# Rainfall Prediction System

**Deepkumar Patel**
**deepkimar.p@ahduni.edu.in**

**Devarsh Suthar**
**devarsh.s1@ahduni.edu.in**

**Dhruv Panchal**
**dhruv.p1@ahduni.edu.in**

**Vishvas Patel**
**vishvas.p@ahduni.edu.in**

***Abstract*—India is primarily an agricultural country, with crop productivity and rainfall playing a large role in its economy. Rainfall prediction is required and mandatory for all farmers in order to analyze crop productivity. Predicting the condition of the atmosphere using science and technology is known as rainfall prediction. Hence, rainfall prediction becomes essential as heavy rainfall can lead to many disasters. Hardware components often fail to predict rainfall accurately. As a result, using machine learning techniques, farmers will benefit greatly from accurately predicted rainfall results. For these purposes, we have compared and evaluated various classification models such as Random Forest and Support Vector Classification (SVC) in this paper. Also we have applied Permutation importance and Chi-square feature selection methods to improve their accuracy.**

***Keywords*—Rainfall | Prediction | Machine Learning | Classification | Permutation Importance | Chi-square | Accuracy**

## I. Introduction

Rainfall prediction is a major problem for the meteorological department as it is closely associated with the economy and human life. Accuracy of rainfall forecasting has great importance for countries like India whose economy is highly dependent on agriculture. There are hardware devices for predicting rainfall by using the weather conditions such as dew point, temperature, humidity, visibility, cloud cover, pressure etc. But these traditional methods cannot work efficiently. However, by using machine learning techniques we can produce accurate results by analyzing historical rainfall data. We extracted data of Ahmedabad city from the past 11 years (i.e, 2009-2020 ) from an API [2]. The extracted data includes hourly forecasts of 19 parameters including precipitation(MM), temperature, dew point, wind speed, pressure, visibility, etc. There were many redundant columns in the rainfall data such as snowfall. These unnecessary columns were discarded. In this article, for the feature selection, we used the Random forest classifier with permutation importance to achieve the most important features. After that, we trained our model with the selected important features and analyzed the accuracy score. We used the Random forest classifier because it provides higher accuracy by aggregating the prediction values of the decision trees. Similarly, we analyzed accuracy of the support vector classifier (SVC) with the permutation importance feature selection. The main reason to use SVC is it provides the best fit hyperplane to the data. The permutation feature importance assisted us in identifying the prominent features from our dataset. Further, to improve the accuracy and to create the strong predictive model, we used the gradient boosting classification. It helps us improve the prediction by minimizing the prediction error. Future, we compared all the classifiers and analysed the results. The graphs for the same can be found below in the results section of this article.

## II. Literature Overview

Maulana, Rositha .et al [1] has described the rainfall prediction using monthly data from 1901-2009 using regression methods like Multiple linear regression, Support vector machine, Lasso regressions and obtained 99% accuracy.

Nikhil Oswal [3] defined the rainfall prediction for the date-wise data using a variety of machine learning techniques and models such as Logistic regression, Decision tree, k-mean, Random forest and obtained 84% accuracy.

Chandrasegar Thirumalai .et al [4] has described the rainfall prediction rate for future years using the amount of rainfall in previous years. The article includes a number of machine learning techniques

and models for predicting rainfall. This paper is carried on the heuristic prediction of rainfall using machine learning techniques. The paper also measures the different categories of data by linear regression method in metrics for effective understanding of agriculture in India.

Geetha, A . et al [5] has implemented the system to predict weather phenomena such as fog, rainfall, cyclones, thunderstorm, etc. They used data mining techniques and tools to implement this model. Decision tree algorithms are used to achieve an accuracy of about 80.67%.

# III. IMPLEMENTATION

## A. *Platform*

- To perform classification analysis and comparisons, we used Python as a platform.

## B. *Data Preparation*

- To generate Ahmedabad Rainfall data, we created an API on a global weather online platform. We created a python script that gathered day-by-day data from 2009 to 2020.
- After converting the data into CSV files, we removed some columns(features) which are irrelevant (for example, SnowFall_cm, which displays the day's snowfall in centimeters). This column was irrelevant because snowfall does not occur in Ahmedabad.
- Then we separated dependent (label) and independent variables (Features).
- Moreover, we used permutation importance feature selection which allows us to reduce the number of features by focusing on the most important ones. We also compared and analyzed various classification algorithms.

