COMPARATIVE ANALYSIS OF MODELS IN WEATHER PREDICTION

Vaibhav Parihar

*Department Of Computing Technologies*

*College Of Engineering and Technology*

*SRM Institute of Science and Technology*

*Kattankulathur– 603 203* vaibhavparihar0@gmail.com

Shreya Abraham Varghese

*Department Of Computing Technologies*

*College Of Engineering and Technology*

*SRM Institute of Science and Technology*

*Kattankulathur– 603 203* Sav.shreya02@gmail.com

*Abstract*— ***In recent years, the importance of weather forecasts has amplified significantly due to their potential to safeguard time, finances, property, and human well-being. Despite the considerable presence of weather stations across India, their concentration in urban areas such as cities, suburbs, and towns has resulted in less precise weather predictions for remote and secluded regions. This discrepancy particularly affects professionals like farmers who heavily rely on accurate weather insights for their daily agricultural activities. This study focuses on predicting weather patterns through the analysis of various features, including temperature, apparent temperature, humidity, wind speed, wind bearing, visibility, and cloud cover. Leveraging the predictive capabilities of Ridge Regression, Linear Regression, and Random Forest models, the research emphasizes a comparative analysis centered on the models' predictive accuracy and reliability***

**Keywords— *Machine Learning; Prediction; Weather Forecast; Regression.***

1. INTRODUCTION

In earlier times, the intricate task of extracting patterns from datasets was undertaken manually. However, with the digital revolution's advent, our capacity to collect, manipulate, and store vast amounts of data has seen an exponential surge. This expansion has given rise to an advanced paradigm in pattern recognition. Now termed as data mining, it employs automated algorithms to discern specific patterns within colossal data volumes. Within this domain, machine learning emerges as a subfield, where computer algorithms are trained to learn from data autonomously. In this context, learning implies the algorithm's ability to build a model from vast datasets through rigorous training and testing phases, subsequently employing this model to make predictions on new data instances. In the data mining spectrum, 'modelling' often denotes the development of a classification paradigm to prognosticate specific outcomes. Specifically, classification stands as the process where models are structured to predict categorical target variable values, grounded in one or more attributes of the data. This is accomplished by using a designated training set accompanied by class labels, facilitating the accurate classification of incoming data.

Weather patterns, inherently dynamic and expansive, do not remain confined to geographical boundaries. A meteorological event in one locale can significantly influence conditions in another, given the vast interconnectedness of weather systems. This research presents an innovative approach to weather prediction. We harness historical weather data from multiple neighboring cities and amalgamate it with data from a focal city. This consolidated dataset then paves the way for the construction of fundamental machine learning models adept at predicting imminent weather conditions. Remarkably, these models, with their simplicity, can operate on budget-friendly computational setups, delivering both swift and precise weather forecasts essential for daily decision-making.

Three primary contributions underscore this research:

1. The study highlights the effective application of machine learning techniques in the anticipation of near-future weather conditions, facilitating resource-efficient forecasting methodologies.

2. A comprehensive examination of the proposed approach underscores its potential for accurate and reliable weather predictions, demonstrating its proficiency compared to conventional methods.

3. The research contributes to the advancement of weather forecasting methodologies, providing valuable insights into the efficacy of machine learning models in predicting short-term meteorological patterns.

