Development of Multiple Combined Regression Methods for Rainfall Measurement

Nusrat Jahan Prottasha
Daffodil International University Dhaka 1207, Bangladesh

Md. Jashim Uddin
Noakhali Science and Technology University, 3814, Dhaka

Md. Kowsher
Stevens Institute of Technology, Hoboken, NJ 07030 USA

Rokeya Khatun Shorna
Daffodil International University, 1207, Dhaka

Niaz Al Murshed
Jahangirnagar University, 1342, Dhaka

Boktiar Ahmed Bappy
Jhenaidah polytechnic institute, 7300, Dhaka
Corresponding author: Nusrat Jahan Prottasha, Email: jahannusratprotta@gmail.com

Rainfall forecast is imperative as overwhelming precipitation can lead to numerous catastrophes. The prediction makes a difference for individuals to require preventive measures. In addition, the expectation ought to be precise. Most of the nations in the world is an agricultural nation and most of the economy of any nation depends upon agriculture. Rain plays an imperative part in agribusiness so the early expectation of rainfall plays a vital part within the economy of any agricultural.

2021. In Raju Pal & Praveen K. Shukla (eds.), *SCRS Conference Proceedings on Intelligent Systems*, 79–95. Computing & Intelligent Systems, SCRS, India.
https://doi.org/10.52458/978-93-91842-08-6-7

Overwhelming precipitation may well be a major disadvantage. It's a cause for natural disasters like floods and drought that unit of measurement experienced by people over the world each year. Rainfall forecast has been one of the foremost challenging issues around the world in the final year. There are so many techniques that have been invented for predicting rainfall but most of them are classification, clustering techniques. Predicting the quantity of rain prediction is crucial for countries' people. In our paperwork, we have proposed some regression analysis techniques which can be utilized for predicting the quantity of rainfall (The amount of rainfall recorded for the day in mm) based on some historical weather conditions dataset. we have applied 10 supervised regressors (Machine Learning Model) and some preprocessing methodology to the dataset. We have also analyzed the result and compared them using various statistical parameters among these trained models to find the bestperformed model. Using this model for predicting the quantity of rainfall in some different places. Finally, the Random Forest regressor has predicted the best r2 score of 0.869904217, and the mean absolute error is 0.194459262, mean squared error is 0.126358647 and the root mean squared error is 0.355469615.

Keywords: Rainfall, Supervised Learning, Regression, Random Forest Tree, AdaBoost Regressor, Gradient Boosting Regressor, XGBoos

1 Introduction

This research paper proposed a scientific method to predict rainfall quantity based on some different weather conditions considering preceding weather records and present weather situations using some regression analysis techniques [1]. Rainfall determining is exceptionally vital since overwhelming and irregular rainfall can have numerous impacts on many other things like annihilation of riverbank, crops, agriculture, and farms. One of the very deleterious departures is flooding due to the over rain. According to Wikipedia in late summer 2002, enormous storm downpours driven to gigantic flooding in eastern India, Nepal, and Bangladesh, killing over 500 individuals and clearing out millions of houses [2]. Each year in Bangladesh approximately 26,000 square kilometers (10,000 sq mi) (around 18% of the country) is flooded, killing over 5,000 individuals and wrecking more than 7 million homes. On the other hand, Western Sydney is now the "greatest concern" from the worst floods in decades to have ravaged eastern Australia. Rodda et al. [3] presented a very rational method of the rainfall measurement problem. The application of science and innovation that predicts the state of the environment at any given specific period is known as climate determining or weather forecasting. There are many distinctive strategies for climate estimate and weather forecasting. But rainfall prediction is rare. Some of the research has shown some classification method to predict whether it would be rain tomorrow

or not. But instead of a classification method for predicting rain, we need to the quantity of the rainfall in a particular place. There is numerous equipment implement for foreseeing rainfall by utilizing the climate conditions like temperature, humidity, weight. These conventional strategies cannot work productively so by utilizing machine learning procedures. we can create an exact comes about rain forecast. Ready to fair do it by having the historical information investigation of rainfall and can anticipate the precipitation for future seasons.

