# Use of Statistical Downscaling Technique for Rainfall Modelling by Machine Learning Techniques


## Guided By: Dr. Meenu Ramadas

**Prof. of Civil Engineering, IIT Bhubaneswar**

## Soumyajit Manna

**UG Student of Civil Engineering**

## IIT Bhubaneswar



**Date of submission: 09.07.2021**

## Abstract:

Statistical Models are developed to discuss and predict the daily rainfall using downscaling. Different Supervised Machine Learning Regression models are used in the model to predict the rainfall. The provided dataset was on the latitude 20ºN and longitudinally in the region between 82.5ºN to 87.5ºN. It is observed that multi-layered algorithm is a better way than single layered in term of accuracy in any situation. And this study includes a brief discussion on each and every model used here. Accuracy is also determined by error estimation process. The analysis of observations is kept in the conclusion section. One of the major points of discussion is over-estimation in lower percentile data whereas underestimation in higher-percentile datapoints. The probable reason of this behaviour is discussed also along with difficulties.



***Key points:*** *Statistical models, prediction, rainfall, machine learning, regression*

# Contents

# 1. Introduction:

India's economy is mostly dependent on the production of sector I which is mostly agriculture. So, it very clear to depict that agriculture is one of the main pillars of survival. Though different technologies are improved and used in several cases, precipitation amount is highly important in some cases even. Where underground water is almost unavailable, the water required for daily using purposes is very difficult to avail. So, dependency on rainfall is higher in these areas.

Similarly, on the other hand, rainfall is very important in water harvesting which is a very important discussion in water resource engineering. Rainwater harvesting is one of the very important projects undertaken by the Government.

And in point of the view of disaster management as well, rainfall prediction is important. A huge amount of precipitation will destroy the crops as well as creates a flood. And generally, the rainfall pattern is a periodic way and the period is 1 year. And river flow level is also related to the amount of rainfall. Not only these but also drought can be predicted from the analysis of rainfall. So, it is very clear to conclude that rainfall is a very important phenomenon of climate that affects socio-economically human beings.

Daily precipitation is dependent on several factors of atmosphere. So, dependency on so many factors make it difficult to be predicted for meteorological department. Several approaches have been done till now to predict the rainfall. Still, it is very difficult to reach our destination. Machine Learning techniques/ algorithms can provide some comfort for prediction.

And, there is another challenge for localized prediction is local data collection. As meteorological stations cannot be established densely to collect data at least for every standard coordinate, downscaling method will be helpful to analyse large-scale low-resolution data to map high resolution data. In this method, primarily dataset of some boxes was used to predict rainfall amount for the anticipated places. The downscaling methods can be classified into two ways: dynamic downscaling and statistical downscaling. Dynamic downscaling involves deriving a high-resolution RCM (Regional Climate Model) from the coarse-grid GCM (General Circulation Model), statistical downscaling deals with developing a linear or non-linear relationship between the large-scale atmospheric variables (predictor) to the small-scale surface variable of interest (precipitation in our case). The statistical downscaling models are based on the major assumptions that regional climate is largely affected by the global-scale circulation patterns (von Storch 1995, 1999) and the relationship between the predictor and the predictand variables is invariant under future climate scenarios.

This paper aimed at using statistical downscaling along with statistical model for prediction of Rainfall (precipitation) amount.

## 2. Study Area:

To implement the goal of the project the 20º N latitude has been selected and longitudinally from 82.5ºE to 87.5ºE was taken as study area. This area is part of India, more specifically in the state of Odisha and its surroundings. The region has been marked by a box in the attached figure 1.

*Figure 1: Region selected for the study.*
*Source: Google Earth*

Geographically, this region is a combination of land and water. The land portion is part of Mahanadi basin and the water portion is named as Bay of Bengal. Bay of Bengal and its coastal area (Coastal area selected for study named as Utkal Plain / Northern Circas) is a very cyclone prone zone. In the parallel of coastal area, a discontinuous range of mountain named as Eastern Ghats is presents and behind of this Chottanagpur Plateau (Discreted plateau) region is present. So, the region consists of variety of geographical phenomenon.

# 3. Data:

Data sets has been used for the study has primarily two sets. One set contains some datasets named as BOX and others as Place. There are provided 3 Boxes for the coordinates with coordinates respectively 82.5º E, 85º E and 87.5º E (all are on 20ºN latitude). Let's name then A, B and C respectively.

And it contains places as a point of coordinates. On the 20ºN Latitude 82.5ºE, 83.5ºE, 84.5ºE, 85.5ºE and 86.5ºE. These are named as Place 1 to Place 5, respectively, from left to right.

As distance between two consecutive coordinates on a same latitude separated by 1º is approximately 100km so, Place 1 to 5 can be considered as a region of 100km X 100 km region and BOX A, B & C as 250km x 250km.