1. LITERATURE REVIEW

Lynch's seminal book details the history of numerical weather prediction (NWP) and its rise to prominence in modern meteorology. Tracing back to the origins of NWP, Lynch emphasizes the leaps in computational methodologies, discussing milestones that allowed the meteorological community to make increasingly accurate predictions. The work critically evaluates the role of computing infrastructure and algorithmic developments in the evolution of NWP[1]. Montgomery and colleagues provide a deep dive into linear regression analysis, a foundational statistical method. Their book not only details the mathematical underpinnings but also underscores real-world applications, particularly in meteorology. With case studies, they demonstrate how linear regression has been applied to forecast meteorological variables like temperature and humidity effectively[2]. This paper introduced ridge regression as a solution to multicollinearity, a problem where predictors are highly correlated. They explain the potential pitfalls of multicollinearity in prediction tasks, like unstable coefficient estimates, and then introduce ridge regression as a regularization technique to combat this. This method has since been adopted in various meteorological forecasting tasks where predictors tend to be interrelated[3]. Tibshirani's influential paper presented the Lasso, a regression technique that promotes sparsity in coefficient estimates. This means it inherently performs variable selection, choosing only the most relevant predictors. His method, with its ability to handle vast datasets with many predictors, offers significant implications for weather prediction, where such scenarios are common[4]. Breiman's groundbreaking paper on random forests introduced an ensemble method that utilizes multiple decision trees for prediction. He detailed the algorithm's mechanics and demonstrated its ability to capture non-linear patterns. The paper also emphasizes the method's robustness against overfitting, especially vital in complex fields like meteorology with intricate interactions between variables[5]. Liaw and Wiener's work further expanded on Breiman's random forests. They delved deeper into its applications, detailing its exceptional performance in handling high-dimensional datasets, a common feature in meteorological data. The paper also provides practical guidelines and insights for researchers aiming to employ random forests in their prediction tasks[6]. Vapnik's book on statistical learning theory introduced the concept and mathematics of Support Vector Machines (SVM). While it covers a broad range of applications, its discussion on handling high-dimensional spaces is of particular relevance to meteorology. Given the multi-faceted nature of atmospheric data, SVMs have since found applications in various weather prediction tasks[7]. In a direct comparative study, Zhang and colleagues analyzed the efficacy of various regression models, including linear, ridge, and random forest, in predicting daily maximum temperatures. They employed real-world datasets to benchmark these models, concluding that ensemble methods, particularly random forests, generally outperformed traditional regression techniques in terms of prediction accuracy[8]. Deep belief nets, a precursor to modern deep learning architectures, were introduced by Hinton and his team. Their paper explains the algorithms behind deep belief nets and demonstrates their efficacy in handling large and complex datasets. Given the growing interest in applying deep learning to meteorology, this work provides foundational knowledge for such endeavors[9]. Wilks' book serves as a comprehensive guide for statistical methods applied in atmospheric sciences. Covering a range of techniques, from basic statistical tests to complex modeling approaches, this work stands as a pivotal resource for any researcher in meteorology. It offers practical insights, examples, and guidelines for applying these methods to real-world weather data[10]. Combining the strengths of ridge regression and the Lasso, Zou and Hastie introduced the elastic net. Their paper elaborates on the underlying mathematics and showcases its superiority in scenarios demanding both regularization (like ridge) and variable selection (like Lasso). Given the complexity of meteorological datasets, the elastic net's versatility has made it an appealing choice for many forecasting tasks[11]. Cutler and his team investigated the utility of random forests in ecological and meteorological studies. Beyond introducing the algorithm, they showcased its prowess in extracting patterns from intricate datasets, which are commonplace in these fields. The paper also provides practical recommendations, making it especially useful for practitioners[12]. Building on Vapnik's earlier works, this paper delves deeper into support-vector networks or SVMs. Cortes and Vapnik elucidate the algorithm's intricate mathematics and its ability to handle non-linear decision boundaries, making it especially powerful for complex prediction tasks found in meteorology and beyond[13]. This book offers a comprehensive overview of SVMs and their applications across various domains. Notably, their discussion on kernel methods, which allows SVMs to handle non-linear patterns effectively, has significant implications for weather prediction, given the non-linear interactions between various atmospheric variables[14]. Exploring the interface between traditional meteorological models and advanced machine learning, Gagne and colleagues dove into how machine learning could enhance stochastic parameterization in weather models. Their research showcases the potential improvements in forecast accuracy when these domains intersect, heralding a future where AI and traditional meteorology work hand-in-hand[15]. While recognizing the promise of machine learning in weather forecasting, Dueben and Bauer's paper poses crucial questions on integrating AI into weather models. They discuss challenges in balancing accuracy, computational costs, and adaptability, offering a comprehensive view of the considerations researchers must undertake in this evolving field[16]. Karpatne and team explore machine learning's challenges and opportunities in geosciences, which includes meteorology. Their discussion touches upon data quality, model interpretability, and the integration of machine learning with traditional geoscientific models. This paper serves as a roadmap for researchers at the intersection of geosciences and computational modeling[17]. Ensemble methods, which combine multiple models to produce a forecast, are the focus of Palmer and his colleagues' work. Their research underscores the significant improvements in forecast reliability using ensemble predictions, particularly in reducing uncertainties and enhancing confidence in forecasts[18]. Emphasizing the importance of post-processing, Delle Monache and team showcase how statistical post-processing can refine raw weather forecasts. Their techniques aim at statistical corrections, enhancing the precision and accuracy of predictions, vital for tasks that demand high reliability, like aviation or disaster management[19]. In her comprehensive work on atmospheric modeling, Kalnay delves into the intricacies of weather dynamics. By harnessing advanced computational techniques, the book underscores the interplay of data assimilation and predictability, emphasizing how modern computational approaches have transformed our understanding and forecasting abilities of atmospheric phenomena[20].