In our paper, we presented some predictive regression analysis techniques to quantify rainfall quantity at a place. Here we used more than 10 years of historical data to train our model. The dataset contains various weather conditions of different places. This method can be utilized to predict the rainfall (The amount of rainfall recorded for the day in mm) and avoid the annihilation caused by it to life, agriculture, farm, and property. If we can quantify the rainfall most people can make some decisions before overwhelmed rain-affected. The contributions of this work are summarised as:

- We have assessed a pipeline of making choices for evaluating the finest reasonable rain prediction.
- We have utilized 10 supervised regressors (Machine Learning Model). Because different regressors give us different results. So, it's essential to find out the right model according to the requirements.
- We have discussed a big comparison among all trained models to figure out the best performer.

The paper is organized as takes after: Section II clarifies the related work of different classification strategies for the forecast of rain classification. Section-III depicts the technique and materials utilized. Section-IV depicts the experimental analysis including performance and result. Section V talks about the conclusion of this research work where section VI described about the plan of future.

2 Related Works

In this paper, through a systematic investigation Rodda et al. [3] have presented the rainfall measurement problem, they claim there's an orderly mistake in the estimation of precipitation made in an ordinary way, a mistake which may influence any gauges utilizing these estimations. Besides Prabakaran et al. [4] proposed a method that speaks to a numerical strategy called Linear Regression

to anticipate the rainfall in different areas in southern states of India. To improvement Wang et al. [5] showed a case study they proposed an application of generalized regression neural network (GRNN) model to anticipate yearly precipitation in Zhengzhou . On the other hand, Sethi et al. [6] presented an exploiting data mining technique for the early prediction of rainfall called multiple linear regression (MLR). Sunyoung Lee et al. [7] presented a divide and conquer approach to predict the rainfall based on the locational information only. Also, M Adil et al. [8] developed the Clusterwise Linear Regression (CLR) technique for the prediction of monthly rainfall . In addition, Mohammed Moulana et al. [9] represented machine learning techniques to precipitation prediction the purpose of this project is to offer non-experts simple get to the methods, approaches utilized within the division of precipitation forecast and give a comparative think about among the different machine learning methods. Asha et al. [10] proposed a mutual neural classification model for predicting rainfall. Sakthivel et al. [11] described neural networks and the rapid miner-based rain prediction system. Naidu et al. [12] presented the changes in rainfall patterns in numerous agro-climatic zones using machine learning approaches. Besides, Dinh et al. [13] utilized an LSHADE-PWI-SVM method for the integration of machine learning classifiers conjointly metaheuristic optimization . On the other hand, Malathi et al. [14] showed a Information Gain based Feature Selection Method for Weather Dataset for the prediction of rainfall. Also, SamsiahSani et al. [15] evaluated many machine learning classifiers based on Malaysian data for rainfall prediction. Ahijevych et al. [16] presented a random forest (RF) that is utilized to produce 2-h figures of the probability for the start of mesoscale convective frameworks (MCS-I). Allen et al. [17] performed property and agribusiness, as well as handfuls of fatalities and Wonders related to extreme electrical storms. Brooks et al. [18] displayed the current dissemination of serious rainstorms as a work of large-scale natural conditions. Gentine et al. [19] representing uncertain sodden convection in coarse-scale climate models remains one of the most bottlenecks of current climate recreations. McPhaden et al. [20] described the participation of the pivotal for agriculture-dependent. Hazell et al. [21] represented to reduce the risk of life and also maintain the agriculture farms in a better way Then, Mollinga et al. [22] elucidates farmers to take early measurements of floods, and manage the water resources properly. Shah et al. [23] discussed to related this task to predict rain.