## C. *Feature Selection*

1. Permutation Importance
   - The simple Idea of permutation importance is to shuffle rows of feature columns one by one. This will act as a noise column in the dataset, and if the predictor's score goes down, it simply means that that feature is important to the data. When a feature is permuted and the prediction score increases, it indicates that the feature is not very important.
   - We use this technique to train models and extract important features from our data, which includes binary label data (Yes/No) and retrain models again to achieve far better accuracy than average.
   - For the random forest classifier, we obtained important features such as DewPointC, cloudcover, uvIndex, and visibility.

2. Chi-Square
   - The Formula for Chi-Square is

$$X_c^2 = \sum \frac{Oi - Ei}{Ei}$$

where:

c = degree of freedom
O = observed values
E = expected values

   - We performed a chi-square test on our dataset to find the k best features. In chi-square, there are hypotheses and null hypotheses, and the chi-square chooses a hypothesis if the contribution of the above equation, referred to as the chi-square value, is greater than 0.05 from the contingency table. As a result, it interprets whether the selected feature is important and dependent on the Label.

## D. *Random Forest Classifier*

- Random forest is a flexible, easy-to-use machine learning algorithm that, in most cases, produces accurate results. Because of its simplicity and diversity, it is also one of the most widely used algorithms. It can be used for classification as well as regression. However, we used it for the classification.
- It is a supervised learning model that consists of decision trees that are trained with the "bagging" method. The row sampling (with replacement) and feature sampling are provided to the decision trees. Each decision tree then gives the prediction.
- If we create a decision tree to its complete depth, a single decision tree has low bias and

high variance. However, when we combine all the decision trees to the majority vote, the High Variance gets converted into the low variance (as we are not dependent on a single decision tree output).

- Each prediction given by the decision trees is aggregated to overcome the high variance.
- Another advantage of the random forest classifier is that when new data is introduced or modified, the random forest is unaffected since the data is distributed among the decision trees. Hence, the data change does not make much impact on the accuracy of the output.

### *E. Support vector classifier (SVC)*

- A Support Vector Machine is a modern and supervised machine-learning method that can be most commonly used in classification problems. Support Vector Classification (SVC) is a form of classification analysis that uses supervised machine learning models and associated learning algorithms to analyze data.
- This method creates hyper-planes in a high or even infinite-dimensional space using linear algorithms.
- In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a specific coordinate
- Support Vector Machine helps us to perform an efficient rainfall prediction with a low error rate.

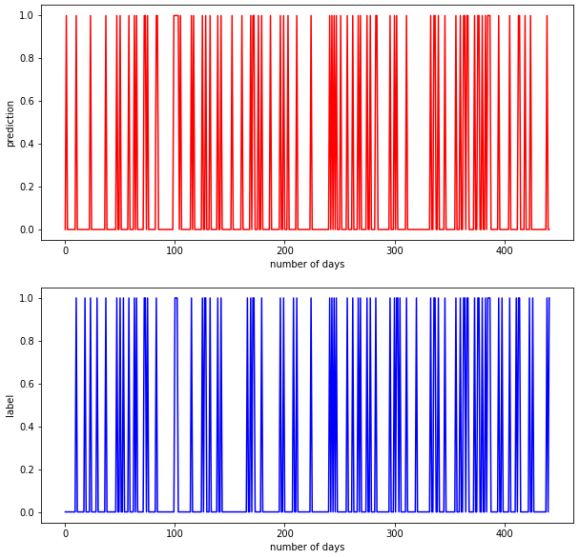
### *F. Logistic Regression*

- Logistic regression is a simple yet very effective classification algorithm to solve a classification problem.
- Logistic Regression Classification algorithms analyze the categorical or binary data and predict the probability of the target variable.
- In this article, we have categorized the rainfall outcomes into four categories: no rain, drizzle, moderate rain, and heavy rain.

## IV. Results

* In all the results shown below, the red graph represents predictions and the blue graph represents the label values.

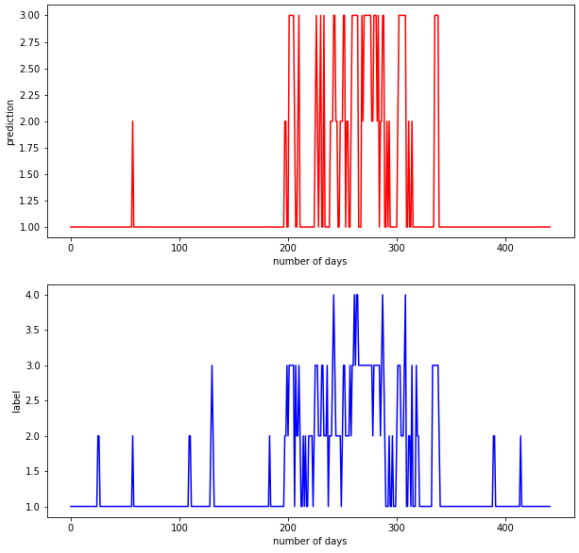
### Random forest (Permutation Importance)




**[Figure 1]**

- Figure 1 depicts a binary classification ( Rain and No Rain).
- The obtained accuracy is 93.44%.