These datasets (BOXes) consist of 31 variables specified by NCEP those are specified in table 1.

| Variable | Description |
|---|---|
| *dswr* | Direct shortwave radiation |
| *lftx* | Surface lifted index |
| *mslp* | Mean sea level pressure |
| *p_f* | Near surface geostrophic airflow velocity |
| *p_u* | Near surface zonal velocity component |
| *p_v* | Near surface meridional velocity component |
| *p_z* | Near surface vorticity |
| *p_th* | Near surface wind direction |
| *p_zh* | Near surface divergence |
| *p5_f* | Geostrophic airflow velocity at 500 hPa height |
| *p5_u* | Zonal velocity component at 500 hPa height |
| *p5_v* | Meridional velocity component at 500 hPa height |
| *p5_z* | Vorticity at 500 hPa height |
| *p5th* | Wind direction at 500 hPa height |
| *p5zh* | Divergence at 500 hPa height |
| *p8_f* | Geostrophic airflow velocity at 850 hPa height |
| *p8_u* | Zonal velocity component at 850 hPa height |
| *p8_v* | Meridional velocity component at 850 hPa height |
| *p8_z* | Vorticity at 850 hPa height |
| *p8th* | Wind direction at 850 hPa height |
| *p8zh* | Divergence at 850 hPa height |
| *Pottemp* | Potential temperature |
| *Pr_wtr* | Precipitable water |
| *prec* | Precipitation total |
| *p500* | 500 hPa geopotential height |
| *p850* | 850 hPa geopotential height |
| *r500* | Relative humidity at 500 hPa height |
| *r850* | Relative humidity at 850 hPa height |
| *rhum* | Near surface relative humidity |
| *shum* | Near surface specific humidity |
| *temp* | Mean temperature at 2 m |

*Table 1: NCEP variables and definitions*

Among these variables related 16 variables are used for the analysis.

Source of this dataset: Indian Meteorological Department (IMD) daily rainfall at 2.5 degree resolution during 1948 to 2017.

# 4. Methodology:

This project consists of majorly two parts: Machine Learning algorithms and Statistical Downscaling.

## 4.1 Statistical Downscaling:

Downscaling is the general name for a procedure to take information known at large scales to make predictions at local scales. Downscaling is a very much preferred way than upscaling which provides unnecessary blurriness in data.

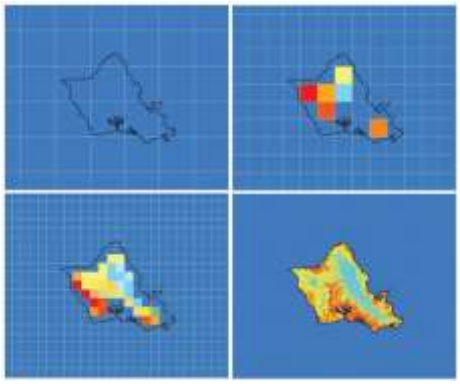


*Figure 2: downscaling (the island of Oahu)*
*Source: Pacific Islands Regional climate assessment (pirca)*

Let say, picture at the top left corner of figure 2 (the island of Oahu) as A, top right B, bottom left C and bottom right D. To validate above statement, let's look into figure 2. In image A, the girds are comparatively larger than in figure C. Topographical features are being represented here. Each and every grids are coloured with colour code of the major feature presented in that square where multiple features are accumulated in a single box. When finer grids are provided more accuracy has been accumulated. So, the finer colour combination has been shown in the image D.

downscaling can be categorised into two class: Dynamic Downscaling and Statistical Downscaling. By GFDL, Dynamical downscaling refers to the use of high-resolution regional simulations to dynamically extrapolate the effects of large-scale climate processes to regional or local scales of interest. In statistical downscaling, high level atmospheric data is used for lower-level prediction. The major assumption (von Storch 1995, 1999) in this method is that regional Climate Change is highly affected by global-scale circulation patterns.

There are several methods for statistical downscaling: regression methods, weather pattern-based approaches, stochastic weather generators (source: Wikipedia). Regression model [5] consists of Support vector Machine which includes Support Vector Regression & Support Vector Classification and Multivariate Analysis which can be called as Multiple Linear Regression.

*Multiple regression* (Multiple Linear Regression) is a form of regression analysis in which the regression function establishes the relationship between one dependent variable y and more than one independent variables (generalised as n number of variables) ($x_1$, $x_2$, …, $x_n$). A linear regression equation is in the following form:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n \ldots\ldots\ldots \quad (1)$$

Parameters ($b_0$, $b_1$, $b_2$, …, $b_n$) are estimated using the least squares method. Mertler and Vannatta (2005) described this multiple regression in detail.

In this study, Multiple Linear Regression (MLR) which is very straight forward has been used for statistical downscaling. The detailed method and target objective has stated as following.

This consists of 2 steps – BOX areas determination and Place plotting. 250km X 250 km Area (G.D.) is considered for the each of the block and similarly, for each of the places it is 100km X 100 km. BOX areas need to plotted at first. Places also need to be plotted according to the scale and need to analyse in which BOX it is in else it is intersecting multiple boxes. For intersecting ones, it is required to consider all of the BOXes as predictands.

## 4.2 Machine Learning Algorithms:

Machine Learning is a scientific technique or approach of recent days in which the machine (the computer) is trained with some mathematical tools and algorithms to receive desired output from the input.

Machine Learning algorithms developed up until now, can be divided into these categories:

- Supervised learning
- Unsupervised learning
- Reinforcement Learning

*Supervised learning* is a machine learning technique for creating a function from training data. The training data consist of pairs of input objects (typically vectors), and desired outputs. The output of the function can be a continuous value (as in regression), or can predict a class label of the input object (for classification). The task of the supervised learner is to predict the value of the function for any valid input point after having seen a finite number of training examples. The main algorithms of supervised learning are: neural networks, nearest neighbour algorithm, decision tree learning, Random Forest and support vector machine (SVM) [6].

In our study, different supervised learning has been used for prediction. The algorithms are described here in brief.

*Multiple Linear Regression:*

(Discussed in the previous section)

*Decision Tree Regression:*

A decision tree is a tree where each internal node specifies a test on an attribute of the data in question and each edge between a parent and a child represents a decision or outcome based on that test.

