IJITCE

ISSN 2347- 3657

International Journal of
Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9,Issue 2,April 2021

Analyzing the Performance of Machine Learning in Weather Forecasting

Dr. M V Rathnamma[1], B Mahesh Reddy[2], T Anitha[3], Dr. V. Lokeswara Reddy[4]

Abstract

Forecasting storms with significant rainfall is a difficult task for meteorologists. Several Machine Learning (ML) models, including the Lasso regression, ridge regression, elastic net regression, random forest, gradient boosting, and the decision tree regress or, are examined here. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Square Error (MSE), and R2 score have all been used to assess the quality of the models. Using the rainfall dataset, this study compares and contrasts several machine learning regression methods. After carefully considering the outcomes of six alternative ML models, we settled on Lasso regression of the linear model as the most effective. The Lasso model obtained a 99.21% R2 score, 13.68 MAE, 6432.41 MSE, and 80.20 RMSE when used on 80% of the training data set and 20% of the test dataset, respectively.

Introduction

Predicting rain properly is a major challenge in hydrological research. Farmers may suffer losses and crop destruction as a result of natural disasters and storms. Predicting rainfall properly and providing extra warning to farmers ahead of time might help them avoid these issues. Many human endeavors, including agriculture, building, energy generation, forestry, tourism, and so on, are considered to be influenced by precipitation as a result of its status as an environmental factor. Due to its greater focus on interconnected phenomena like landslides, floods, avalanches, earthquakes, and so on, rainfall forecasting is of greater importance. On [1]. Predicting rainfall is crucial because of the strong correlation it has with other natural disasters such as landslides, floods, avalanches, earthquakes, and so on. The public suffers greatly for decades after such calamities [2]. To a certain degree, natural

[1]Associate Professor, Department of CSE ,K.S.R.M College of Engineering(A), Kadapa
[2,3] Assistant Professor, Department of CSE, K.S.R.M College of Engineering(A), Kadapa
[4]Professor, Department of CSE, K.S.R.M College of Engineering(A), Kadapa

disasters may be avoided [3] if an accurate model of rainfall is developed. In order to construct ML models capable of producing reliable forecasts, we relied on a number of different regression-based machine learning methods. The model of the system's behavior is studied and refined with the help of machine learning. Models for the Indian panther ecology were developed using machine learning modeling approaches, allowing for the prediction of key system factors. From gathering data to testing hypotheses, this article strives to cover the whole machine learning process from start to finish. Regression coefficient (R2), mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) are useful measures for assessment.

Following is the article's outline: The second section provides a comprehensive literature evaluation of several regression techniques for predicting precipitation. In the third section, we go into the specific machine learning regression models used in this study. Part IV explains the cutting-edge experiments carried out in the direction of the use of deep learning techniques in weather forecasting. Part V elaborates on the last thoughts and the way forward.

Literature Review

We have significant difficulties in making reliable weather forecasts. Methods for predicting rain use a suite of mathematical models and established statistical trends. Predicting the weather in a precise and timely manner is a significant challenge for scientists. Multiple computational models and empirical data are used in precipitation forecasting. In both economic and climatic forecasting, regression is a common statistical and empirical technique. Rainfall-runoff connection in a Japanese catchment zone was predicted using an Artificial Neural Network by El-shafie [5] et al. Feed-forward backpropagation was proposed, using hyperbolic tangent neurons in the activation layer and linear neurons in the target layer. Other statistical measures, such as correlation coefficients and mean square error, are used to assess model effectiveness. An improved model was suggested. Nikhil Sethi, [6] and others. For agricultural purposes, the suggested approach of predicting future rainfall based on an understanding of climatic parameters is very useful. Only one model, based on machine learning algorithmic multiple linear regressions, is proposed in this paper. Authors Ashwani [7] et al. Artificial neural network (ANN) and decision tree It had been estimated using algorithms applied to meteorological data collected during a certain time frame. Accurate ratings based on the standard implementation characteristics of algorithms were used to evaluate the models' capabilities and choose the best one for weather forecasting. An alternative framework was created by Liu et al. It is used to identify situations in which Feature Selection (FS) models and Naive Bayes (NB) prediction techniques may be put to use. Rainfall event, also known as a binary prediction module, and rainfall categorization, which may include light, steady, or severe rainfall, are the two main types of predictive modules. The goal of using GA is to choose inputs that provide a workable choice to reduce the effort required for a dataset to achieve the same or best function.

