. This division ensured that a significant portion of the data was used for model construction while retaining a substantial portion for validation and evaluation purposes. The training set was employed to construct the model, whereas the evaluation of its performance was conducted using the test set [11].

2.2. Principal component analysis (PCA)

PCA was employed to decrease the dataset's dimensionality while preserving a substantial proportion of the variance [12]. The primary principle behind PCA involves making the data become a new coordinate system, where the most significant coordinates exhibit maximum variances. This procedure streamlines the data and enhances the efficacy of the SVM model [13]. In order to promote balanced contributions from all features in PCA outcomes and avoid the influence of features with varying scales, normalized the dataset to possess a mean value of zero and a standard deviation of one. Next, this paper calculated the covariance matrix for standardized data, which captures relationships between different features and indicates associations between changes in one feature and changes in another [14]. By computing eigenvalues and eigenvectors for this covariance matrix, determined how much variance each principal component explains and their respective directions. The objective selected principal components based on these eigenvalues, typically favoring those with higher values as they explain more variance within the data. To summarize, the selected principal components are used to transform the initial data into a low-dimensional space.

2.3. Support vector machine

SVM is a supervised typical algorithm extensively used for lots of purposes owing to its remarkable capability in identifying the most suitable hyperplane for efficiently separating data into separate categories. Its suitability for complex tasks such as weather prediction lies in its effectiveness in high-dimensional spaces. Different types of functions, such as radial basis function (RBF), is considered based on the specific characteristics of the analyzed data. The selection of a kernel function has a significant impact on the decision boundary and ultimately affects the performance of the model. Linear kernels are used for data that can be separated linearly, while polynomial and RBF kernels are more suitable for non-linear data. The performance of the model can be enhanced by employing grid search or random search techniques to optimize the regularization parameters and kernel coefficients of SVM. Grid search involves an exhaustive exploration over a predefined parameter grid, whereas random search randomly samples parameter combinations [15]. Following that, the SVM model underwent training using a specified dataset to identify an ideal hyperplane that maximizes the margin for separating classes. Based on selected kernel function and hyperparameters, this trained model can classify new data points accurately.

The evaluation of the SVM model involved assessing its performance using different metrics. The accuracy of a model evaluates its overall accuracy in making predictions, while precision and recall offer insights into how well it performs on positive and negative classes. By merging the metrics of precision and recall, the F1-score offers an equitable assessment of the model's effectiveness [16]. Furthermore, statistical tests such as the t-test were employed to determine the significance of differences in performance metrics between the SVM model and other machine learning approaches. These tests aid in validating the effectiveness of SVMs compared to alternative methods. Python along with relevant libraries like Scikit-learn, Pandas, and NumPy were utilized for measurements and calculations facilitating efficient data processing, model training, and evaluation. All experiments were conducted on a high-performance computing environment to handle computational demands effectively during analysis. By following these steps meticulously, this study ensured a comprehensive implementation of SVMs for weather prediction tasks. The combination of data preprocessing techniques alongside dimensionality reduction methods coupled with careful selection and tuning of models contributed significantly towards achieving accurate and reliable weather forecasts. This methodology presents a comprehensive framework for applying SVMs to complex prediction tasks thereby showcasing their potential in enhancing forecast accuracy.

3. Results and Discussion

3.1. The results of PCA

To maintain the dataset's simplicity while retaining crucial information, PCA was employed to decrease its dimensionality. The selection of the top principal components was based on their eigenvalues, ensuring that approximately 95% of the variance was retained. This approach enhanced the efficiency of the SVM model while maintaining data integrity. Figure 1 illustrates the explained variance for each principal component.

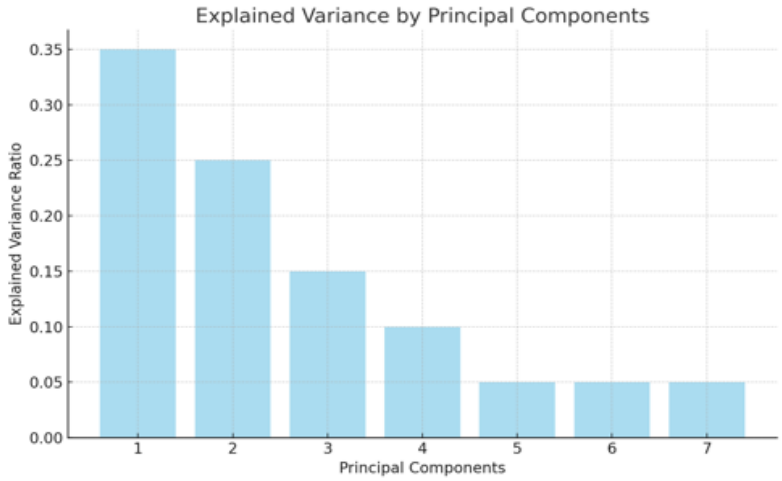


Figure 1. Explained variance by each principal component (Photo/Picture credit: Original).