Let's explain with a diagram.

Here in the figure 3, a decision tree has shown and it is very clear that how to reach till terminal nodes to get some decision.

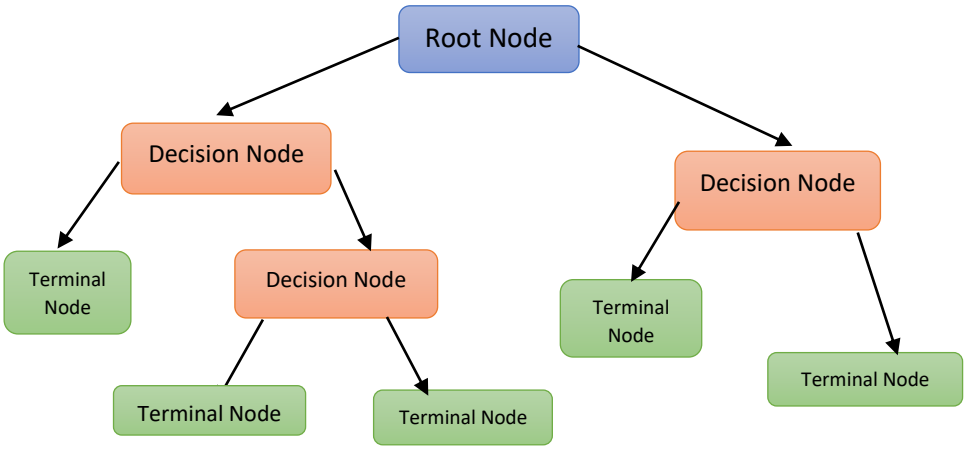



*Figure 3: Decision Tree Algorithm*

A decision tree is arriving at an estimate by asking a series of questions to the data, each question narrowing our possible values until the model get confident enough to make a single

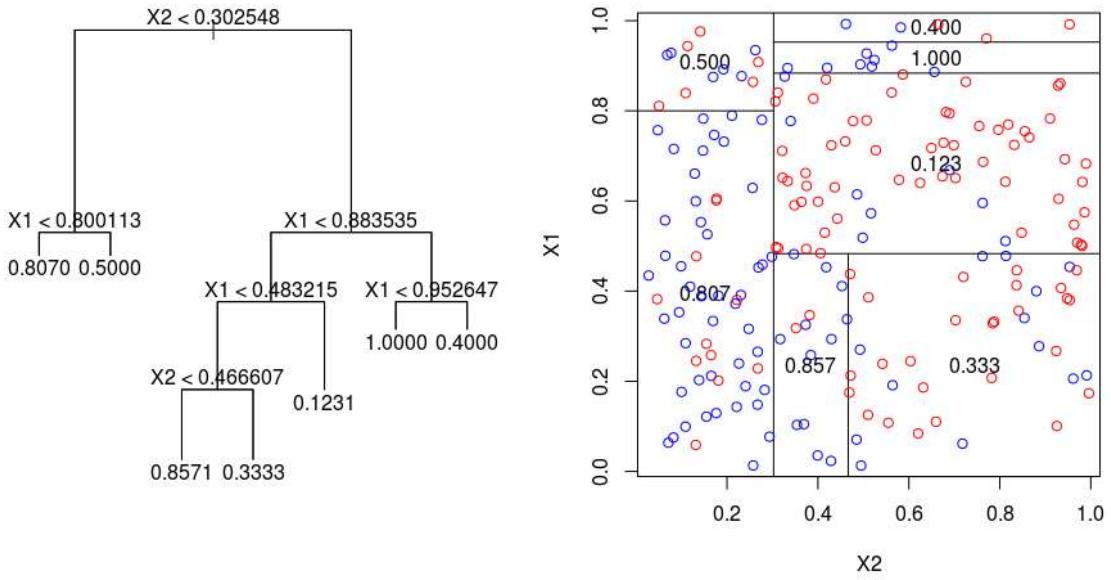



*Figure 4: Decision Tree Regression Example*

prediction. The order of the question as well as their content are being determined by the model. In addition, the questions asked are all in a True/False form. Let's discuss with an example [10].

For each True and False answer there are separate branches. No matter the answers to the questions, we eventually reach a prediction (leaf node). Start at the root node at the top and progress through the tree answering the questions along the way. So given any pair of X1, X2. As a supervised machine learning model, a decision tree learns to map data to outputs in what is called the training phase of model building [10]. The prediction will be an estimate based on the train data that it has been trained on.

The decision of making strategic splits heavily affects a tree's accuracy. Decision trees regression normally use mean squared error (MSE) to decide to split a node in two or more sub-nodes.

*Random Forest Regression:*

Random forests (RF) construct many individual decision trees at training. Predictions from all trees are pooled to make the final prediction; the mode of the classes for classification or the mean prediction for regression. As they use a collection of results to make a final decision, they are referred to as Ensemble techniques.

*Support Vector Regression:*

The support vector regression (SVR), a supervised learning approach proposed by Cortes and Vapnik [7], is a predictive algorithm. SVR employs a margin of tolerance that is dependent on the uncertainty in data. The main aim of SVR is to reduce the test data error by treating the problem as an optimisation problem that tries to locate the narrowest tube centred around the surface [8,9].

