MACHINE LEARNING-BASED WEATHER PREDICTION: A COMPARATIVE STUDY OF REGRESSION AND CLASSIFICATION ALGORITHMS

Ayazuddin Ahmed[1], Habeebulah Syed Ataullah[2], Mohammed Abdul Fahad[3], Bhargavi Bendalam [4]

[1,2,3] B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

[4]Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

bhargavi@lords.ac.in

Abstract: *Accurate weather forecasting is crucial across multiple industries like agriculture, transportation, and disaster management, making it a key application for machine learning. This study explores the prediction of various weather conditions—rain, sunshine, clouds, fog, drizzle, and snow—using a range of fundamental machine learning techniques and boosting algorithms. Historical meteorological data, including temperature, humidity, wind speed, and pressure, was used to train and evaluate these algorithms. Tests encompassed well-known methods such as decision trees, random forests, naive Bayes, k-nearest neighbors, and support vector machines. Additionally, boosting methods like XGBoost and AdaBoost were employed to improve forecast precision.*

Results indicated that XGBoost and AdaBoost achieved the highest accuracies (87.86% and 87.33%, respectively) compared to other methods tested. Validation through ROC Curve Analysis and Lift Curve Analysis demonstrated superior performance of XGBoost and AdaBoost in terms of true positive rate, false positive rate, and lift.[1,2]

I. Introduction

Weather forecasting plays a critical role in numerous sectors such as agriculture, transportation, and disaster management, where accurate predictions can significantly impact decision-making and planning. Leveraging machine learning algorithms for weather forecasting has emerged as a promising approach due to their ability to analyze large volumes of historical meteorological data. This study investigates the effectiveness of various machine learning methods and boosting algorithms in predicting diverse weather conditions including rain, sunshine, clouds, fog, drizzle, and snow. By utilizing datasets containing key meteorological variables such as temperature, humidity, wind speed, and pressure, the study evaluates methods ranging from decision trees and random forests to naive Bayes, k-nearest neighbors, and support vector machines.[3] Furthermore, the study employs advanced boosting techniques such as XGBoost and AdaBoost to enhance forecast accuracy. The findings highlight XGBoost and AdaBoost as achieving the highest levels of accuracy (87.86% and 87.33%, respectively), validated through rigorous analysis including ROC Curve and Lift Curve analyses. This research underscores the efficacy of machine learning in improving the precision of weather forecasts, crucial for informed decision-making across various industries.[4]

II. Literature Survey

Qasem Abu Al-Haija et al. (2018) developed a deep learning-based system for classifying weather conditions as either unfavorable or normal for autonomous cars. Their framework utilized three convolutional neural networks (CNNs): SqueezeNet, ResNet-50, and EfficientNet, leveraging transfer learning techniques and Nvidia GPU capabilities. Evaluations on DAWN2020 and MCWRD2018 datasets demonstrated strong classification performance, with ResNet-50 achieving the highest detection accuracy (98.48%), precision (98.51%), and sensitivity (98.41%) specifically trained on weather data.[5]

Sebastian Scher et al. (2020) proposed a deep learning approach integrating convolutional neural networks with weather forecast data. Their method introduced a scalar confidence value to medium-range predictions based on atmospheric state, indicating predictability levels relative to normal conditions. Despite being computationally efficient compared to ensemble models, their approach outperformed traditional non-numerical methods in predicting forecast uncertainty.[6]

Dávid Markovics et al. (2021) evaluated 24 machine learning models over two years of data from 16 photovoltaic plants in Hungary for deterministic day-ahead power forecasting using Numerical Weather Prediction (NWP). Their study highlighted the effectiveness of hyperparameter tuning, identifying kernel ridge regression and multilayer perceptron as top-performing models with a maximum prediction skill score of 44.6%.[7]

A H M Jakaria et al. (2019) demonstrated a method for weather prediction using basic machine learning models trained on historical data from multiple weather stations. Their approach focused on generating timely predictions for specific weather conditions, emphasizing the use of regional weather station data beyond the immediate forecast area. Their evaluations indicated sufficient accuracy for practical application alongside advanced forecasting methods.[8].