Research Methodology

ISSN 2347–3657
Volume 9,Issue 2,April 2021

The whole process of the Framework for the machine learning method is explained in figure1. The entire process is classified into four stages had been a, namely data acquisition, data pre-processing, Build the machine learning model, and predict the target variable with the trained model. From figure 1, the framework for the machine learning model has been explained below.

Data Acquisition

In the first stage, Raw Data which is collected from the India Meteorological Department (IMD) Govt. of India [9]. Raw data is observed in figure 2.

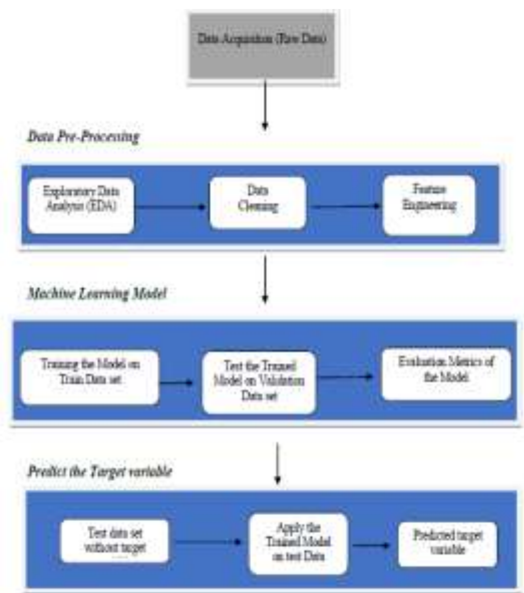


Figure 2.Framework for Machine learning Model

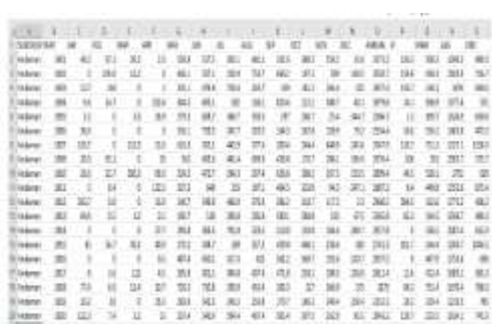
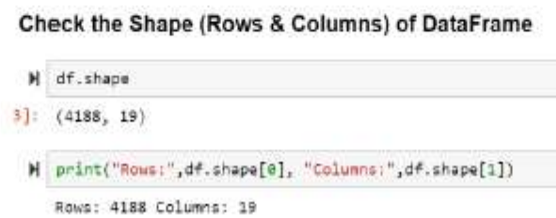


Figure 2. Raw Dataset of rainfall Collected data set includes monthly rainfall feature of 36 meteorological sub-divisions of India during the period of 1901-2017 is observed in figure 2. Given data set contains 4188 instances and 19 features are observed in figure 3. Is Among19 variables „Annual‟ variable is a target variable and remaining 18 variables are 'SUBDIVISION', 'YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL','AUG', 'SEP', 'OCT', 'NOV', 'DEC', 'JF', 'MAM', 'JJAS', 'OND' considered as input variables (predictor variables).