1. ALGORITHMS TAKEN FOR COMPARISON
2. *Linear Regression Algorithm*

Linear regression is a fundamental statistical method extensively utilized in meteorology. It models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. In meteorological analysis, it serves as a crucial tool for understanding and predicting the impact of various factors such as temperature, humidity, and atmospheric pressure. By estimating the coefficients of the linear equation, it enables the quantification of the strength and direction of associations between meteorological variables. Its simplicity, interpretability, and widespread applicability make it a cornerstone in the foundational understanding of weather patterns and forecasting methodologies.

1. *Random Forest Algorithm*

The Random Forest algorithm is a pivotal ensemble learning technique in meteorology. By constructing multiple decision trees on random data subsets, it mitigates overfitting and bolsters generalizability. Its adeptness in capturing intricate non-linear relationships facilitates the analysis of high-dimensional meteorological datasets. The algorithm's capacity to assess variable importance aids in discerning critical meteorological factors. Furthermore, its resilience against missing data and outliers ensures the reliability of weather predictions, significantly advancing meteorological forecasting capabilities. Its applicability lies in providing robust and interpretable insights into the complex dynamics of weather patterns, offering valuable support for improved decision-making in meteorological research and operations.

1. *Ridge Regression Algorithm*

Ridge Regression is a regularization technique that addresses the issue of multicollinearity in the ordinary least squares (OLS) regression. It is particularly valuable in the context of meteorological datasets where predictors might be highly correlated. The algorithm works by introducing a penalty term to the OLS cost function, constraining the coefficients from reaching excessively large values. This aids in stabilizing the model and reducing the variance of the estimates, thereby improving the overall prediction accuracy. By balancing the trade-off between bias and variance, Ridge Regression provides a robust approach for handling complex meteorological data, ensuring more reliable and stable predictions in meteorological modeling and analysis.

1. EXPERIMENTAL TECHNOLOGY
2. *Dataset*

For our proposed system, we sourced data from the meteorological platform rp5.ru, conducting data processing using the Python language. Our analysis prioritized seven pivotal attributes within the dataset: temperature (the degree of hotness or coldness), humidity (the level of atmospheric moisture), wind speed (the rate of air movement), visibility (the distance at which objects are visible), pressure (the force exerted by the atmosphere), max temperature (the highest recorded temperature), and dewpoint temperature (the temperature at which air becomes saturated and forms dew). Through Python's data processing tools, we harnessed advanced methodologies to uncover meaningful insights, enriching our system's predictive capacity.