III.System Analysis

Existing system: The current weather prediction system predominantly depends on traditional meteorological models and numerical simulations to forecast weather conditions. These systems use physical and mathematical equations to predict parameters such as temperature, precipitation, and wind patterns. Despite their reasonable accuracy, these models often face challenges with fine-scale predictions and real-time adjustments. Machine learning methods have emerged as a promising complement to these traditional approaches. This study seeks to evaluate the effectiveness of machine learning algorithms in enhancing weather predictions by comparing regression and classification techniques with existing meteorological models.[9]

Disadvantages of existing system:

The disadvantages of the existing weather prediction system based on traditional meteorological models and numerical simulations include:

- Limited Spatial and Temporal Resolution: Traditional models may not capture fine-scale local variations and events, leading to less accurate predictions for specific regions and short-term forecasts.

- Computational Complexity: Numerical simulations are computationally intensive, requiring substantial resources and time to generate forecasts, making them less suitable for real-time weather predictions.
- Sensitivity to Initial Conditions: These models are highly sensitive to the accuracy of initial conditions, and even small errors in measurements can lead to significant forecast deviations.[9]

Proposed system:

The proposed weather prediction system aims to enhance forecast accuracy and reliability by integrating machine learning techniques with advanced data sources and real-time information.

This system employs regression and classification algorithms to predict various meteorological parameters, such as temperature, precipitation, humidity, and wind speed.

By utilizing a diverse array of data sources including historical weather records, satellite imagery, IoT sensor data, and crowd-sourced observations—the system can produce more precise and adaptable forecasts. Additionally, it features user-friendly interfaces and emphasizes transparency, making weather predictions accessible to a broader audience.

The proposed system has the potential to overcome the limitations of current weather prediction systems, significantly improving forecasting quality and proving invaluable for sectors reliant on weather forecasts, such as agriculture, transportation, and disaster management.

Advantages of proposed system:

Enhanced Accuracy and Reliability: Integrates machine learning with advanced data for more accurate and reliable forecasts.

Improved Fine-Scale Predictions: Uses regression and classification algorithms to predict meteorological parameters precisely.

Diverse Data Utilization: Leverages historical records, satellite imagery, IoT sensor data, and crowd-sourced observations for comprehensive forecasts.

Real-Time Adjustments: Incorporates real-time information for timely adjustments and improved forecast responsiveness.[10]

IV.System Study

Feasibility study: The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.Three key considerations involved in the feasibility analysis are,

Economical feasibility: This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was

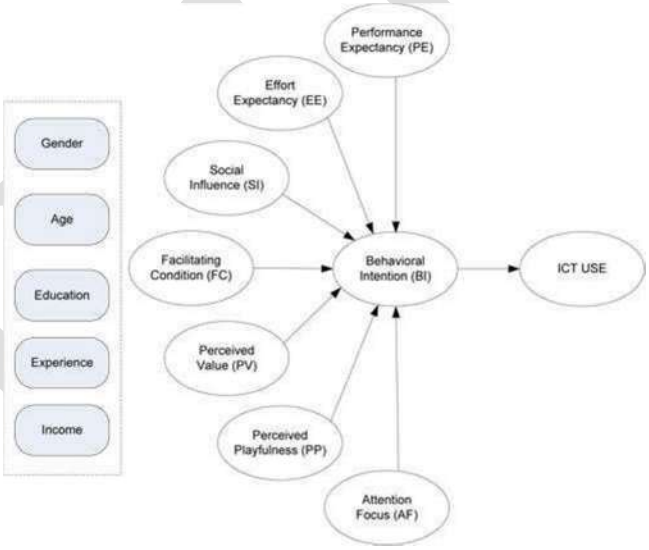
achieved because most of the technologies used are freely available. Only the customized products had to be purchased. [11].