Check the Shape (Rows & Columns) of DataFrame

```
df.shape
(4188, 19)
print("Rows:",df.shape[0], "Columns:",df.shape[1])
Rows: 4188 Columns: 19
```

Figure 3. Shape of the data set *3.2 Data Pre-Processing* For achieving better results, the applied ML models, Data Pre-processing, is the second stage of the research methodology, which has three sub-parts, which are Exploratory Data Analyses (EDA), data cleaning, and feature engineering. *3.2.1 Exploratory Data Analysis (EDA)* Exploratory Data Analysis is about applying methods on data to gain observation before applying machine learning techniques. EDA explains data by means of statistical and visualization techniques. It brings out the essential aspects of the data. EDA also plays a crucial role in helping choose the right ML model to solve a specific problem. *3.2.1.1 Average Annual Rainfall in Each Subdivision* Here, we are finding the subdivisions with the highest and lowest rainfall, from figure 4 we notice that Subdivisions with highest annual rainfall are "COASTAL KARNATAKA," "ARUNACHAL PRADESH" and "KONKAN & GOA" with an approximate yearly

International Journal of
Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9,Issue 2,April 2021

rainfall of 3403 mm, 3397 mm and 2987 mm respectively. Subdivisions with the lowest annual rainfall are "HARYANA DELHI & CHANDIGARH," "SAURASHTRA & KUTCH," and "WEST RAJASTHAN" with an approximate annual rainfall of 528 mm, 496 mm and 294 mm respectively.

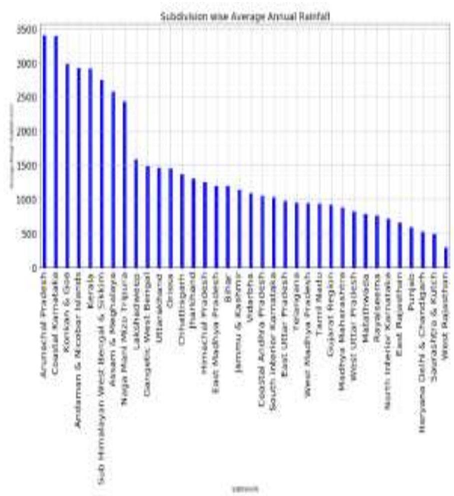


Figure 4. Subdivision wise highest and lowest rainfall

3.2.1.2 Rainfall in Subdivisions From figure 5, we noticed that, majority of rainfall is received in the months of JUNE, JULY, AUGUST, SEPTEMBER (JJAS) from Coastal Karnataka, Arunachal Pradesh, Konkani Goa, and Kerala and which are receiving the highest rainfall.

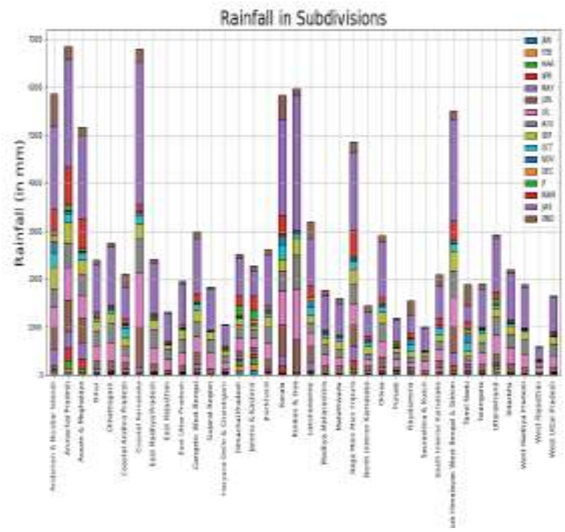


Figure 5. Rainfall in subdivisions of monthly wise

3.2.2 Data Cleaning Data cleaning is the subpart in data pre-processing. Under data cleaning, some of the operations had been applied to handle unnecessary data like duplicates, outliers, and missing values. From Figures 6 & 7, we notice that how much percentage of values is missing in all features. Fill the null values with a mean of that corresponding that features.