3 Methodology

To perform the complete technique, we assume the four significant steps such as data collection, data pre-processing, training model using 10 supervised regressors, and execution examination. Within the information collection step, we have used a dataset [1] from the Kaggle platform which has been split into two parts such as the training part and validation part. Here we have utilized one of the validation parts as the testing data to evaluate the models' performance. Each row has various weights for decision making to suggest the sensible best rain prediction. Afterward, gathering all raw data, firstly we would be made ready for the training model with the help of data pre-processing techniques and this has been used for outliers free and more rigid. It also assists to increase the performance of the models. As a result, we have applied six pre-processing methods such as cleaning data, missing value check, handling the categorical data, handling outliers, handling outliers, feature selection. Next, to establish supervised regressors models, we utilized the regressors such as Linear Regression, Ridge Regression, Polynomial Regression, and Lasso Regression. From all the training methods we have used a total of 10 regressors so that we can compare the performance and figure out the best model. Most of the regressors come up with a good performance. We have described the whole methodology in Figure 1.

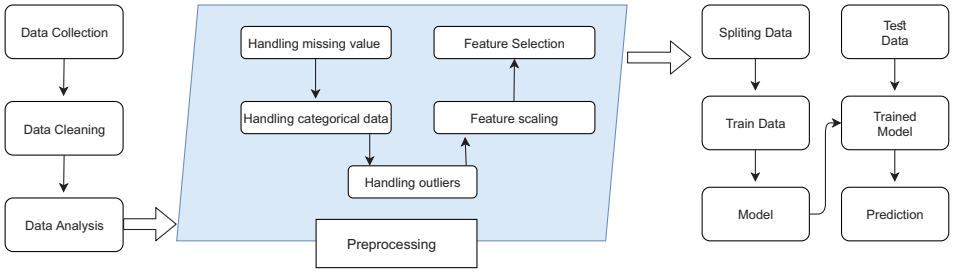

Figure 1: The whole methodology of rainfall prediction including all important steps such as data collection, necessary preprocessing, and training model with performance prediction

3.1 Introduction dataset

Kindly This dataset contains about 10 years of daily weather observations from many locations. We have collected this dataset from Kaggle. It is having 23 di-

[1]https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

Table 1: Considering feature's description of dataset

Feature name	Description
Location	The common title of the area of the climate station.
MinTemp	The least temperature in degrees centigrade.
MaxTemp	The most extreme temperature in degrees centigrade.
Rainfall	The sum of precipitation recorded for the day in millimeters.
WindGustDir	The heading of the most grounded wind blast within 24 h to midnight.
WindGustSpeed	The speed (in kilometers per hour) of the strongest wind blast within 24 h to midnight.
WindDir9am	The course of the wind blast at 9 a.m.
WindSpeed9am	Wind speed (km/hr) found the middle value of over 10 minutes sometime recently 9 am.
WindSpeed3pm	Wind speed (in kilometers per hour) found the middle value of over 10 min sometime recently 3 p.m.
Humidity9am	Relative humidity at 9 am.
Humidity3pm	Relative humidity at 3 pm.
Pressure 9am	Climatic weight (hPa) was decreased to cruel ocean level at 9 a.m.
Temp3pm	Temperature (degrees C) at 3 p.m.
Rain Today	Numbers 1 on the off chance that precipitation (in millimeters) within the 24 h to 9 a.m. surpasses 1 mm, something else 0.

verse observation features of weather condition like 'Location', 'Min Temp', 'Max-Temp', 'Rainfall', 'Evaporation', 'Sunshine', 'Wind Gust Dir', 'Wind Gust Speed', 'Wind Dir 9am', 'Wind Dir 3pm', 'Wind Speed 9am', 'Wind Speed 3pm', 'Humidity 9am', 'Humidity 3pm', 'Pressure 9 am', 'Pressure 3pm', 'Cloud 9am', 'Cloud 3pm', 'Temp 9am', 'Temp 3pm', 'Rain Today'. Here,in the table 1 the description of the data-set has been illustrated.