Technical feasibility: This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

Social feasibility: The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

V. System Design

System architecture:



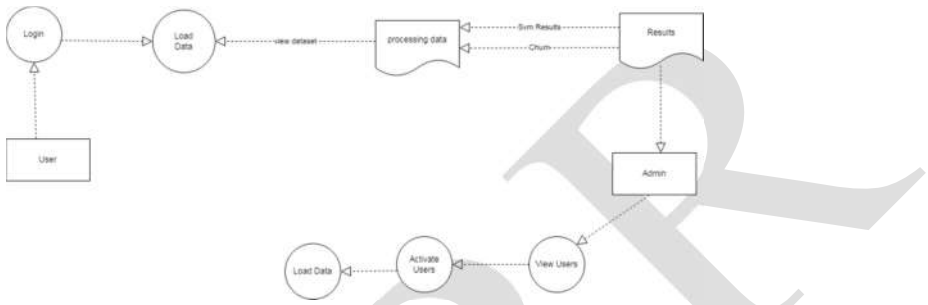
Data flow diagram:

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.[12]

The data flow diagram (DFD) is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.[13]

DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



UML s

UML stands for Unified Modelling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artefacts of software system, as well as for business modelling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

Goals:

The Primary goals in the design of the UML are as follows:

Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.

Provide extendibility and specialization mechanisms to extend the core concepts.

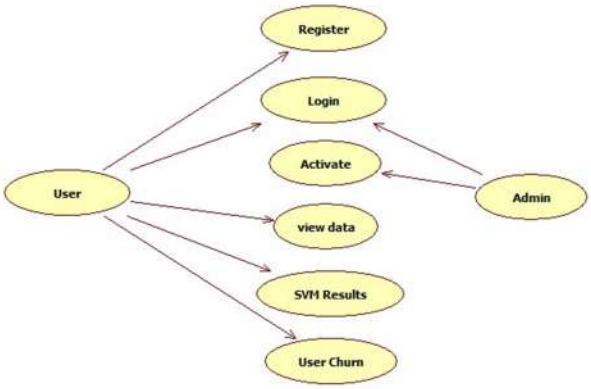
Be independent of particular programming languages and development process.

Provide a formal basis for understanding the modelling language.

Encourage the growth of OO tools market.

Support higher level development concepts such as collaborations, frameworks, patterns and components.

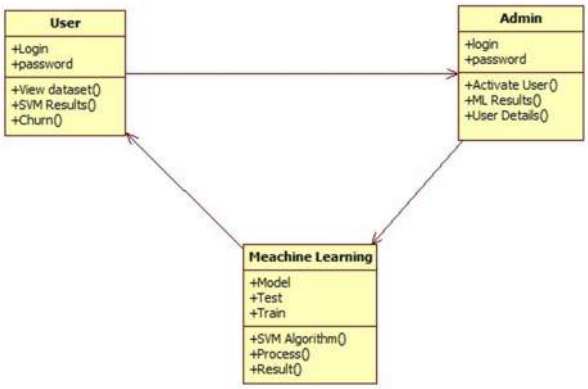
Integrate best practices.



A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. [14]

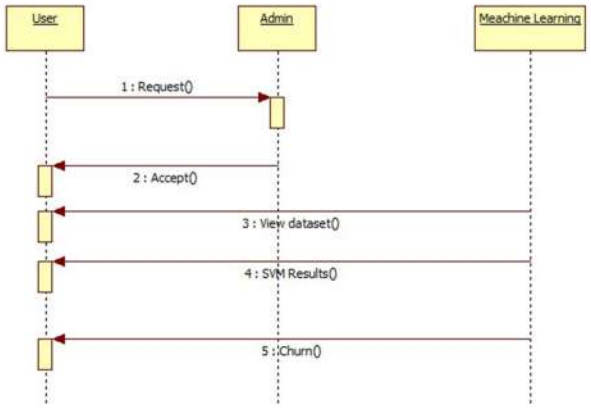
Class diagram:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