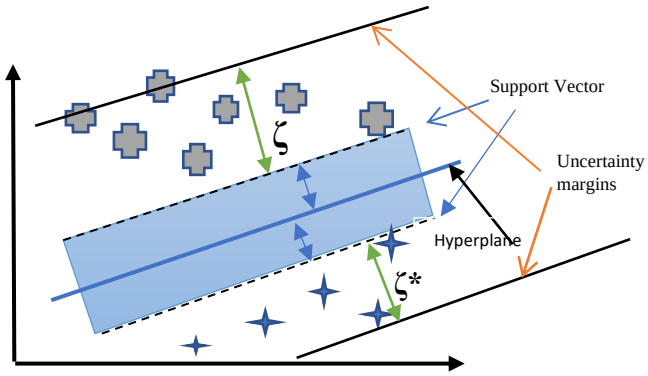



*Figure 5: Support Vector Regression (Linear)*

The distance between the Hyperplane (generalised for n dimension) and the support vector in figure 3 is equal to ε and that value is same for both sides even. In reverse, it should be mentioned that this hyperplane is considered such a way that distance between the each of the support vectors and Hyperplane is same and it is ε. And the distance between uncertainty margin and support vector is ζ and ζ* respectively.

To formulized the SVR problem we need to optimize the following expression:

$$\min_{w,b,\zeta,\zeta^*} \left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\zeta+\zeta^*)\right)\ldots\ldots (2)$$

Where the limit has been provided as following:

$$\begin{cases} y_i - w\phi(x_i) - b - \xi_i \le \epsilon \\ -y_i + w\phi(x_i) + b - \xi_i^* \le \epsilon \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad \ldots\ldots\ldots (3)$$

where y, φ(x), and w represents the estimated values, nonlinear mapping function, and matrix showing the position of the hyperplane (Fig. 3) for N data points. Furthermore, b and C are the bias function and tuning parameters for the SVM. The relationship between the input and desired values is described as follows:

$$y = \sum_{i=1}^{N}(\alpha_i + \alpha_i^*)\ Ker(x_i, x_j) + b \ldots. (4)$$

Where $\alpha_i$, $\alpha_i^*$ and $Ker(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)^T$ denote the Lagrange multipliers. The kernel function (Gaussian) is as follows:

$$Ker(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \ldots\ldots (5)$$

where γ is the kernel parameter, and C, γ and ε are the SVM tuning parameters. The algorithm can be easily adapted for data obtained from heritage buildings to predict values for unknown data points later.

*Multi Layered Perception:*

The artificial neural network (ANN) is a simplified mathematical model of a natural neural network. It is a directed graph where a vertex corresponds to a neuron and an edge to a synapse. Various ANN models have been proposed since its conception in the 1940's, but the multi-layer perceptron (MLP) one of the popular one [11].

A Multi-Layer Perceptron (MLP) or Multi-Layer Neural Network contains one or more hidden layers (apart from one input and one output layer). While a single layer perceptron can only learn linear functions, a multi-layer perceptron can also learn non – linear functions.

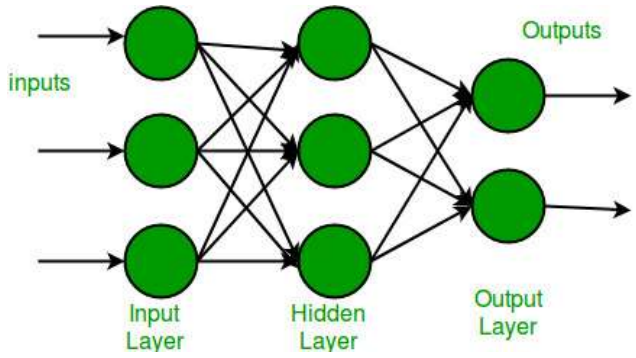
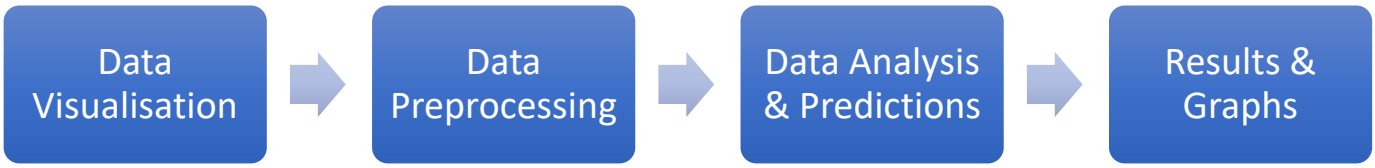



*Figure 6: Multi-Layered Perception*
*Source: Geeks for Geeks*

This neuron takes as input x1, x2, …., x3 (and a +1 bias term), and outputs f(summed inputs+bias), where f(.) called the activation function. The main function of Bias is to provide every node with a trainable constant value (in addition to the normal inputs that the node receives). Every activation function (or non-linearity) takes a single number and performs a certain fixed mathematical operation on it.

## 5. Data Analysis:

Analysis of the given data will be done in the following four steps.




I. Data Visualisation: Provided data can be visualised using two tools.

- Heatmap: Heatmap provides the correlation between all the factors of the given dataset. (Here all the input factors are used for Heatmap).
- Output visualisation: Only the predictable variable is used to plot to be visualised.