3.2 Pre-processing

In machine learning, the data preprocessing is within the framework of exchanging or encoding the crude information in a stage where calculations can be effectively implemented to prepare. We ought to preprocess the information concurring to create it fit for the machine learning model. Well-processed data gives high accuracy and makes the model more solid. Here, we have utilized a few stages of preprocessing strategies, which have been outlined in Figure-2:

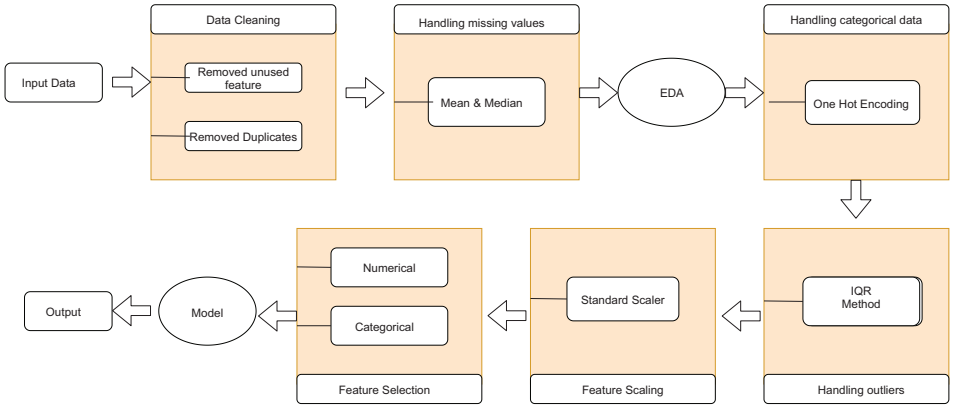

Figure 2: Data Pre-Processing

In our dataset, there are parcels of unused, null, and duplicate values. For this reason, we took some steps to handle these issues. such as,

- Erased duplicate row and column: we discover that numerous information points are repeated in row and column sections. Therefore, we expelled all the duplicate information.
- Erased the row and column, which shows up more than 50% of the null value. Cleaning data occurs when 50% of information comes to the null value. At that point, we have chosen to evacuate the whole rows and columns.

For the most part, missing value is characterized as the value which was not put away within the sample. The missing value may be a common occasion in information. On the other hand, most prescient modeling strategies can't handle any missing value. Thus, this issue must be unraveled before modeling. In some cases, median, mean, mode strategies are utilized to overhaul a missing value. In any case, the foremost direct method for managing the missing value is the mean, median, mode strategy. Here we have utilized this mean, median & mode strategy for managing missing data

- Handled categorical features: Categorical data could be a subjective include whose values are taken on the value of labels. So, we ought to encode this sort of information into numbers so that the machine learning model can execute scientific operations on it. In our dataset, there exist a few categorical features. We have utilized one-hot encoding, one of the foremost prevalent encoding algorithms, to encode the categorical values into numbers. It is the foremost common approach, and it works well unless any categorical variable takes a large number of diverse values. After this encoding, a double matrix is shaped where 1 indicates the presence of any value and 0 indicates the absence of the value.

- Inside our dataset, there were a lot of outliers presented: an outlier is a perception point that's removed from other perceptions. An outlier may be due to variations within the estimation or it may appear exploratory mistake the latter are some of the time excluded from the set of information. An issue of outliers can cause, they tend to be unaffected by littler UI changes that do influence a more whimsical standard population. Bulk orders will thrust through littler convenience changes in a way that your average visitor may not. So to handle the outliers we have used the IQR (interquartile range) method, which is an efficient technique.

- Include scaling is one of the significant strategies that are mandatory to standardize the working data's independent features. All things considered, there are different strategies like Min-Max Scaling, Variance Scaling, Standardization, Mean Normalization, and Unit vectors for include scaling. In our work, we have applied standard scaling as a feature scaling procedure. Here, the exchanged every data point in the range of between -1 and 1.

3.3 Training selective models

The linear model [24]performs well in machine learning linearly. We utilized the four regressors as Linear Regression, Ridge Regression, Polynomial Regression, and Lasso Regression. Tree-model [25]algorithms are considered to be one of the leading and most utilized supervised learning methods. In this work, we utilize a decision tree regressor. We utilized ”gini” for the Gini impurity, and the splitter is chosen as ’best’ to select the part at each node. Ensemble methods [26] are procedures that make multiple models and combine them to create moves forward. Here, we utilized four ensemble-based regressors. These are Random Forest, Gradient Boosting, Adaboost, and XGboost. Afterward, we have utilized three neighbors regressors of statistical pattern recognition. This is K- nearest neighbors [27],five nearest is chosen for every iteration. Besides, the Manhattan distance is chosen for all neighbor classifiers. The support vector machine SVM [28] is used mainly for exploring a hyperplane in ddimensional space that notably fits a hyperplane in data points. In the linear SVM, we used hinge as loss function with l2 penalty.