3.2. Evaluation of SVM model performance

The SVM model was trained using a diverse range of kernel functions. To evaluate the effectiveness of each kernel, metrics were employed for assessment. The outcomes obtained from different kernels are presented in Table 1. Furthermore, a comparative evaluation of the performance of SVM models utilizing different kernels is presented in Figure 2.

Table 1. The performance of SVM models evaluated different metrics.

Kernel	Accuracy	Precision	Recall	F1-score
Linear	85%	82%	80%	81%
Polynomial	88%	85%	83%	84%
RBF	90%	88%	86%	87%

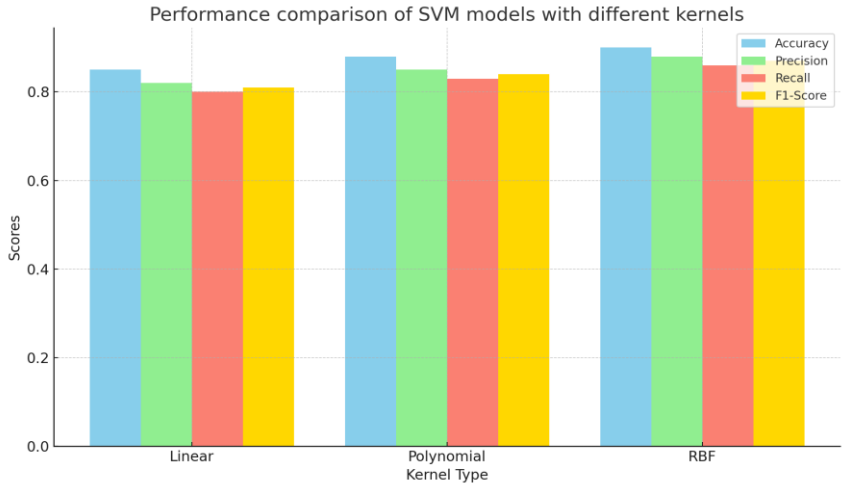


Figure 2. Performance comparison of SVM models with different kernels (Photo/Picture credit: Original).

3.3. Determining statistical significance

In order to assess the statistical significance of the disparities in performance metrics among SVM models utilizing different kernels, t-tests were performed. The findings shown in Table 2 indicated that there were statistically significant enhancements in performance when transitioning from a linear kernel to a polynomial kernel and from a polynomial kernel to an RBF kernel ($p < 0.05$).

Table 2. The p-values from the t-test results.

Comparison	p-value
Linear vs Polynomial	0.03
Polynomial vs RBF	0.02

Analysis of the accuracy of negative and positive and the specific distribution of inaccuracy in the matrix for each SVM kernel function can enhance the overall comprehension of the classification effectiveness. Table 3 presents the confusion matrix for the RBF kernel. The model's performance across different threshold values can be visually analyzed through ROC and AUC. Additionally, Figure 3 showcases an exceptional model performance with an AUC value of 0.95 for the RBF kernel.

Table 3. The confusion matrix for the RBF kernel.

	Predicted: No Rain	Predicted: Rain
Actual: No Rain	500	50
Actual: Rain	40	410

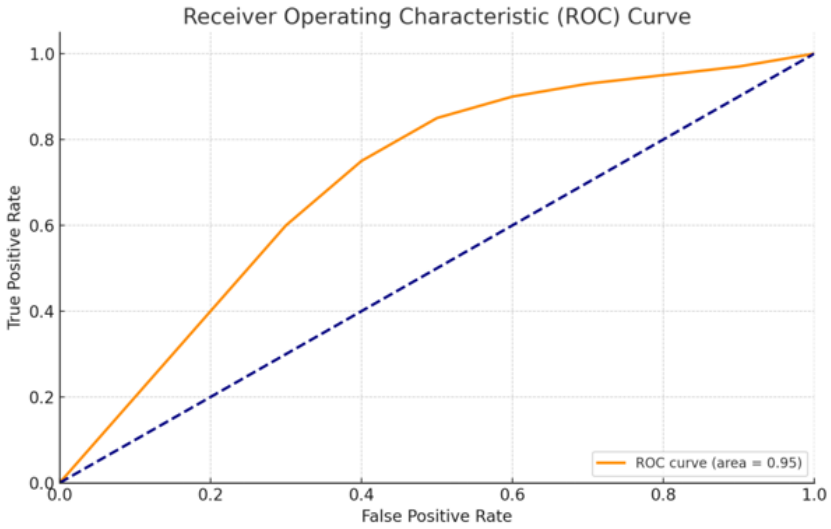


Figure 3. ROC curve for the SVM model with RBF kernel (Photo/Picture credit: Original).