II. Data Pre-processing:

- Dimensionality Reduction: Here are some of the benefits of applying dimensionality reduction to a dataset:
  - Space required to store the data is reduced as the number of dimensions comes down
  - Less dimensions lead to less computation/training time
  - Some algorithms do not perform well when we have a large dimension. So, reducing these dimensions needs to happen for the algorithm to be useful
  - It takes care of multicollinearity by removing redundant features. For example, you have two variables – 'time spent on treadmill in minutes' and 'calories burnt'. These variables are highly correlated as the more time you spend running on a treadmill, the more calories you will burn. Hence, there is no point in storing both as just one of them does what you require
  - It helps in visualizing data. As discussed earlier, it is very difficult to visualize data in higher dimensions so reducing our space to 2D or 3D may allow us to plot and observe patterns more clearly.

  Any of the following methods can be followed to reach the destination:

  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Generalized Discriminant Analysis (GDA)

  PCA involves the following steps:

  I. Construct the covariance matrix of the data.
  II. Compute the eigenvectors of this matrix.
  III. Eigenvectors corresponding to the largest eigenvalues are used to reconstruct a large fraction of variance of the original data.
- Missing values checking: As a part of data pre-processing, it is required to check missing values. There are certain techniques can be followed to handle these types of glitches. So, if there are less numbers (ignorable) of data are missing in a column or more, those rows can be omitted from the huge amount of data pool. If there are significant number of data points are missing in a column or multiple, along with if those columns are necessary then two processes can be applied. The first one is to fill with mean (central tendency) or generating synthetic data points in the range of standard deviation and put in the vacant places.
- Feature Scaling: Sometimes it is noticed that values of some data points are really bigger and on the other hand large value handling is very difficult. So, features are scaled within some smaller limit. For this work, there are two methods: Min-Max Scaler and Standard Scaler.

- Min-Max Scaler: In this approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers. A Min-Max scaling is typically done via the following equation:
  $X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$ ..... (6)
  $X_{sc}$ denotes the variable after scaling, $X_{min}$ & $X_{max}$ signifies the minimum and maximum value of that particular dataset of particular variable.
- Standard Scaler: Standard Scaler follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance. In the presence of outliers, Standard Scaler does not guarantee balanced feature scales, due to the influence of the outliers while computing the empirical mean and standard deviation. This leads to the shrinkage in the range of the feature values.
- Robust Scaler: By using Robust Scaler, we can remove the outliers and then use either Standard Scaler or Min-Max Scaler for pre-processing the dataset. It scales features using statistics that are robust to outliers. This method removes the median and scales the data in the range between 1st quartile and 3rd quartile. i.e., in between 25th quantile ($Q_1$) and 75th quantile ($Q_3$) range. This range is also called an Interquartile range.
  The median and the interquartile range are then stored so that it could be used upon future data using the transform method. If outliers are present in the dataset, then the median and the interquartile range provide better results and outperform the sample mean and variance. The formula can be described as following:
  $X_{sc} = \frac{X_i - Q_1(x)}{Q_3(x) - Q_1(x)}$ ..... (7)

III. Data Analysis and Prediction:

- Different Machine Learning Algorithms are applied (Discussed in detail in Methodology.)
- Prediction Done.

IV. Results and Graphs:

- Error estimation: The most popular and usable method for error estimation is Mean Squared Error. Mean squared can be formulated as following and the formula is self-explanatory.
  $MSE = \frac{1}{N}\sum_{i=0}^{N}(\text{predicted} - \text{input})^2$ ..... (8)
- Graphical Representation

Each and every step should be done for prediction of each place.

A very important point needs to be discussed and finalise the effects of three BOXes on places.

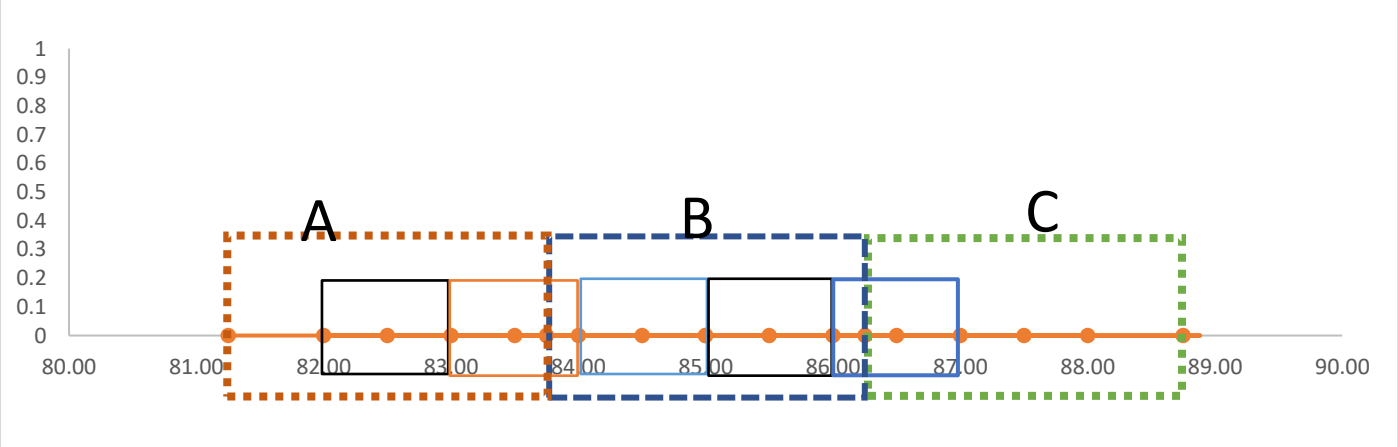



*Figure 7: Visualisation of BOX areas and places*

In the above figure, all the Boxes are drawn with dotted line where as the area for the places are drawn with the solid boxes. So, it is clearly depicted from the above figure that Place 1 in in fully inside by BOX A and will be affected by BOX A. Similarly, everything put in the table below:

| Places | Box(es) |
|---|---|
| Place 1 | A |
| Place 2 | A, B |
| Place 3 | B |
| Place 4 | B |
| Place 5 | B, C |

