IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 4, April 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204201 |

Analytical Prediction of Rainfall using Machine Learning

Dr.Sunil B.Wankhade[1], Rutuja Chavarkar[2], Nandini Yadav[3], Sanjeev Yadav[4],Shubham Sah[5].

Head of Department, Department of Information Technology, Rajiv Gandhi Institute of Technology (Mumbai University), Mumbai, Maharashtra, India[1]

U.G. Student, Department of Information Technology, Rajiv Gandhi Institute of Technology (Mumbai University), Mumbai, Maharashtra, India [2,3,4,5]

ABSTRACT: Rainfall prediction is the process to forecast whether it will rain or not affecting the weather in a specific region. Accurate rainfall prediction is essential for many applications, including agriculture, water resource management, and prevention against floods.Machine learning models have become increasingly popular for analytical prediction of rainfall due to their ability to identify patterns and relationships in large datasets. The process involves data collection, cleaning, and preprocessing, feature engineering, model selection, training, evaluation, and deployment..

The heart is the fundamental and most important part of the human body. The life of a human beingmostly depends on the heart. The heart is a very complex part of the human body; hence, understanding it precisely is avery big deal for doctors as well as researchers. Sometimes it gets very difficult for doctors to understand the problemsrelated to the heart of the patient, resulting in the criticality of the patient's health. In Asia, 18.9% of the deaths recordedwere due to cardiac arrest, of which 20% were male and 17.8% were female. One in every five deaths nowadays iscaused by heart disease. Hence, we can understand the seriousness of cardiological diseases. That's why, to overcomethese problems, diseases related to the heart can be predicted by using artificial intelligence, deep learning, machinelearning, and neural networks. It invites automation in the health industry for predicting patients' health and also helpsdoctors monitor the issues in a more precise and easy way. So, we are building a system for predicting heart diseaseswith the help of machine learning and deep learning. There are various algorithms for building such models, such asKNN (K-NearestNeighbors), DecisionTree,SVM algorithm, RandomForest,K-means,Logistic Regression, andmany more. But here we are using KNN (K-nearest neighbors) for building the model for heart disease prediction. Ourgoal is to improve the model's accuracy and provide a very precise and correct outcome of whether the person will haveheart disease or not by using inputs from their daily life habits. We have used a built-in dataset from Kaggle. It consistsof around 18 attributes such as gender, BMI (body mass index), blood pressure, smoker, drinker, etc. This paperproposes a model for heart disease prediction using KNN (K-Nearest Neighbors) algorithm and deployed it in a webapplication.

KEYWORDS:Heartdisease,Prediction,Machinelearning,K-nearestneighbors

I. INTRODUCTION

Rainfall is the vital factor playing crucial role in agriculture, water resource management, prevention of floods, water supply and climate modelling. Predicting whether it will rain tomorrow or not is the core objective of rainfall prediction model. Machine Learning models have been the emerging approach for the effective and efficient way towards the rainfall prediction. In the recent years there has been advancement in the ML models.To develop an ML-based rainfall prediction model, historical data on rainfall and other relevant meteorological and environmental factors, such as temperature, humidity, wind speed, and pressure, are collected and pre-processed to ensure

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204201 |

consistency and accuracy.Different ML models such as Cataboost, Random Forest, Logistic Regression, KNN and SVRare used. Based on which it is checked which model is better for the rainfall prediction. These ML models can analyze large datasets and identify complex patterns and relationships between the input features and rainfall. The data is collected fromthe previous dataset. Different meteorological and historical factorsare considered for predicting rainfall patterns. Model is then trained and evaluated. To make the predictions of future rainfall patterns based on input features we deploy the model Continuously there is the need to monitor the performance to improve its accuracy.