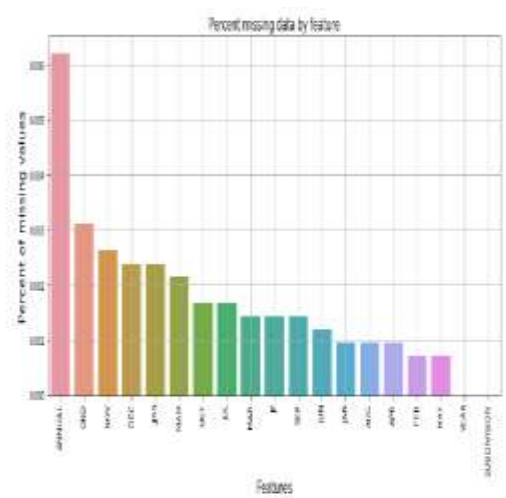


Figure 6. Percentage of missing values by features

International Journal of
Information Technology & Computer Engineering
ISSN 2347–3657
Volume 9,Issue 2,April 2021

	Total	Percent
ANNUAL	26	0.006209
OND	13	0.003104
NOV	11	0.002627
DEC	10	0.002388
JJAS	10	0.002388
MAM	9	0.002149
OCT	7	0.001671
JUL	7	0.001671
MAR	6	0.001433
JF	6	0.001433
SEP	6	0.001433
JUN	5	0.001194
JAN	4	0.000955
AUG	4	0.000955
APR	4	0.000955
FEB	3	0.000716
MAY	3	0.000716
YEAR	0	0.000000

Figure 7. Percentage of Missing values with the total number of values in features

4.1) Linear Models

There are three exclusive linear regression models, which are Lasso, Ridge, and Elastic Net regression. The easiest method to forecast output by applying a linear function of input features.

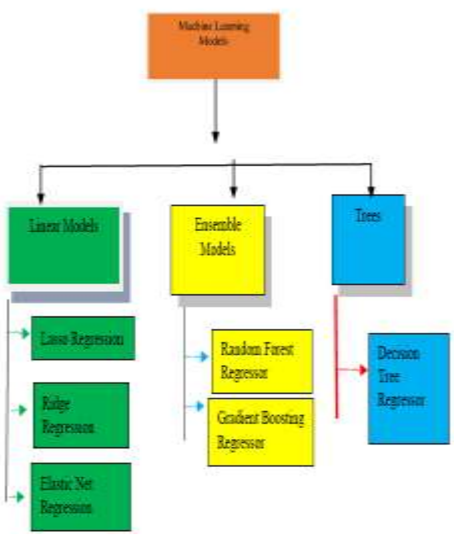


Figure 8. Regression of Machine Learning models there is an association between one or more independent or input features (X) and dependent or target feature (y) for simple Linear Regression (SLR). The regular equation for linear regression is assumed as yi= mixi+b. For multiple explanatory variables, where 'y' represents the target feature, and 'X' represents independent variable where i=0,1,2,.., n, indicates the explanatory or independent variables, 'm' termed as a slope. The process has been explained as Multiple Linear Regression (MLR)[10].

$$\hat{y} = m_0 x_0 + m_1 x_1 + \cdots + m_n x_n + b \qquad \text{-----------(1)}$$

$$= \sum_{i=1}^{r} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{r} \left(y_i - \sum_{j=0}^{c} m_j \cdot x_{ij} \right)^2 \text{-----------------(2)}$$

The cost function for simple linear regression is defined in equation (2) from this equation; assume that there is being 'r' rows or instances and 'c' columns or features. The whole data set has been classified into a train and validation data set. Lasso and ridge regression models are used to minimize the complexity of the model and prevent over-fitting problems. *4.1.1) Ridge Regression:* Add the penalty to the square of the magnitude in the coefficient in the ridge regression.

$$= \sum_{i=1}^{r} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{r} \left(y_i - \sum_{j=0}^{c} m_j \cdot x_{ij} \right)^2 + \lambda \sum_{j=0}^{c} m_j^2 \text{---------(3)}$$

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657
Volume 9,Issue 2,April 2021

The above equation (3) is the cost function of Ridge regression. So, ridge regression had been set a constraint on the coefficients (m). [11] Factors had been regularized when we apply the penalty term (lambda (λ)), then the optimization function is penalized. So, ridge regression minimizes the coefficients, and it helps to decrease the model complication. The significant advantage of ridge regression is „coefficients shrinking" and reducing the „model complication." supposing, when putting λ=0, the cost function of ridge regression becomes similar to the cost function of linear regression (eq.2).