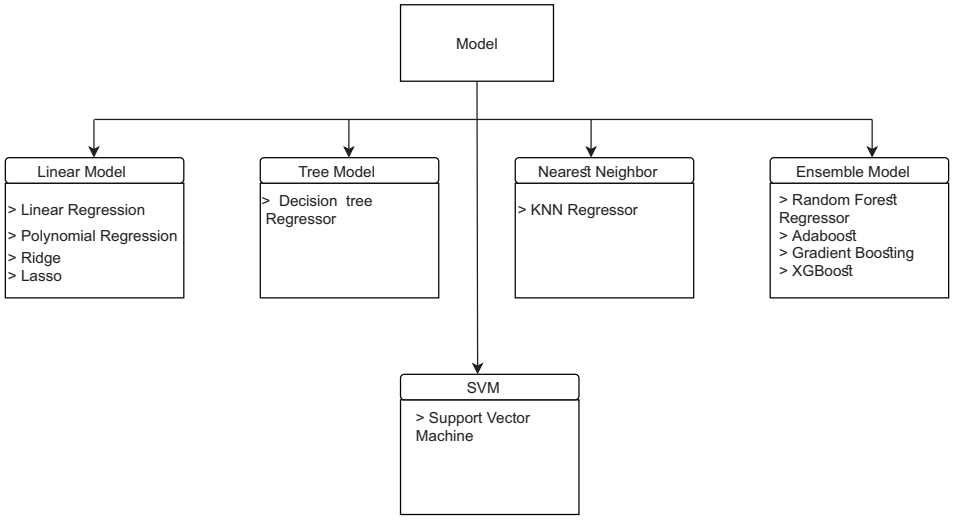

Figure 3: Training Algorithms

4 Experiment

In the advancement of our test from the proposed work, we have to begin with amassed the demonstrate and prepared it. 10 different regressors from supervised learning based on distinctive learning techniques have been executed to anticipate precipitation's most pertinent mode. This area depicted distinctive test errands for the execution investigation and assessment and compared all calculations. Then, we have outlined the test setup utilized to execute the entire errand and utilized 11 statistical assessment measurements for investigation execution. At long last, we have moreover compared with other works related to this issue concerning the finest form of our work.

4.1 Experiment Setup

we have completed the complete computation in [2]google collab, a python reenactment environment given by Google. This environment comes with parallel computation facilities for quick execution. We have utilized the foremost well-known libraries to create simple and expressive information structures that work well and instinctively quickly. At long last, sklearn library contains specialized machine learning and statistical modeling instruments, counting classification, regression, and clustering calculations for modeling. We have utilized a machine learning system named [3]sci-kit learn to implement the regression algorithm. At long last, we utilized [4]matplotlib and [5]seaborn for information visualization, graphical representation, additionally for information investigation.

4.2 Statistical measurement

R2 score : The R2 score could be a very critical metric that's utilized to assess the performance of a regression based machine learning model. It is articulated as R squared and is additionally known as the coefficient of assurance. It works by measuring the sum of variance within the expectations clarified by the dataset. Basically put, it is the contrast between the tests within the dataset and the expectations made by the demonstrate. As we can see from all models Random Forest regressor achieves the best r2 score which is 0.869904217. The second and third

[2]https://colab.research.google.com/
[3]https://scikit-learn.org
[4]https://matplotlib.org
[5]https://seaborn.pydata.org

positions are achieved by GradientBoostingRegressor and XGBoost which are 0.863496747 and 0.863215393. The condition is shown underneath in condition 1:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{7.1}$$

Mean absolute error: If we consider with respect to error rate then first comes to mean absolute error. In measurements, mean absolute error may be a degree of blunders between combined perceptions communicating the same wonder. Mean Absolute Error (MAE) is another loss function utilized for relapse models. MAE is the entirety of outright contrasts between our target and anticipated factors. So it measures the normal greatness of errors in a set of forecasts, without considering their bearings. Random Forest regressor gets the least mean absolute error rate which is 0.194459262 compare to others. The declaration of the F1 score is displayed in equation 2 :