II. RELATED WORK

Several studies have been conducted using machine learning data to predict heart disease. Researchers have used different data mining techniques to achieve different levels of accuracy. Avinash Goland and colleagues studied heart disease classification using several ML algorithms, including Decision Tree, ANN, and K-Means. They found that Decision Tree has the highest accuracy and suggested that it can be further optimized by combining the technique and parameter tuning. T Nagamani and his team proposed a system that uses data mining techniques and the MapReduce algorithm to achieve greater accuracy compared to artificial fuzzy neural networks. Fahd Saleh Alotaibi designed an ML model that compared five different algorithms and found that Decision Tree was the most accurate. Anjan Nikhil Repaka and colleagues proposed a system using NB techniques for classification and AES algorithms for secure data transmission. Teresa Prince R. conducted research on various classification algorithms including Naive Bayes, KNN, Decision Tree and Neural Network. Finally, Nagaraj M. Lutimath and his team performed heart disease prediction using Naive Bayes classifier and SVM and found that SVM outperformed Naive Bayes in terms of accuracy. To determine the best algorithm for predicting heart disease, different classification algorithms were analyzed based on their accuracy, precision, recall, and F-measure scores. We analyzed decision tree, random forest, logistic regression, k-nearest neighbor, support vector, and Naive Bayes classification algorithms based on accuracy, precision, recall, and F-Measure scores and determined the best classification algorithm to use. prediction of heart disease.

III. METHODOLOGY

The initial and foremost step in the process is to acquire data. In our case, the data was gathered from Kaggle. The next step involves transforming the data into a format suitable for experimentation. This includes preprocessing tasks such as replacing null and missing values with the mean of the column, detecting and removing outliers, and visualizing data using box plots. Subsequently, data analysis is carried out to observe variations in rainfall patterns, derive conclusions, and determine the correlation between different parameters. The data is visualized using bar graphs, heat maps, histograms, and other tools for easy understanding. Next, we attempt to forecast average rainfall by dividing the data into training and testing datasets. Various statistical and machine learning methods, including SVR, KNN, logistic regression, CatBoost, and Random Forest, are employed for prediction and analysis. Finally, we aim to create a user-friendly and interactive webpage that allows users to input different parameters and obtain predictions. Fig.1 depicts the Methodology:

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204201 |

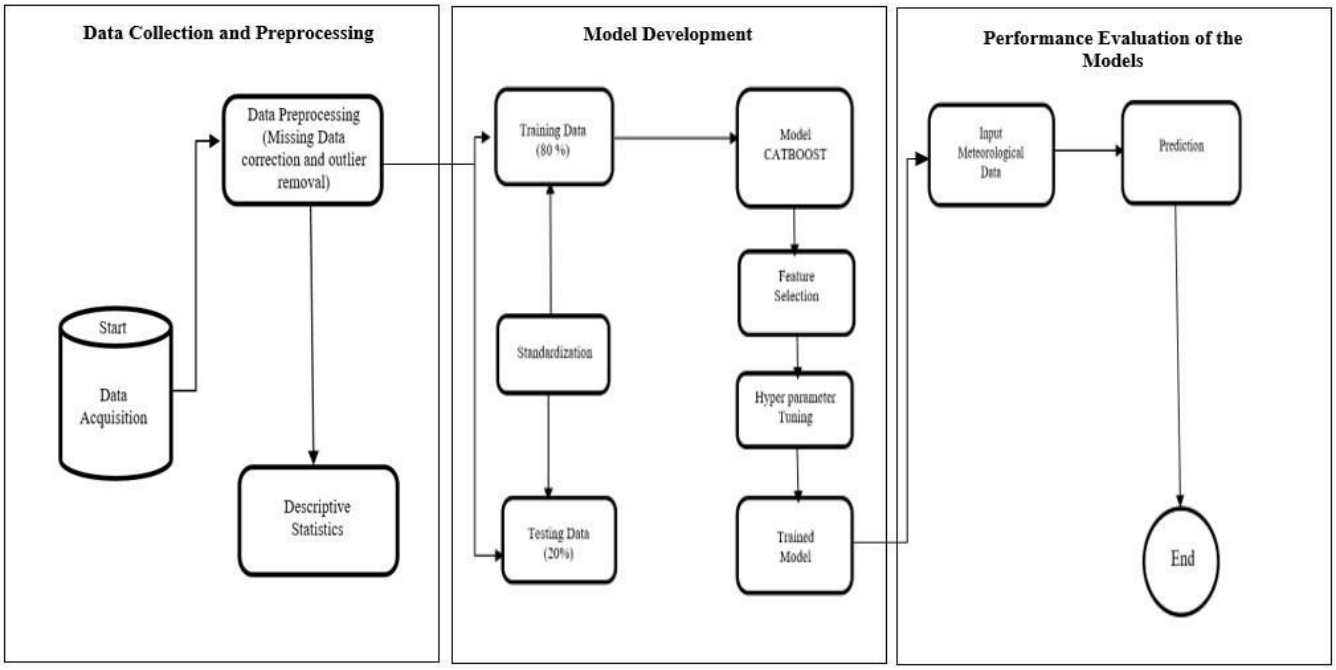

Fig 1: Methodology

IV. EXPERIMENTALRESULTS

Data collection:-

The authors collaboratively agreed that to achieve accurate predictions, the dataset they will be working with should contain a sufficient number of rows or tuples and should have relevant attributes that can be effectively used to determine their correlation for prediction purposes. After comparing various datasets, they concluded that the one available on Kaggle meets these criteria. It consists of 145,460 tuples of data with 23 distinct attributes, out of which 16 contain numerical values and the remaining 7 columns have categorical values. Additionally, the 16 numerical values can be classified into 2 discrete and 14 continuous values.