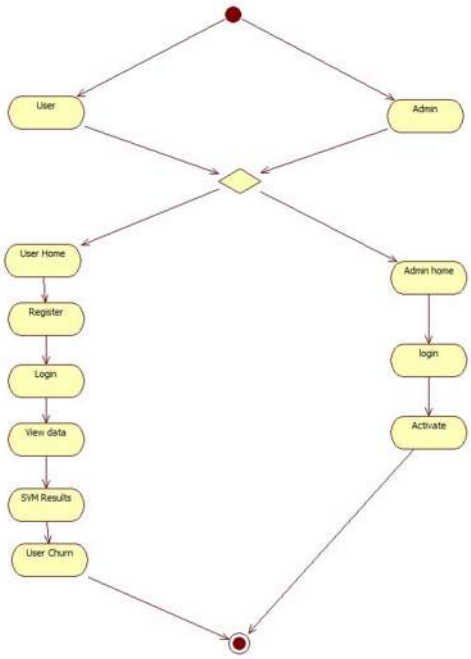
Sequence diagram:

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams



Activity Diagram:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



VI.Modules Description

MODULES:

- User
- Admin

- Data Pre-processing
- Machine Learning

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the user. Once admin activated the user then user can login into our system. User can upload the dataset based on our dataset column matched. For algorithm execution data must be in float format. Here we took Three Customer Behaviour dataset for testing purpose. User can also add the new data for existing dataset based on our Django application. User can click the Classification in the web page so that the data calculated Accuracy and F1-Score, Recall, Precision based on the algorithms. User can click Prediction in the web page so that user can write the review after predict the review that will display results depends upon review like positive, negative or neutral.[15]

Admin:

Admin can login with his login details. Admin can activate the registered users. Once he activate then only the user can login into our system. Admin can view the overall data in the browser. Admin can click the Results in the web page so calculated Accuracy and F1-Score, Precision, Recall based on the algorithms is displayed. All algorithms execution complete then admin can see the overall accuracy in web page.

Data Preprocessing:

A dataset can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc.[16] Features are often called as variables, characteristics, fields, attributes, or dimensions. The data preprocessing in this forecast uses techniques like removal of noise in the data, the expulsion of missing information, modifying default values if relevant and grouping of attributes for prediction at various levels.

Machine learning:

Based on the split criterion, the cleansed data is split into 60% training and 40% test, then the dataset is subjected to four machine learning classifiers such as Support Vector Machine (SVM). The accuracy, Precision, Recall, F1-Score of the classifiers was calculated and displayed in my results. The classifier which bags up the highest accuracy could be determined as the best classifier.[17,18]

VII. Conclusion

Integrating machine learning techniques with diverse, real-time data sources significantly enhances the accuracy, reliability, and adaptability of weather forecasts. By leveraging advanced algorithms and comprehensive data, the proposed system addresses the limitations of traditional meteorological models and offers precise, user-friendly predictions. This innovation is invaluable for sectors like agriculture, transportation, and disaster management, providing critical insights for better decision-making and resource management.

VIII. References

[1] J. C. Villarreal Guerra, Z. Khanam, S. Ehsan, R. Stolkin, and K. McDonald-Maier, "Weather Classification: A new multi-class dataset, data augmentation approach and comprehensive evaluations of Convolutional Neural Networks," 2018 NASA/ESA Conference on Adaptive Hardware and Systems, AHS 2018, pp. 305–310, Nov. 2018, doi: 10.1109/AHS.2018.8541482.

[2] R. G. Tiwari, S. K. Yadav, A. Misra, and A. Sharma, "Classification of Swarm Collective Motion Using Machine Learning," Smart Innovation, Systems and Technologies, vol. 316, pp. 173–181, 2023, doi: 10.1007/978-981-19-5403-0_14/COVER.