4.1.2) Lasso Regression LASSO (Least Absolute Shrinkage and Selection Operator) regression [12] cost function can be written as

$$\sum_{i=1}^{r}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{r}\left(y_i - \sum_{j=0}^{c} m_j \cdot x_{ij}\right)^2 + \lambda \sum_{j=0}^{c} |m_j| \text{------------}(4)$$

The above equation (4) is the cost function for Lasso regression. So, coefficients of Lasso regression are similar to ridge Constraints on ridge regression coefficients. If λ=0, then equation 4 becomes equation 2, means Lasso regression becomes like cost function of simple linear regression. The difference between lasso and regression is the magnitude of coefficients. Some of the independent variables are removed from the dataset and select the most significant features for calculating the output. So, the main advantage of Lasso regression is to avoid over fitting and choose the best features.

Results and Discussion

Performance Measure In this section, we study the regression of machine learning algorithms. According to results of lasso, ridge and elastic net of linear models, random forest regress or, gradient descent regress or of ensemble models and decision trees of trees are explained before, and then we compare the results. As stated, in the paper total 4188 instances out of which 80% of data that is 3350 data samples for training and 20% of data that is 838 data samples are chosen for testing purpose. The results in this paper have been taken from test data hat is 838 data samples. The evaluation metrics for regression algorithms are R^2 score, Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

R Squared (R^2 or R^2 score) R^2 tells us, "How well a regression line predicts actual values." R-squared is the proportion of the target variable difference that is described by the linear model. The R-squared value lies between 0 and 100%. If the R-squared value is significant means about 100%, then the model properly fits data. If R-squared value is very fewer means about to '0', then the model not properly fits data and gives the wrong predictions. Here, y is the best fit line values; y is the mean of the actual values.

$$R^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

5.1.2) Mean Absolute Error (MAE) There are many ways of measuring the model's performance. MAE is one of the metrics for brief and evaluating the quality of a machine learning model. The error is calculated in MAE as an average of the absolute difference between the actual values and the predicted values.

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657
Volume 9,Issue 2,April 2021

Where y_i is the real value and $\hat{y}_i$ is the predicted value.

$$MAE=\frac{1}{r}\sum_{i=1}^{r}|y_i-\hat{y}_i|$$

Analysis of Results

This section analyses the result of the extensive experiment conducted on different Machine learning (ML) algorithms such as Lasso, Ridge, Elastic Net of Linear Models, Random Forest, and Gradient Boosting Regress or of Ensemble Models and Decision Tree Regress or for rainfall datasets. Table 1 shows the performance measurements of ML solutions on rainfall datasets. From Table 1, the lasso regression model provides better R^2 Score performance.

Table 1.*Regression Performance of six-ML Algorithms*

Models	Train & Test (%)	R^2 Score	MAE	MSE	RMSE
lasso	70-30	98.44	10.80	12673.84	112.57
ridge	70-30	98.48	10.92	12339.40	111.08
enet	70-30	98.42	13.01	12819.53	113.22
rf	70-30	97.87	45.78	17312.12	131.57
gb	70-30	97.97	42.54	16468.26	128.32
dtr	70-30	96.63	83.30	27387.45	165.49
lasso	75-25	98.14	11.30	14847.94	121.85
ridge	75-25	98.18	11.37	14573.74	120.72
enet	75-25	98.11	13.62	15112.67	122.93
rf	75-25	97.37	45.76	21012.06	144.95
GB	75-25	97.84	37.10	17292.54	131.50
dtr	75-25	95.53	88.30	35759.71	189.10
lasso	**80-20**	**99.21**	**13.68**	**6432.41**	**80.20**
ridge	80-20	99.10	16.67	7307.15	85.48
enet	80-20	99.13	15.58	7110.44	84.323
rf	80-20	98.58	45.19	11602.47	107.71
gb	80-20	98.76	40.20	10136.62	100.68
dtr	80-20	95.91	82.50	33521.73	183.08
lasso	90-10	98.65	15.32	11025.61	105.00
ridge	90-10	98.60	17.80	11441.76	106.96
enet	90-10	98.58	16.71	11571.53	107.57
rf	90-10	98.04	47.19	16032.62	126.62
gb	90-10	98.11	41.34	15450.79	124.30
dtr	90-10	96.84	83.05	25817.36	160.67