Data pre-processing:-

Upon successful loading of the dataset, our team checked for null values present in the data, as it is a common issue with most datasets. Out of the 23 attributes in the dataset, 21 had null values. Four of these attributes had more than 35% of their rows missing data, which made replacing them with only one value (such as mean or median) difficult as it would significantly affect the variance and visualization. To handle this, we defined a function that randomly fills a value in place of the null or NaN. For the remaining attributes with continuous numerical values, we replaced null values with the median value of the column. Categorical attributes with null values were replaced with distinct dummy values. Next, we converted all categorical values into numerical values for prediction purposes. For instance, we replaced 'yes' and 'no' with one and zero, respectively, while giving distinct representative numbers to attributes like location and wind direction. Following this, we checked for outliers in the attributes using box plots to find inconsistencies and errors that could affect statistical results. To remove outliers, we defined a range using the interquartile range, and any value within this range was not considered an outlier. If an attribute value was not within this range, we replaced it with the lowest or highest range value accordingly. Once the pre-processing stage was completed, we drew visualizations to obtain further insights into the data.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204201 |

Model Validation:-

A crucial step in model development is model validation, which involves comparing the performance of a trained model to a test set of similar data that was not used for training. The purpose of model validation is to identify the most effective model that produces the best performance. To begin with, we performed a train-test split, where 80% of the dataset was used for training and the remaining 20% for testing. This is done to ensure that the model is evaluated with new data and not the same data it was trained on. In addition, we addressed the issue of imbalanced data, which occurs when one class has a disproportionately high number of observations compared to the other. To tackle this, we used resampling techniques, specifically the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples for the minority class. After resampling the data, we applied seven different algorithms to determine which would provide the best results. These algorithms included CatBoost Classifier, Random Forest Classifier, Logistic Regression, Gaussian Naïve Bayes, K Neighbours Classifier, XGBoost Classifier, and Support Vector Classifier. We compared each algorithm's accuracy, RoC curve, and AuC score and concluded that the CatBoost Classifier was the most suitable algorithm for prediction.

Feature Selection:-

During the process of Data cleaning, feature selection is a crucial step. Selecting relevant features not only eliminates un- important ones but also enhances the performance of the model. In our project, we have utilized the mutual_info_classif method. This method is based on mutual information (entropy) gain and applies to classification problems. This method improves the accuracy of the model due to its univariate filtering. In univariate methods, features are calculated separately, which can result in the selection of suboptimal features when they are grouped. However, univariate filtering methods are relatively quick and can be used for screening, leading to better performance and less training. We calculated the mutual information values of every attribute and sorted them in decreasing order (higher values indicate more dependency).

Interface Building:-

To enhance the user experience and make it easy to understand the significance of building prediction models, we have developed a user-friendly web interface using Flask. This web framework is a Python module that simplifies the process of building web applications. One of the key features of our interface is the dataset analysis, which was carried out using PowerBI. This business intelligence tool enables users to visualize and analyze raw data and present it in the form of actionable insights. Our analysis involved the use of line charts, heatmaps, bar graphs, decomposition trees, area charts, and radial charts, as well as slicers for periodic or time-oriented analysis. Figures 4 and 5 display our PowerBI analysis dashboard, which contains all of the charts and plots. We have created two dedicated predictor pages that enable users to make predictions using their own data. On one page, we make predictions using a limited set of attributes that we extracted using feature selection, while on the other page, users can enter data for all 22 entities to make predictions.