*Table 2: Places and Box(es) combination for prediction*

## *Place 1:*

I. Visualisation:

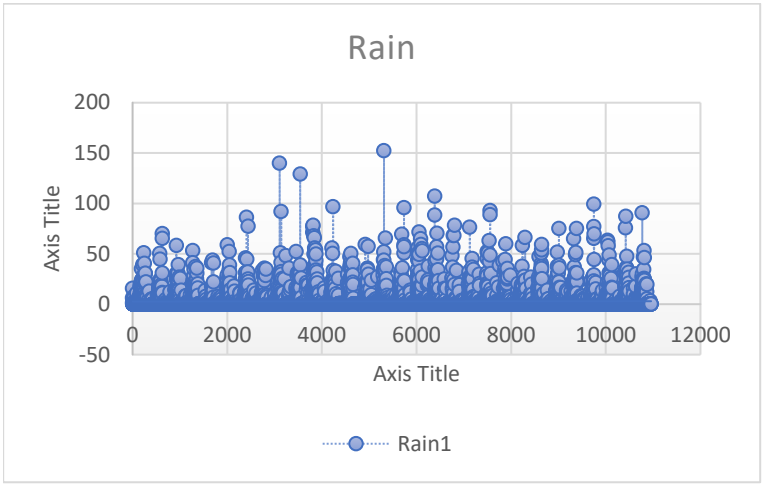



*Figure 8: Day-wise Rainfall graph for place 1 in duration of 1961-1990*

Here the plot for Rainfall has been provided. It is clearly shown that it is periodic in nature and some point of time, it is quiet high, so that time can be predicted as Monsson.

II. Data Pre-processing:

PCA:

The graph attached left side, contains that each dimension contains how much percentage of data. And it is clear that the first component contains around 80% data, second one 7.5% and third one 5% and fourth one 5% around and rest ones decreasing such a way. So, if we consider 4 elements to be handled, then we'll use 97.5% data which is reasonably good and 2.5% data is ignored here to make the computation easy.

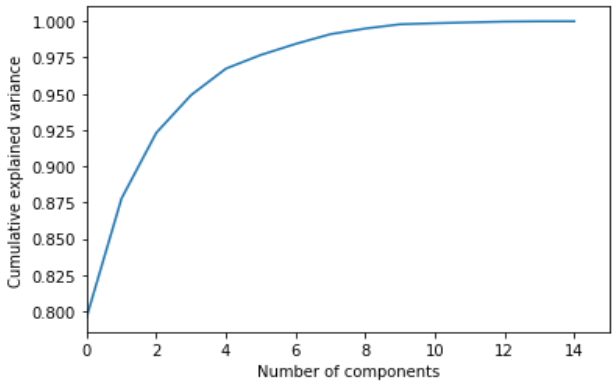



*Figure 9: Effect of data storing in each dimension*

Feature Scaling:

Min Max scaler is used here to scale the data between 0 to 1.