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right| \tag{7.2}$$

Mean squared error: If we consider with respect to mean squared error, The mean squared error (MSE) tells how near a relapse line is to a set of focuses. It does this by taking the separations from the focuses to the relapse line these separations are the errors and squaring them, we call It mean squared error. From all the models Random forest achieves a minimum mean squared error 0.126358647. The articulation is shown beneath in 3 :

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 \tag{7.3}$$

Root mean squared error: Now if we consider the root mean squared error, Root Mean Square Error (RMSE) means the standard deviation of the residuals which is prediction error. Residuals are a degree of how distant from the relapse line information focuses are RMSE could be a degree of how to spread out these residuals are. Here root mean squared error of Random Forest is 0.355469615 which is less compare to others. The verbalization is shown in 4:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{7.4}$$

By all the statistical performance analysis we can see Random forest is the efficient regressor model and performing well in this use case.

4.3 Result & performance analysis

Table 2: Performance Metrics of different regressors

Model Name	r2 score	MAE	MSE	RMSE
Random Forest	0.869904217	0.19445926	0.126358647	0.355469615
Decision Tree	0.742284572	0.21508858	0.250312287	0.500312189
Linear Regression	0.837495137	0.22694578	0.157836744	0.397286728
KNN Regressor	0.401557082	0.48855924	0.581252029	0.762398865
AdaBoost Regressor	0.786451397	0.37659111	0.207414199	0.455427491
Gradient Boosting Regressor	0.863496747	0.20372662	0.132582057	0.364118191
XGBoost	0.863215393	0.20367076	0.132855329	0.364493249
Ridge Regression	0.837495234	0.157836649	0.132855329	0.397286608
Lasso Regression	-5.91E-05	0.83158029	0.971331339	0.985561434
SVM	0.841801	0.203451	0.130951	0.345151

From Table 2, we showed statistical results and comparisons among all machine learning regressors.For better analysis, we choose some statistical procedures for numerical result computing such as r2 score, mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE). After developing the models and testing all regressors, We can see that the Random Forest has predicted the best accuracy of 0.869904217 among all others, and the mean absolute error is 0.194459262 which is the lowest, mean squared error is 0.126358647 and the root mean squared error is 0.355469615. Considering all errors and accuracy, it took the best place. Secondly, the gradient boosting regressor has gained better accuracy with the second place which is 0.863496747 with the mean absolute error is 0.203726623, mean squared error is 0.132582057 and the root mean squared error is 0.364118191. Thirdly, the XGBoost regressor has acquired better accuracy, which is 0.863215393, along with the mean absolute er-

ror is 0.203670766, mean squared error is 0.132855329 and the root mean squared error is 0.364493249. Also, from the section on linear algorithms, we can figure out that Linear Regression and Ridge Regression showed the almost same accuracy and so on. So in this analysis, we can although Random forest and Gradient Boosting Regressor have acquired almost the same Accuracy but if we consider the evaluation metrics of then so, Random forest has a low error rate compare to Gradient Boosting. So, here we have considered the Random forest approach. Overall all of regressors showed a standard and acceptable performance.

The bar chart is a graph for representing all regressors algorithms with Statistical measurement. The bar can be vertically or horizontally. Here is the bar graph of our selective algorithms, down below.