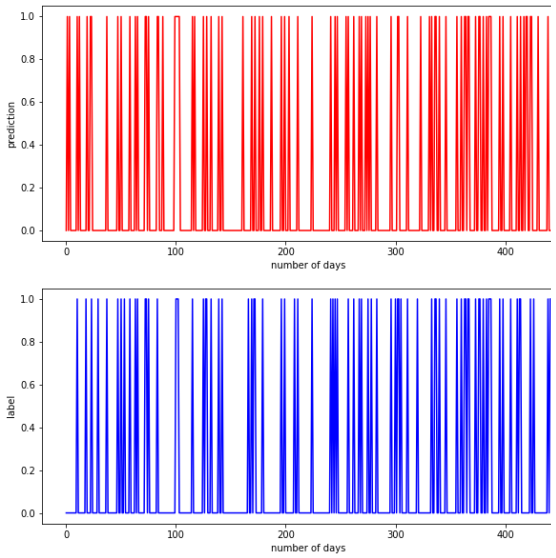
### Random forest (Chi-square)




**[Figure 2]**

- Figure 2 depicts four classifications ( No Rain, Drizzle, Moderate Rain, and Heavy Rain ).
- The obtained accuracy is 83.71%.

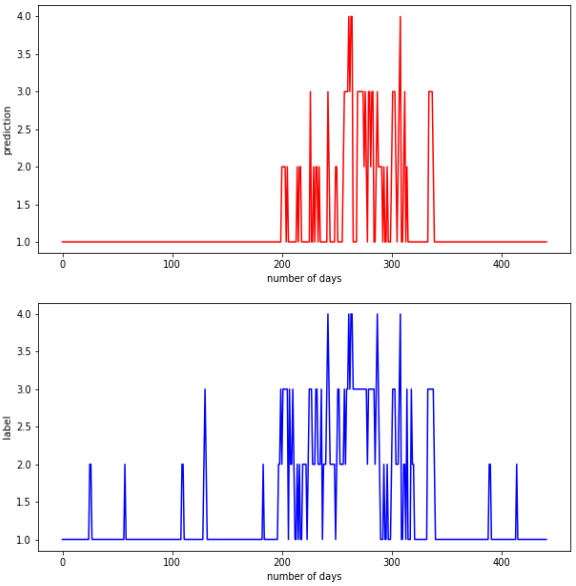
## SVC (Permutation Importance)




**[Figure 3]**

- Figure 3 displays a binary classification ( Rain and No Rain).
- The obtained accuracy is 91.17%.

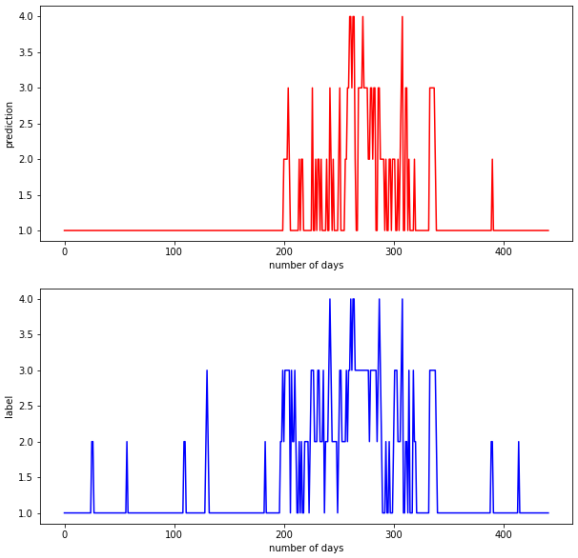
## Logistic Regression




**[Figure 5]**

- Figure 5 depicts four classifications ( No Rain, Drizzle, Moderate Rain, and Heavy Rain ).
- The obtained accuracy is 81.45%.