[3] R. G. Tiwari, A. K. Agarwal, R. K. Jindal, and A. Singh, "Experimental Evaluation of Boosting Algorithms for Fuel Flame Extinguishment with Acoustic Wave," 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pp. 413–418, Nov. 2022, doi: 10.1109/3ICT56508.2022.9990779.

[4] W. Wang and D. Sun, "The improved AdaBoost algorithms for imbalanced data classification," Inf Sci (N Y), vol. 563, pp. 358– 374, Jul. 2021, doi: 10.1016/J.INS.2021.03.042.

[5] P. Bahad and P. Saxena, "Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics," pp. 235–244, 2020, doi: 10.1007/978-981-15-0633-8_22.

[6] R. Mitchell, A. Adinets, T. Rao, and E. Frank, "XGBoost: Scalable GPU Accelerated Learning," Jun. 2018, doi: 10.48550/arxiv.1806.11248.

[7] V. Gautam et al., "A Transfer Learning-Based Artificial Intelligence Model for Leaf Disease Assessment," Sustainability 2022, Vol. 14, Page 13610, vol. 14, no. 20, p. 13610, Oct. 2022, doi: 10.3390/SU142013610.

[8] Q. A. Al-Haija, M. A. Smadi, and S. Zein-Sabatto, "Multi-Class Weather Classification Using ResNet-18 CNN for Autonomous IoT and CPS Applications," Proceedings - 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020, pp. 1586–1591, Dec. 2020, doi: 10.1109/CSCI51800.2020.00293.

[9] S. Scher and G. Messori, "Predicting weather forecast uncertainty with machine learning," Quarterly Journal of the Royal Meteorological Society, vol. 144, no. 717, pp. 2830–2841, Oct. 2018, doi: 10.1002/QJ.3410.

[10] D. Markovics and M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," Renewable and Sustainable Energy Reviews, vol. 161, p. 112364, Jun. 2022, doi: 10.1016/J.RSER.2022.112364.

[11] A. H. M. Jakaria, M. M. Hossain, and M. A. Rahman, "Smart Weather Forecasting Using Machine Learning:A Case Study in Tennessee," Aug. 2020, doi: 10.1145/nnnnnnn.nnnnnnn.

[12] "WEATHER PREDICTION | Kaggle." https://www.kaggle.com/datasets/ananthr1/weather-prediction (accessed Apr. 18, 2023).

[13] V. Rattan, R. Mittal, J. Singh, and V. Malik, "Analyzing the application of SMOTE on machine learning classifiers," 2021 International Conference on Emerging Smart Computing and Informatics, ESCI 2021, pp. 692–695, Mar. 2021, doi: 10.1109/ESCI50559.2021.9396962.

[14] R. G. Tiwari, A. K. Agarwal, R. K. Kaushal, and N. Kumar, "Prophetic Analysis of Bitcoin price using Machine Learning Approaches," in Proceedings of IEEE International Conference on Signal Processing,Computing and Control, 2021. doi: 10.1109/ISPCC53510.2021.9609419.

[15] J. S. Cavenaugh, "Bootstrap Cross-validation Improves Model Selection in Pharmacometrics," Stat Biopharm Res, vol. 14, no. 2, pp. 168–203, 2020, doi: 10.1080/19466315.2020.1828159.

[16] N. K. Trivedi, V. Gautam, H. Sharma, A. Anand, and S. Agarwal, "Diabetes Prediction using Different Machine Learning Techniques," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022, pp. 2173–2177, 2022, doi: 10.1109/ICACITE53722.2022.9823640.

[17] M. Vuk and T. Curk, "ROC curve, lift chart and calibration plot," Advances in Methodology and Statistics, vol. 3, no. 1, pp. 89–108– 89–108, Jan. 2006, doi: 10.51936/NOQF3710.

[18] N. Ujjwal, A. Singh, A. K. Jain, and R. G. Tiwari, "Exploiting Machine Learning for Lumpy Skin Disease Occurrence Detection," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1–6, Oct. 2022, doi: 10.1109/ICRITO56286.2022.9964656