Here, lasso=Lasso Regression, ridge=Ridge Regression, enter=Elastic Net Regression, fro=Random Forest Regression, gab= Gradient Boosting, dry=Decision Tree Regressor. The comparison of R^2 Score for different ML models is graphically presented in Fig.10. Among six ML models, and Lasso regression model has the highest R^2 Score with 99.21% compared to the remaining ML models at 80% train data set and 20% test data set.

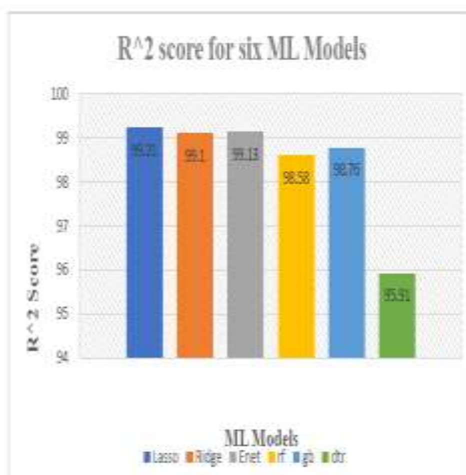


Figure 10. Comparison of R^2 Score (%) for ML Models

The comparison of Mean Absolute Error (MAE) for different ML models is graphically presented in Fig.11. Among six ML models, the Lasso regression model has the lowest MAE value, with 13.68 compare to remaining ML models at 80% train data set and 20% test data set.

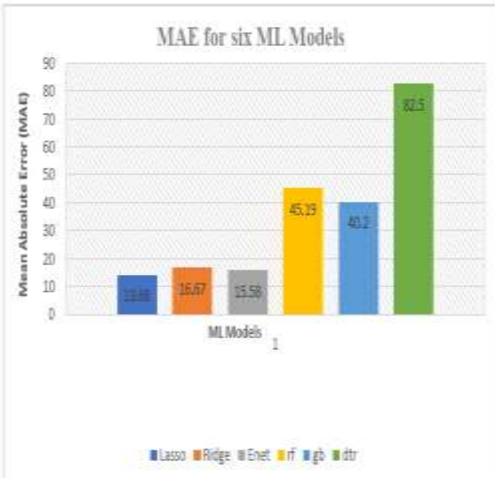


Figure 11. Comparison of MAE for ML Models The comparison of Mean Square Error (MSE) for different ML models is graphically presented in Fig.12. Among the six ML models, the Lasso regression model has the lowest MSE value with 6432.41 compare to the remaining ML models at 80% train data set and 20% test data set.

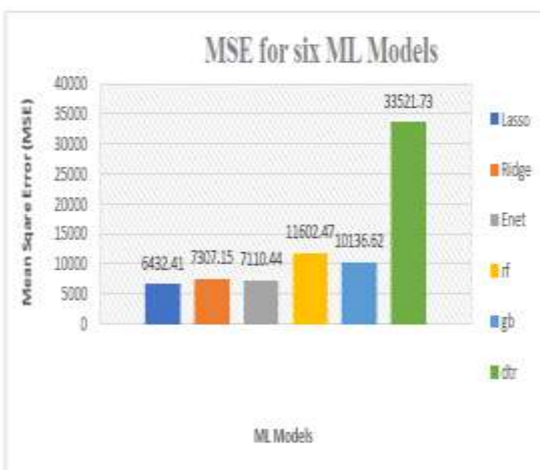


Figure .12. Comparison of MSE for ML Models

The correlation of Root Mean Square Error (RMSE) [z] for different ML models is graphically presented in Fig.13. Among six ML models, the Lasso regression model has the lowest RMSE value, with 80.2 compared to the remaining ML models at 80% train data set and 20% test data set.

