# Rainfall Pattern Prediction

Deepthi Sudharsan
*CSE - AI*
*Amrita Vishwa Vidyapeetham*
Ettimadai, India

Isha Indhu S
*CSE - AI*
*Amrita Vishwa Vidyapeetham*
Ettimadai, India

Kavya S Kumar
*CSE - AI*
*Amrita Vishwa Vidyapeetham*
Ettimadai, India

Meghna Menon
*CSE - AI*
*Amrita Vishwa Vidyapeetham*
Ettimadai, India

Rama Sailaja
*CSE - AI*
*Amrita Vishwa Vidyapeetham*
Ettimadai, India

*Abstract*—**Rainfall Pattern detection is one of the most important real-life problems, especially in the agricultural sector. Weather pattern forecasting is essential to understand and analyze factors like the right crop that needs to be grown, the best time to harvest, precautions to undertake on the field, etc. Weather plays a crucial role in the whole agricultural process. It is not that easy to predict the course of weather accurately and here is where Machine Learning and AI come to play. Highly trained models tuned with the right hyper-parameters can yield accurate results in forecasting. The project covers the detection of rainfall patterns using some machine learning models.**

*Index Terms*—**Rainfall Pattern Detection, Machine Learning**

## I. Introduction

Rainfall pattern prediction revolves around the forecasting of the trend in rainfall (like intensity and time) based on its previous results. Rainfall prediction that can predict heavy or no rainfall can, in turn, prevent any huge risk such as property damage, flood, or drought. Accuracy in such predictions is very crucial and for that purpose traditional methods prove inefficient. Therefore, we require a reliable technique like machine learning models to take up such tasks. Machine Learning models when perfectly tuned can predict with the finest accuracies. Accurate forecasting can help the agricultural sector. Like farmers for instance can choose which crops to grow, when and how much they should sow and harvest the yields etc. with minimal loss, maximum profit, and optimum methodologies and standards.

## II. Data

### A. Dataset

As is known, data is extremely important to any machine learning algorithms, the same applies to this project. The dataset used by the project is for the rainfall distribution of Indian states from the year 1901 to the year 2015. The rainfall distribution of each month in these years has been provided for each state. There are a few months for a few states where the information is not available are marked as “NA” during classification these entries are not considered. There are 36 regions into which India has been divided in this data, a few of the union territories have been considered as a part of the state, and large states have been divided further into regions. This data has been stored as comma separated values. In python the pandas package is used to read the csv file. For this project the predictions are done state-wise, to make this possible the data belonging to the particular states are separated into another data frame. Some exploratory data analysis is performed on the data to look at the distribution of the rainfall. For the state of Tamil Nadu, the rainfall distribution of each month for every year, has been plotted on a histogram along with the complete data of Tamil Nadu.

## III. Methods

Four machine learning algorithms have been used to implement this project, namely

- Linear Regression
- Lasso Regression
- Random Forest

### A. Linear Regression

The most elementary regression model is the simple linear regression which explains the linear relationship between the dependent variable (Target variable) and one independent variable using a straight line

$$Y = \beta_0 + \beta_1 X$$

The best fit line is identified by minimizing the error term RSS (Residual sum of error) which is the sum of squares of residual for each point. Residual for a point is identified by subtracting the actual value with the predicted value. Linear regression makes the following four assumptions:

1) There should be a linear and additive relationship between the dependent and independent variable (Linearity of residual ). If a linear model is fit to a non-linear, non-additive set, then the regression model will fail to capture the trend, thus resulting in an inefficient model. This will lead to erroneous prediction on unseen data.
2) The error terms should not be dependent on one another, and there should be no correlation between the residual (error) terms.
3) The mean of residual should follow a normal distribution with mean equal to zero or close to zero. This is done

to check whether the selected line is actually the line of best fit.

4) The error term must have a constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred as heteroscedasticity.

### B. Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator. The equation for lasso regression is as follows

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

### C. Random Forest

Random Forest is a classifier that contains a number of individual decision trees on numerous subsets of the given dataset and takes the average to enhance the predictive accuracy of that dataset. Rather than depending on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. One big advantage of random forest is that it can be used for both classification and regression problems. In ML language, random forests are also called an ensemble or bagging method. Random Forest does the tasks in two-phases. First being, to create the random forest by combining N decision tree, while the second task is to make predictions for each tree created in the first phase.

The Working process can be explained in the following steps:
Step-1: Selecting random K data points from the training set.
Step-2: Building the decision trees associated with the selected data points (Subsets).
Step-3: Choosing the number N for decision trees that you want to build.
Step-4: Repeating Step 1 and 2.
Step-5: Finding the predictions of each decision tree, for new data points, and assigning the new data points to the category that wins the majority votes.

## IV. Implementation

Implementation Steps are given below:

1) Data Pre-processing step
In this step the data is pre-processed where dataset is loaded.
2) Fitting the Random Forest algorithm to the Training set
Fitting the Random forest algorithm to the training set. To fit it, we will import the Random Forest Classifier class from the sklearn.ensemble library.
3) Predicting the test result
Since the model is fitted to the training set, now we can predict the test result. We will create a new prediction vector $\mathrm{y}_{\mathrm{pred}}$, for prediction.
4) Visualizing the test set result.
Now we will visualize the test set result.