IV. Graphs & Results:

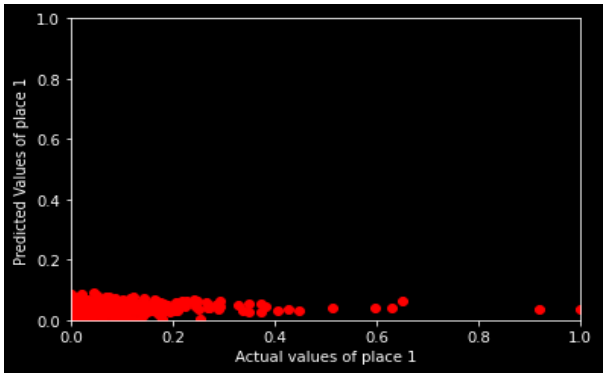



*Figure A: Prediction using LR*

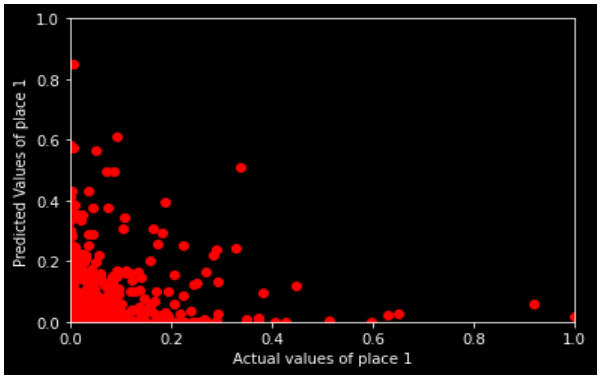



*Figure B: Prediction using DT*

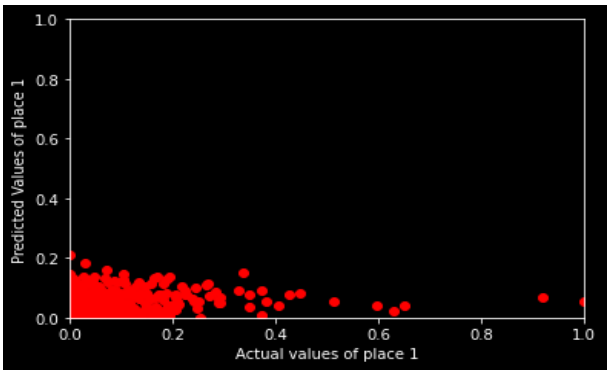



*Figure C: Prediction using RF*

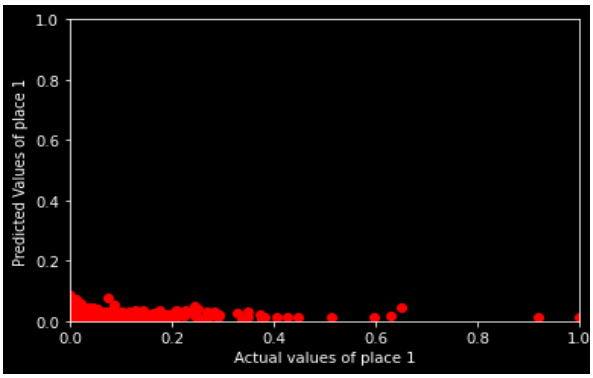



*Figure D: Prediction using SVR*

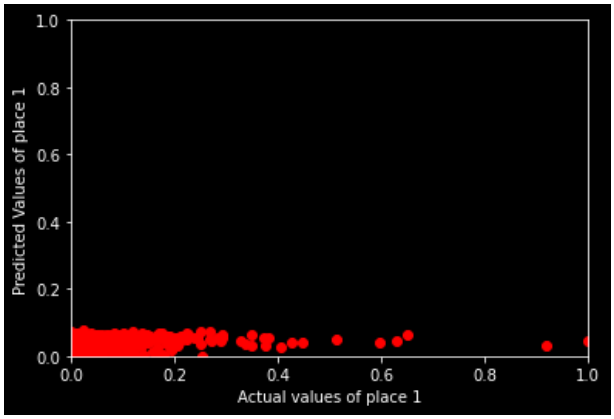



*Figure E: Prediction using ANN*

*Figure 10: Plots of predictions for place 1 using different Algorithms (Actual value and predicted value)*

*Figure A: Plot of predictions using Linear Regression*
*Figure B: Plot of predictions using Decision Tree Regression*
*Figure C: Plot of predictions using Random Forest Regression*
*Figure D: Plot of predictions using Support Vector Regression*
*Figure E: Plot of predictions using multiple Layered Perception Regression.*

Mean Squared Error values:

| **Algorithm** | **MSE Values** |
|---|---|
| Linear Regression | 0.00318206 |
| Decision Tree Regression | 0.00645830 |
| Random forest Regression | 0.00320269 |
| Support Vector Regression | 0.00347041 |
| Multiple Layered Perception (ANN) | 0.00311993 |

*Table 3: Mean Squared Error Values of Place 1 for different Algorithms*

*Place 2*:

I. Data visualisation:

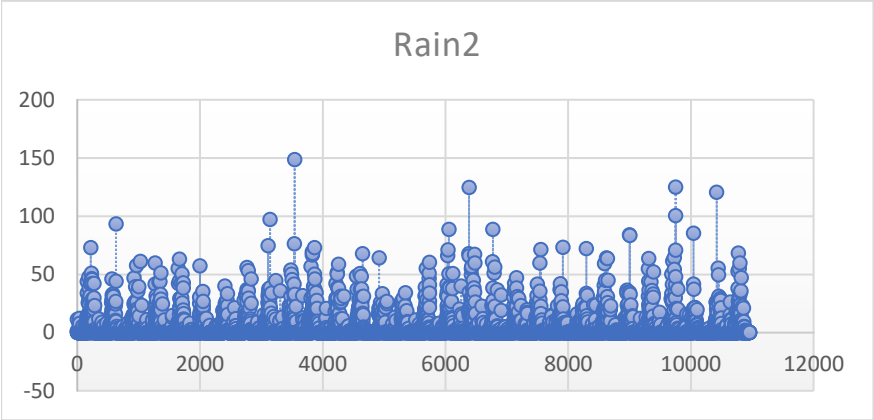



*Figure 11: Day-wise Rainfall graph for place 2 in duration of 1961-1990.*

II. Data Pre-processing:

Similar way of place 1, the data going to be transformed in lesser dimension and according to the image attached in the left image if any number is considered in between 5 to 7, it will be reasonably good and satisfactory because it is coving more than 95% of data.

So, 6 components are taken for final dimensions to do PCA.

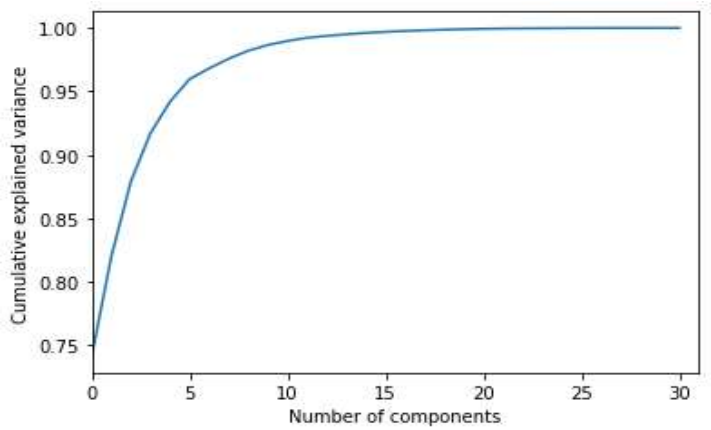



*Figure 12: PCA for place 2*

Feature scaling:

Min-Max scaling done.