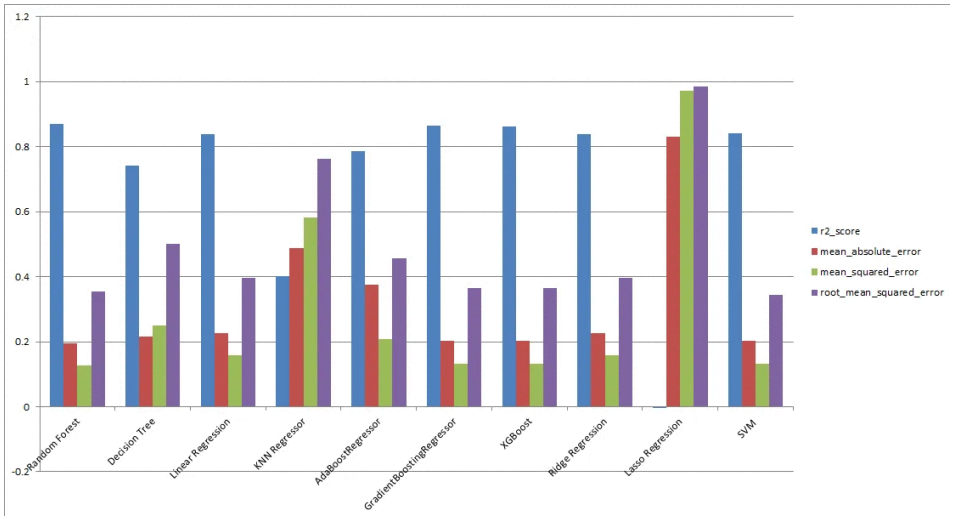

Figure 4: Selective algorithms

5 Conclusion

In this work, we have presented an initial attempt to determine how much rain will come when it's raining time. In the data collection phase, we adopted real data from Australia from the Kaggle platform. The primary purpose of this task is to find out the best regression technique for the prediction of rain. For this reason, we have used a variety of regression analysis techniques that can be utilized for predicting the quantity of rainfall so that anyone can use the best predictive model in real-life applications. To perform this task, we selected five significant

steps, these are data collection, data prepossessing, training model using regression analysis techniques, and performance analysis. In pre-processing part, we have described cleaning data, Missing value check, EDA, Handling outliers, Feature selection, Feature scaling respectively. Besides, we used ten supervised regressors (machine learning models) for predicting rainfall. Among all models the are gives good accuracy in our predicting regression. Here, in the figure 4 the graphical performance including compassion among all trained models has been depicted

6 Future Work

In future work, we will focus on the real-life application of rainfall prediction, so that anyone especially farmer can use it easily and forecast the weather of rain. Also, we have plan to use the neural network based deep hybrid approaches to improve the performance. Undoubtedly, we have plans to evaluate the other country's data for forecasting the rain.

Bibliography

[1] Ortiz-García, E. G., S. Salcedo-Sanz, and C. Casanova-Mateo. Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. Atmospheric research, 139:128–136, 2014.

[2] Ian Tyrrell. *River Dreams: The people and landscape of the Cooks River*. NewSouth, 2018.

[3] John C Rodda. The rainfall measurement problem. *IAHS Publication No*, 78:215–231, 1967.

[4] Gujanatti Rudrappa, Nataraj Vijapur, Rajesh Pattar, Ravi Rathod, Rashmi Kulkarni, Vudu Sree Chandana, and Sateesh N Hosmane. Machine learning models applied for rainfall prediction. *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS*, 11(3):179–187, 2021.

[5] Zhi-liang Wang and Hui-hua Sheng. Rainfall prediction using generalized regression neural network: case study zhengzhou. In *International conference on computational and information sciences*, pages 1265–1268. IEEE, 2010.

[6] Nikhil Sethi and Kanwal Garg. Exploiting data mining technique for rainfall prediction. *International Journal of Computer Science and Information Technologies*, 5(3):3982–3984, 2014.

[7] Sunyoung Lee, Sungzoon Cho, and Patrick M Wong. Rainfall prediction using artificial neural networks. *journal of geographic information and Decision Analysis*, 2(2):233–242, 1998.

[8] Adil M Bagirov, Arshad Mahmood, and Andrew Barton. Prediction of monthly rainfall in victoria, australia: Clusterwise linear regression approach. *Atmospheric research*, 188:20–29, 2017.

[9] Mohammed Moulana, Kolapalli Roshitha, Golla Niharika, and Maturi Siva Sai. Prediction of rainfall using machine learning techniques. *International Journal of Scientific & Technology Research*, 9:3236–3240, 2020.