### A. Outputs

Fig. 1. Lasso Regression Errors

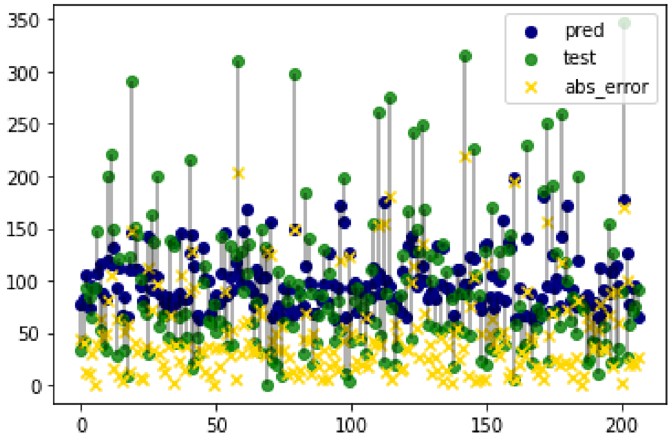
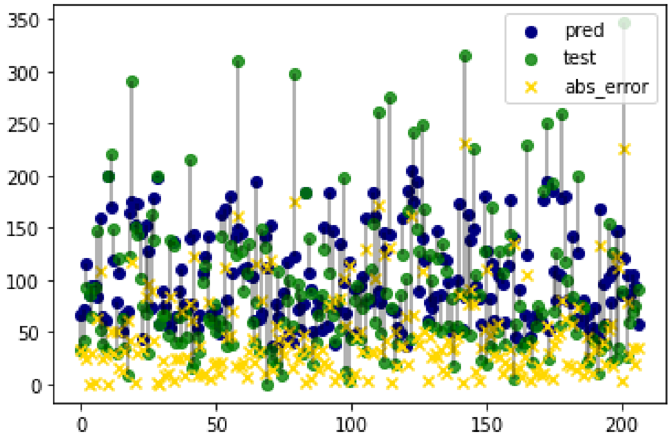


Fig. 2. Linear Regression Errors



Dumbbell plots have been used to mark the differences between the actual points and the points predicted by the classifiers. The perpendicular lines seen in Fig. 2, Fig. 1, Fig. 3 are the error values of the predictions.
In the Fig. 4, it can be observed that the predictions by the Random Forest Classifier are very close to the actual values of the rainfall distribution.

Fig. 3. Random Forest Errors

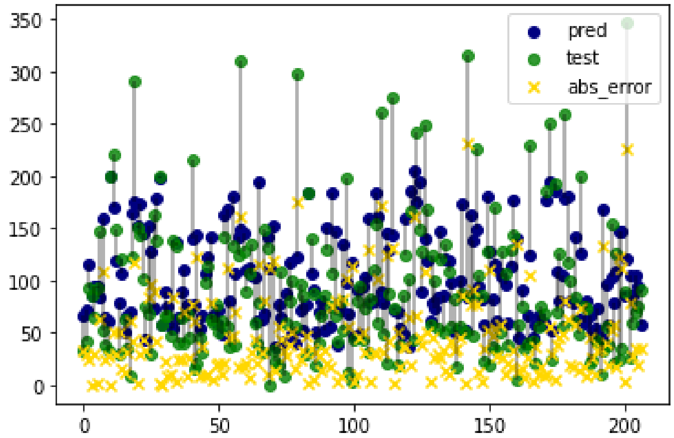
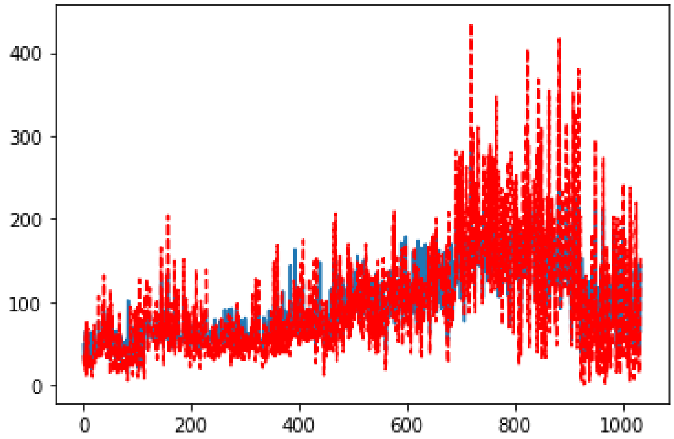


Fig. 4. Similarities between actual data and predicted values (RF)



## V. Conclusions

This project has been designed to predict the rainfall for one state in India, Tamil Nadu. Rainfall Prediction is one of the most difficult and uncertain tasks that has a significant impact on human society. Accurate and timely rainfall prediction can proactively help reduce human and financial loss. This project was started with the hopes that it will help the farmers and other people in the agricultural industry to choose their crops wisely for the harvest season so that they would not have to face any loss.

## VI. Future Scope

As of now the model has been trained for a year for the above mentioned state.It can further be expanded to the other states of India. The program will also be made general by making it pick the best model using cross validation and accuracy score to predict the rainfall. The accuracy of the model can also be boosted using methods like algorithm tuning, feature selection etc. Other machine learning algorithms can also be used to predict the results.

## Acknowledgment

This project would not have been possible without the support and guidance of Dr. Geetha who has provided us with sufficient knowledge and guidelines even during these unprecedented and trying times.

We would also like to thank our Computer Science and Engineering (Artificial Intelligence) department for giving us this opportunity to nurture and hone our skills.

Furthermore, we would like to thank the Amrita Vishwa Vidyapeetham management for providing ample resources to avail our project needs and a platform for online lectures during this lockdown.

## References

[1] https://www.javatpoint.com/machine-learning-random-forest-algorithm
[2] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.html
[3] https://github.com/vgaurav3011/Rainfall-Prediction/blob/master/Exploration_Rainfall_Data.ipynb