V. FUTURESCOPE

There are several potential ideas and changes that could be implemented to enhance the effectiveness of a rainfall prediction system using machine learning. One potential idea is to integrate the system with a flood prediction and alert model specific to rivers, which could provide additional warnings and alerts to people living in areas prone to flooding. Another idea is to develop a separate prediction model to forecast whether it will rain on the following day, using weather data specific to major cities. To improve the accuracy and usefulness of the system, additional features could also be added, such as real-time updates and major alerts across the country, which could help people take necessary precautions in the event of heavy rainfall or flooding. Additionally, the impact of other meteorological parameters on rainfall prediction could be studied to further refine the system's accuracy. To extend the system's capabilities further, it could be adapted to forecast hourly rainfall data, using a range of different parameters. This could be done using a

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2023 ||

| DOI:10.15680/IJIRSET.2023.1204201 |

larger data set for a specific region or area, and utilizing big data frameworks to improve computation speed and accuracy.

VI. CONCLUSION

Rainfall prediction plays a vital role in agriculture production, as the growth and yield of crops are directly affected by the amount of rainfall. Therefore, accurate rainfall prediction is essential to guide farmers in their agricultural practices. In this study, we propose a gradient boost approach known as CatBoost for predicting rainfall in a selected location in Australia. Our proposed method uses a decision tree algorithm (CatBoost model) to forecast rainfall for the Australian dataset and provides improved results in terms of accuracy and prediction. The results of the study demonstrate that the CatBoost model performs better in months with high annual averages when compared to alternative approaches. However, future research can be focused on combining the diversities of different models to further improve the accuracy and effectiveness of the rainfall prediction system. Additionally, more locations and datasets can be incorporated into the system to expand its scope and usefulness. In this research paper, we have extensively explored and applied several preprocessing steps and implemented various machine learning algorithms to predict rainfall in Australian regions. Our study involved a comparative analysis of the overall performance of seven different algorithms to select the most accurate and effective one. Through this analysis, we found that the CatBoost classifier was the most efficient machine learning model, with an accuracy of 91.12% and an AuC score of 0.89.

Additionally, we have developed a user-friendly front-end system that allows users to input different parameters and observe how they affect the model's prediction accuracy for the next day's rainfall. The system includes an interactive dashboard developed using PowerBI that provides a better understanding of the data and the built system.Our study concluded that Australia's rainfall pattern is irregular and uncertain. Through feature extraction and data visualization techniques, we were able to identify and figure out the relationships among the data and the factors that significantly affect the amount of rainfall received in different regions of Australia.

REFERENCES

[1] Kala, A., & Vaidyanathan, S. G. (2018, July). Prediction of rainfall using artificial neural network. 2018 International Conference on Inventive Research in Computing Applications (ICIRCA).
[2] Oswal, N. (2021). Predicting Rainfall using Machine Learning Techniques. Institute of Electrical and Electronics Engineers (IEEE).
[3] Wicaksono Putra, A. B., Malani, R., Suprapty, B., & Onnilita Gaffar, A. F. (2020, July). A deep auto encoder semi convolution neural network for yearly rainfall prediction. 2020 International Seminar on Intelligent Technology and Its Applications (ISITIA).
[4] Ramya Bhaskar Sundaram, "An End-to-End Guide to Understand the Math behind XGBoost", 2018.Available:https://www.analyticsvidhya.com/blog/ 2018/09/an-endto-end-guide-to-understand-the-mathbehind-xgboost/[Submitted on 6 September 2018]
[5] Anamika Jha "CatBoost – A new game of Machine Learning:", 2020. Available:https://affine.ai/catboost-anew-game-of-machinelearning/ [Submitted on 21 October 2020]
[6] Aman Kharwal, "Rainfall Prediction with Machine Learning",2020.Available:https://thecleverprogrammer.com/2020/09/11/rainfallprediction-with-machinelearning/ [Submitted on 11 September 2020]
[7] Vishal Morde. "XGBoost Algorithm: Long May SheReign!",2019.Available:https://towardsdatascience.com/httpsmedium-com-vishalmorde-xgboost-algorithm-longshe-may-rein-edd9f99be63d [Submitted on 8 April 2019]
[8] Marchuk, G. *Numerical Methods in Weather Prediction*; Elsevier: Amsterdam, the Netherlands, 2012.
[9] Imam Cholissodin, S. Prediction of rainfall using improved deep learning with particle swarm optimization. *Telkomnika* **2020**, *18*, 2498–2504.
[10] Ravuri, S.; Lenc, K.; Willson, M.; Kangin, D.; Lam, R.; Mirowski, M. Skillful Precipitation Nowcasting using Deep Generative Models of Radar. *Nature* **2021**, *597*, 672–677.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details