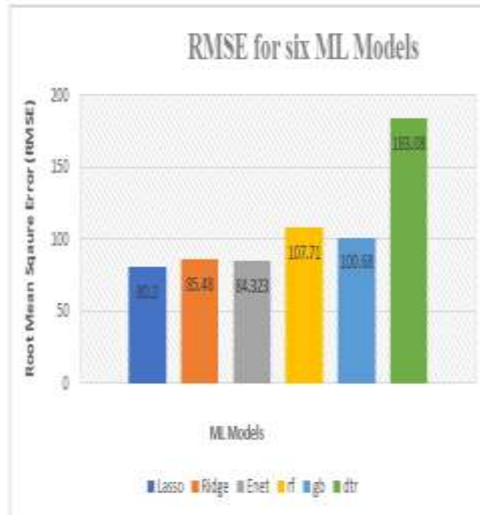


Figure 13. Comparison of RMSE for ML Models

Conclusions and Future Works

The automated regression of precipitation is a common use of ML algorithms. This study synthesizes and illustrates the theory behind six ML algorithms, and it experimentally compares the regression performance of all ML techniques on a standard rainfall dataset. With an R2 score of 99.21% using just 80% to 20% of the training and validation dataset, lasso regression outperformed the other five techniques. In addition, the efficacy of each ML algorithm is measured by contrasting its predictions with the actual targets. Regression methods may be used inside the feature to fine-tune precision.

References

N. Ghana ([1]) Published in the International Journal of Advanced Science and Technology, Volume 29 Issue 8s (2020), pages 746-758, is "A Multiple Linear Regression Model to Predict Rainfall

ISSN 2347–3657
Volume 9,Issue 2,April 2021

Using Indian Meteorological Data" by E. Ramada Ankara.

Alcntara-Ayala, an Irishman. Natural hazards, vulnerability, and catastrophe avoidance in low-income nations. October 2002 issue of Geomorphology, 47(24):107-124.

[3]. N. Nicholls Neville. Climate and Atmospheric Risks: New Methods for Keeping Tabs

Disaster Prevention via Prediction. Risk Analysis, 2001, 23, 2, 137–155

"Machine Learning-Based Modeling of Human Panther Interactions in the Ravalli Hills of Southern Rajasthan," by Puente Sharma and Nadir Chitty, appears in the Indian Journal of Ecology 46(1): 126-131 (Reference 4).

Performance of artificial neural network and regression approaches for rainfall-runoff prediction, International Journal of the Physical Science, volume 6(8), pages 18 April 2011; A.El-shafie, M.Mukhlisin, Ali A. Rajah, and M.R. Tasha.

The article "Exploiting Data Mining Technique for Rainfall Prediction" by Nikhil Seth et al. can be found in [6].

Published in 2014 in International Journal of Computer Science and Information Technologies, Volume 5, Issue 3, Pages 3982-3984, ISSN: 0975-9646.

Authors Ashbin Mondale and Jadhawar B.A. (March/April 2015) International Journal of Engineering Research and General Science Volume 3 Issue 2 ISSN 2091-2730 "Weather Forecast Prediction: A Data Mining Application"

(Referenced work) Liu, B. N. L. Li, and Dillon, T. S. Together, the innovative input solution approach and the enhanced naive Bayesian classifier technology allow for more accurate rainfall forecasting. Part C: Applications and Reviews, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 2, No. 2 (2001):249-256.

https://data.gov.in/resources/sub-divisional-monthly-rainfall-1901-2017 is where you can get this information [9].

Regression model in machine learning research: Shen Rong and Zhang Bao-wen, MATEC Web of Conferences 176, 01033 (2018).