2021. In Raju Pal & Praveen K. Shukla (eds.), *SCRS Conference Proceedings on Intelligent Systems*, 79–95. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-93-91842-08-6-7

[10] P Asha, A Jesudoss, S Prince Mary, KV Sai Sandeep, and K Harsha Vardhan. An efficient hybrid machine learning classifier for rainfall prediction. In *Journal of Physics: Conference Series*, volume 1770, page 012012, 2021.

[11] S Sakthivel et al. Effective procedure to predict rainfall conditions using hybrid machine learning strategies. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(6):209–216, 2021.

[12] Diwakar Naidu, Babita Majhi, and Surendra Kumar Chandniha. Development of rainfall prediction models using machine learning approaches for different agro-climatic zones. In *Handbook of Research on Automated Feature Engineering and Advanced Applications in Data Science*, pages 72–94. IGI Global, 2021.

[13] Tuan Vu Dinh, Hieu Nguyen, Xuan-Linh Tran, and Nhat-Duc Hoang. Predicting rainfall-induced soil erosion based on a hybridization of adaptive differential evolution and support vector machine classification. *Mathematical Problems in Engineering*, 2021.

[14] R Malathi and M Manimekalai. Ant colony–information gain based feature selection method for weather dataset. *Annals of the Romanian Society for Cell Biology*, pages 3838–3850, 2021.

[15] Nor SamsiahSani, Israa Shlash, Mohammed Hassan, Abdul Hadi, and Mohd Aliff. Enhancing malaysia rainfall prediction using classification techniques. *J. Appl. Environ. Biol. Sci*, 7(2S):20–29, 2017.

[16] David Ahijevych, James O Pinto, John K Williams, and Matthias Steiner. Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Weather and Forecasting*, 31(2):581–599, 2016.

[17] John T Allen. Climate change and severe thunderstorms. In *Oxford research encyclopedia of climate science*. 2018.

[18] Harold E Brooks. Severe thunderstorms and climate change. *Atmospheric Research*, 123:129–138, 2013.

[19] Pierre Gentine, Mike Pritchard, Stephan Rasp, Gael Reinaudi, and Galen Yacalis. Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, 45(11):5742–5751, 2018.

[20] Michael J Mcphaden, Gary Meyers, K Ando, Y Masumoto, VSN Murty, M Ravichandran, F Syamsudin, Jérôme Vialard, Lianbo Yu, and W Yu. Rama: the research moored array for african–asian–australian monsoon analysis and prediction. *Bulletin of the American Meteorological Society*, 90(4):459–480, 2009.

[21] Peter BR Hazell. The appropriate role of agricultural insurance in developing countries. *Journal of International Development*, 4(6):567–581, 1992.

[22] Peter P Mollinga, Ruth S Meinzen-Dick, and Douglas J Merrey. Politics, plurality and problemsheds: A strategic approach for reform of agricultural water resources management. *Development Policy Review*, 25(6):699–719, 2007.

[23] Chirag Shah, Chathra Hendahewa, and Roberto González-Ibáñez. Rain or shine? forecasting search process performance in exploratory search tasks. *Journal of the Association for Information Science and Technology*, 67(7):1607–1623, 2016.

[24] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear model selection and regularization. In *An introduction to statistical learning*, pages 225–288. Springer, 2021.

[25] Raksha Agarwal and Niladri Chatterjee. Langresearchlab_nc at cmcl2021 shared task: Predicting gaze behaviour using linguistic features and tree regressors. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 79–84, 2021.

[26] S Benítez-Peña, E Carrizosa, V Guerrero, MD Jiménez-Gamero, B Martín-Barragán, and C Molero-Río. On sparse ensemble methods. 2021.

[27] Kim de Bie, Ana Lucic, and Hinda Haned. To trust or not to trust a regressor: Estimating and explaining trustworthiness of regression predictions. *arXiv preprint arXiv:2104.06982*, 2021.

[28] Mauricio González-Palacio, Lina Sepúlveda-Cano, and Ronal Montoya. Simplified path loss lognormal shadow fading model versus a support vector machine-based regressor comparison for determining reception powers in wlan networks. In *International Conference on Information Technology & Systems*, pages 431–441. Springer, 2021.