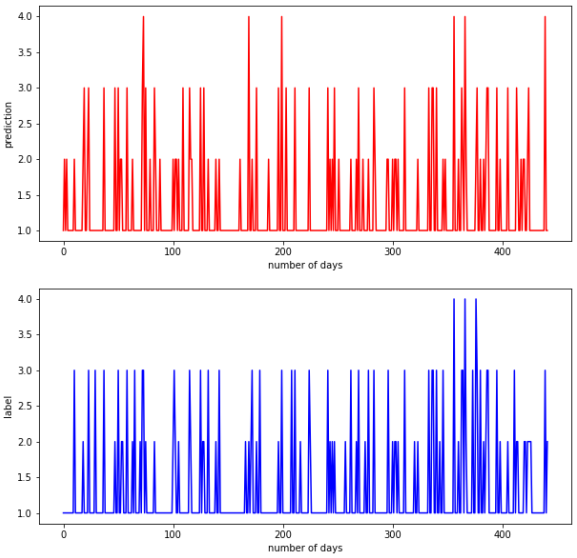
## SVC (Chi-square)




**[Figure 4]**

- Figure 2 depicts four classifications ( No Rain, Drizzle, Moderate Rain, and Heavy Rain ).
- Obtained Accuracy is 82.13%.

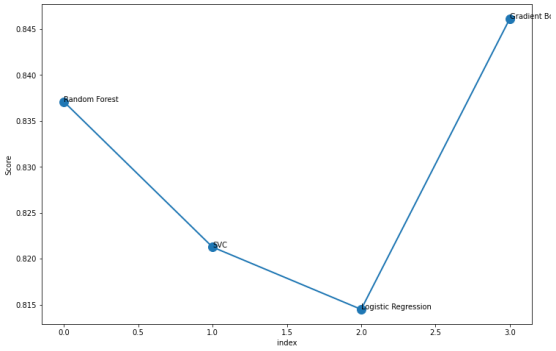
## Gradient Boosting




**[Figure 6]**

- Figure 6 depicts four classifications ( No Rain, Drizzle, Moderate Rain, and Heavy Rain ).
- The obtained accuracy is 84.61%.

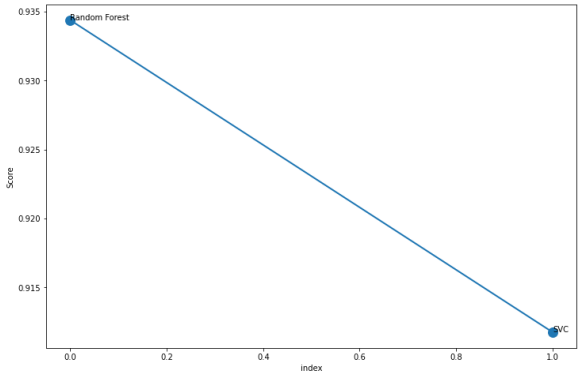
**Comparison of the four classifiers for the four rain classifications**




**[Figure 7]**

- Figure 7 depicts the accuracy-wise comparison of the four classifiers ( Random forest, SVC, Logistic regression, and Gradient boosting) for the four classifications (No Rain, Drizzle, Moderate Rain, and Heavy Rain).

**Comparison of the two classifiers for the two classifications**




**[Figure 8]**

- Figure 8 depicts the accuracy-wise comparison of the two classifiers ( Random forest and SVC) for the two classifications (Rain and No Rain).

# V. Conclusion

**The accuracy of various classifiers such as Random forest, SVC, logistic, and Gradient boosting has been analyzed in our article. The default accuracy of the random forest classifier (Before applying the permutation importance) is 92.7%. The accuracy improves to 93.4% after applying the permutation importance feature selection. Whereas, the default accuracy of the Support vector classifier (Before applying the permutation importance) is 86%. The accuracy improves to 91% after applying the permutation importance. As a result, the permutation importance feature selection helped us improve the model's accuracy. We discovered from the comparison graphs that training a model for only two classifications (i.e Rain and No-Rain) gradually increases the model accuracy. However, if we want a more detailed outcome and we increase the number of classifications (No rain, Drizzle, Moderate rain, and heavy rain ), the obtained accuracy is in the range of 80%~85%. Furthermore, in the case of four classifications, gradient boosting achieves the highest accuracy. Random Forest is the second best model in terms of accuracy.**

## References


[1] http://www.ijstr.org/final-print/jan2020/Prediction-Of-Rainfall-Using-Machine-Learning-Techniques.pdf

[2] https://towardsdatascience.com/obtain-historical-weather-forecast-data-in-csv-format-using-python-5a6c090fc828

[3] https://arxiv.org/pdf/1910.13827v1.pdf

[4] Thirumalai, Chandrasegar, et al. "Heuristic prediction of rainfall using machine learning techniques." 2017 International Conference on Trends in Electronics and Informatics (ICEI). IEEE, 2017.

[5] Geetha, A., and G. M. Nasira. "Data mining for meteorological applications: Decision trees for modeling rainfall prediction." 2014 IEEE International Conference on Computational Intelligence and Computing Research. IEEE, 2014