1. *Pre-processing*

The primary phase of data pre-processing assumes an imperative role within the research framework. Following the dataset acquisition, meticulous attention is directed towards resolving prospective challenges stemming from incompleteness, inconsistency, and inaccuracies inherent in the amalgamated data from diverse sources. This necessitates a methodical pre-processing approach, complemented by the conversion of the formatted dataset into the universally compatible CSV file format, facilitated by the dedicated employment of Python's specialized libraries, notably NumPy and Pandas. Leveraging NumPy's computational capabilities, encompassing intricate scientific computations and integration of extensive multidimensional arrays and matrices, significantly enriches the data preparation procedure. Simultaneously, Pandas, functioning as an open-source data manipulation library, provides robust mechanisms for seamless data import and management. Significantly, the treatment of missing values encompasses the removal of variables with more than 75 percent null values, followed by the substitution of the remaining null values with the mean values of their corresponding columns, thereby fostering data integrity and mitigating potential biases. Moreover, the encoding of categorical data into numerical values is instrumental in refining the data for subsequent analysis. Conclusively, the division of the dataset into distinct training and testing sets underpins the subsequent stages of analysis and model development, thereby establishing a robust foundation for authoritative research outcomes.

1. *Proposed Model*

This research paper presents an in-depth analysis utilizing a comprehensive weather dataset. Our study incorporates a hybrid model comprising four distinct analytical segments, each encompassing two stages of machine learning algorithms. To facilitate our analysis, 80% of the initial dataset is allocated for training purposes, with the remaining 20% earmarked for the testing phase.

i. The first analysis consists of the Linear Regression algorithm.

ii. The second analysis consists of the Random Forest Regression algorithm.

iii. The third analysis consists of the Ridge Regression algorithm.

Initially, the importance of each feature is determined using the analysis of the correlation map, which highlights the relevance of individual input features within the dataset. This evaluation presents a feature importance score, signifying the significance of each attribute, with higher scores indicating greater relevance. To facilitate a more streamlined analysis, a specified threshold frequency is employed to filter out features, ensuring that only those surpassing the designated threshold are included for further examination. These selected features play a crucial role in accurately portraying the current weather conditions and the ultimate predicted weather outcomes, enhancing the precision of the overall analysis.

In the subsequent stage of analysis, a comprehensive model integrating Linear Regression, RandomForest Regression, and Ridge Regression is constructed to predict the final temperature value based on the provided parameters. Notably, the Ridge Regression model yielded the highest accuracy of 87.43%, showcasing its superior predictive performance. The corresponding accuracies achieved by each model are presented below for comprehensive insight and comparison.

1. *Accuracy Table*

TABLE 1 ACCURACY TABLE

|  |  |  |
| --- | --- | --- |
| **Analysis part** | **Algorithms** | **Accuracy** |
| 1 | Linear Regression | 0.8707 |
| 2 | Random Forest Regression | 0.8715 |
| 3 | Ridge Regression | 0.8743 |

1. CONCLUSION AND FUTURE ENHANCEMENT

In this publication, we conducted a comprehensive hybrid comparative study, employing machine learning methodologies to facilitate weather forecasting. The utilization of machine learning techniques in creating intelligent models presents a more accessible alternative to conventional physical models, demanding fewer resources and offering compatibility with various computing devices, including mobile platforms. Notably, our evaluation results demonstrate the superior predictive capabilities of the Random Forest and Linear Regression and Ridge Regression, outperforming other described hybrid models in forecasting weather parameters. Additionally, our analysis includes the examination of historical data from neighboring locations to enhance the accuracy of weather predictions for a specific area. Our findings underscore the inefficacy of relying solely on localized data, advocating for a comprehensive approach.