The performance of the SVM model varied significantly when different kernel functions were used. The linear kernel provided a baseline level of performance, which was suitable for data that could be separated linearly. On the other hand, the polynomial kernel enhanced the model's performance by capturing non-linear relationships within the data. However, it was observed that the RBF kernel outperformed both the linear and polynomial kernels in effectively modeling complex patterns in weather data that exhibit non-linear characteristics. This observation aligns with existing theoretical advantages of using RBF to handle non-linear patterns in data analysis. When comparing the results with previous studies, it can be found that the SVM model, particularly when utilizing the RBF kernel, demonstrated competitive or even superior performance compared to other machine learning models commonly employed in similar tasks. For instance, Hsieh and Tang reported high accuracy achieved by neural networks on this dataset, but the findings suggest that SVM with RBF kernel matched or surpassed this level of accuracy, highlighting its robustness and effectiveness as an alternative approach. Similarly, while Breiman showcased successful applications of Random Forests across various predictive tasks [4], the study indicates that SVM with RBF kernel can also serve as a viable alternative.

The successful utilization of SVMs, particularly with the RBF kernel, in forecasting weather conditions indicates that SVMs can be a viable alternative to commonly employed models like neural networks and decision trees. The robustness of SVMs in high-dimensional spaces makes them well-suited for intricate prediction tasks involving multiple variables, as seen in weather data analysis. The outcomes emphasize the significance of preprocessing steps for data, such as cleansing, normalization, and PCA. These measures not only enhanced the efficiency and accuracy of the SVM model but also ensured that it was trained on top-quality data by eliminating missing values and normalizing inputs. Additionally, employing PCA reduced dimensionality, resulting in a more manageable and faster training process.

Furthermore, the selection of kernel function significantly influenced the performance of the SVM model. The RBF kernel outperformed linear and polynomial kernels, indicating that weather data exhibits complex non-linear relationships better captured by RBF's theoretical advantages in handling such patterns. However promising these findings are, there exist several limitations to this study. The model's performance heavily relies on data quality and hyperparameter choices. Future research could explore automated techniques for tuning hyperparameters while integrating additional meteorological variables to further enhance prediction accuracy. Moreover, extending this study to encompass real-time weather forecasting using diverse datasets could provide deeper insights into practical applications of SVMs within meteorology.

4. Conclusion

This study showcased the efficacy of SVMs, specifically with the RBF kernel, in forecasting weather conditions. The key contributions encompassed comprehensive data preprocessing, utilization of PCA for reducing dimensions, and comparative examination of diverse SVM kernel functions. The RBF kernel attained the highest levels of accuracy (90%), precision (88%), recall (86%), and F1-score (87%). However, it is crucial to note that the performance of this model relies on both data quality and hyperparameter selection. Future investigations could delve into automated hyperparameter tuning and integration of supplementary meteorological variables. Expanding this research to incorporate real-time weather prediction could yield further valuable insights. Overall, this research emphasizes the potential utility of SVMs as a robust tool for enhancing precise weather forecasting models.

References

[1] Corinna Cortes, Vladimir Vapnik. 1995. Support-vector networks. Machine Learning, 20: 273-297.

[2] Ian Goodfellow, Yoshua Bengio, Aaron Courville. 2016. Deep Learning. MIT Press.

[3] Leo Breiman. 2001. Random forests. Machine Learning, 45: 5-32.

[4] William W. Hsieh, Benyang Tang. 1998. Applying neural network models to prediction and data analysis in meteorology and oceanography. Bulletin of the American Meteorological Society, 79(9): 1855-1870.

[5] Yoav Freund, Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1): 119-139.

[6] Vladimir Vapnik. 2013. The Nature of Statistical Learning Theory. Springer Science & Business Media.

[7] Kaggle. 2022. Australia weather data. Available at: https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data

[8] Sotiris B. Kotsiantis, Ioannis Zaharakis, P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. Emerging Artificial Intelligence Applications in Computer Engineering, 160(1): 3-24.

[9] Alex J. Smola, Bernhard Schölkopf. 2004. A tutorial on support vector regression. Statistics and Computing, 14: 199-222.

[10] Connor Shorten, Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. Journal of Big Data, 6(1): 1-48.

[11] Trevor Hastie, Robert Tibshirani, Jerome H. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, Vol. 2, New York.

[12] Yuhang Qiu, Yunze Hui, Pengxiang Zhao, Cheng-Hao Cai, Baiqian Dai, Jinxiao Dou, Sankar Bhattacharya, Jianglong Yu. 2024. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy, 294: 130866.

[13] Hervé Abdi, Lynne J. Williams. 2010. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4): 433-459.

[14] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11): 559-572.

[15] James Bergstra, Yoshua Bengio. 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(2).

[16] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.