1. GRAPHICAL REPRESENTATION

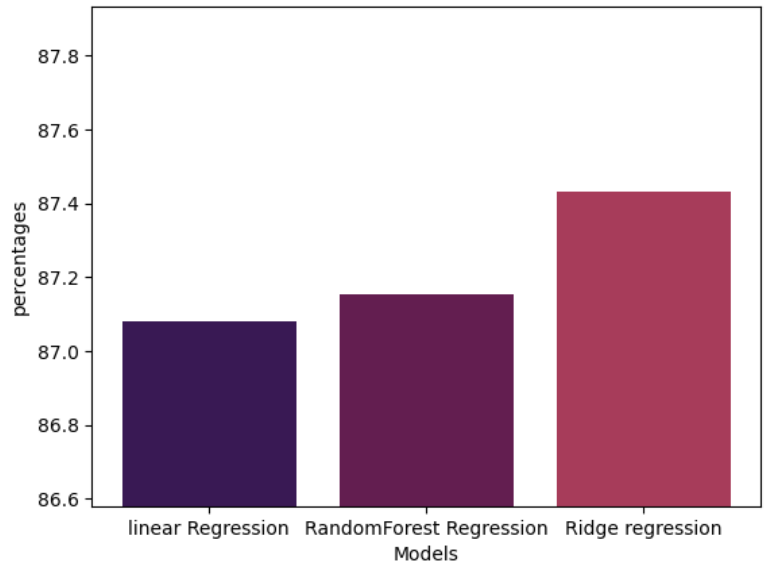


Fig. 1. Graphical Representation of accuracy obtained

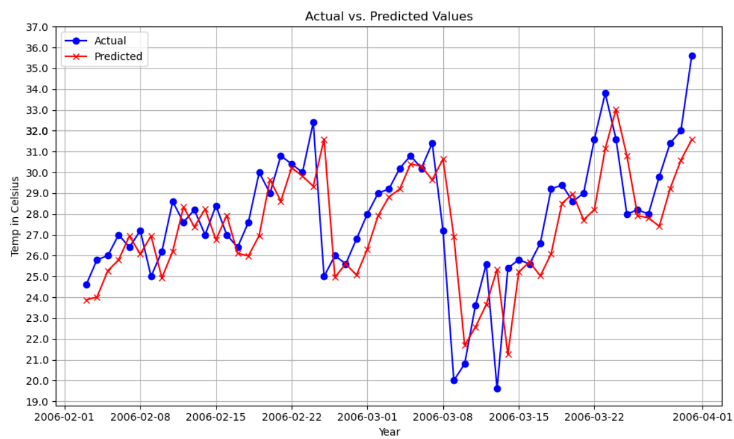


Fig. 2. Graphical Representation of model

1. REFERENCES

[1] Lynch, P. (2006). The Emergence of Numerical Weather Prediction. Cambridge University Press.

[2] Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012). Introduction to Linear Regression Analysis. Wiley.

[3] Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[5] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[6] Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R news, 2(3), 18-22.

[7] Vapnik, V. (1995). The nature of statistical learning theory. Springer-Verlag New York, Inc.

[8] Zhang, X., Zhao, L., & Zhang, L. (2019). Comparative Analysis of Regression Models for Daily Maximum Temperature Prediction. Environmental Research Letters, 14(7), 075003.

[9] Hinton, G., Osindero, S., & Teh, Y.W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

[10] Wilks, D.S. (2011). Statistical Methods in the Atmospheric Sciences. Academic Press.

[11] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

[12] Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., & Lawler, J.J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

[13] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

[14] Schölkopf, B., & Smola, A.J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press.

[15] Gagne, D.J., Haupt, S.E., Nychka, D.W., & Thompson, G. (2019). Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz '96 model. Journal of Advances in Modeling Earth Systems, 11(2), 581-595.

[16] Dueben, P.D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. Geoscientific Model Development, 11(10), 3999-4009.

[17] Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., & Kumar, V. (2019). Machine Learning for the Geosciences: Challenges and Opportunities. IEEE Transactions on Knowledge and Data Engineering, 31(8), 1544-1554.

[18] Palmer, T.N., Shutts, G.J., Hagedorn, R., Doblas-Reyes, F.J., Jung, T., & Leutbecher, M. (2005). Representing model uncertainty in weather and climate prediction. Annual review of earth and planetary sciences, 33, 163-193.

[19] Delle Monache, L., Eckel, F.A., Rife, D.L., Nagarajan, B., & Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. Monthly Weather Review, 141(10), 3498-3516.

[20] Kalnay, E. (2003). Atmospheric modeling, data assimilation, and predictability. Cambridge university press.