Splitting of data:

III. Data analysis:

IV. Results and Graphs:

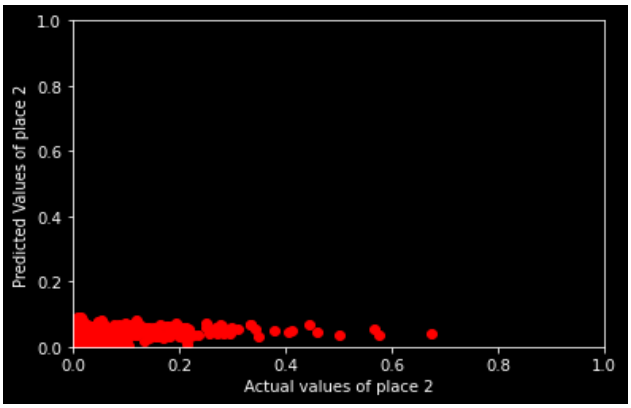



*Figure A: prediction using LR*

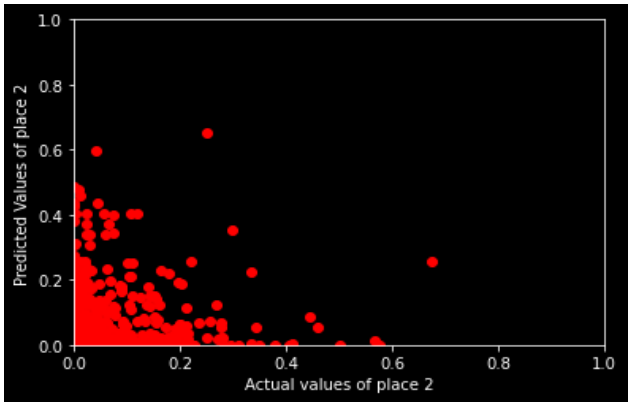



*Figure B: Prediction using DT*

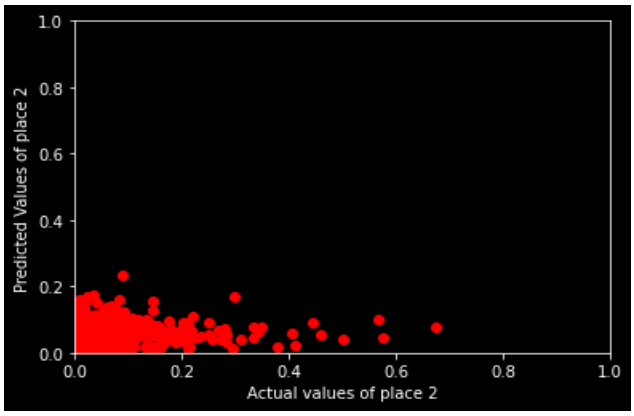



*Figure C: Prediction using RF*

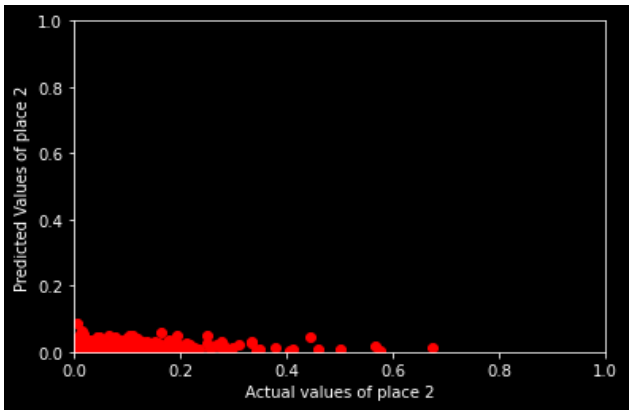



*Figure D: Prediction using SVR*

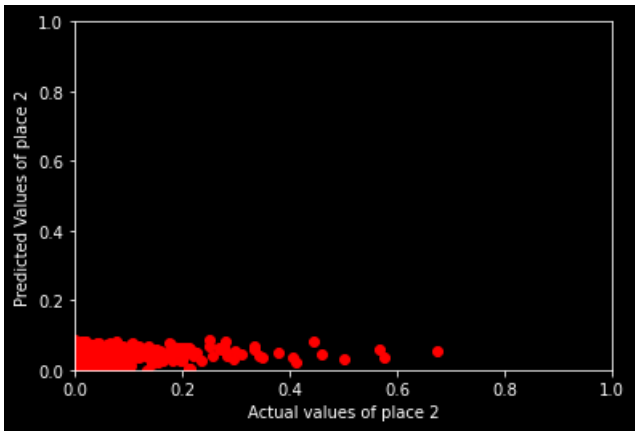



*Figure E: Prediction using ANN*

*Figure 13: Plots of predictions for place 2 using different Algorithms (Actual value and predicted value).*

*Figure A: Plot of predictions using Linear Regression*
*Figure B: Plot of predictions using Decision Tree Regression*
*Figure C: Plot of predictions using Random Forest Regression*
*Figure D: Plot of predictions using Support Vector Regression*
*Figure E: Plot of predictions using multiple Layered Perception Regression.*

Mean Squared Error Values:

| Algorithm | MSE values |
|---|---|
| Linear Regression | 0.00241792 |
| Decision Tree Regression | 0.00511280 |
| Random forest Regression | 0.00262601 |
| Support Vector Regression | 0.00271254 |
| Multiple Layered Perception (ANN) | 0.00234273 |

*Table 4: Mean Squared Error Values of Place 2 for different Algorithms*

*Place 3:*

I. Data Visualisation:

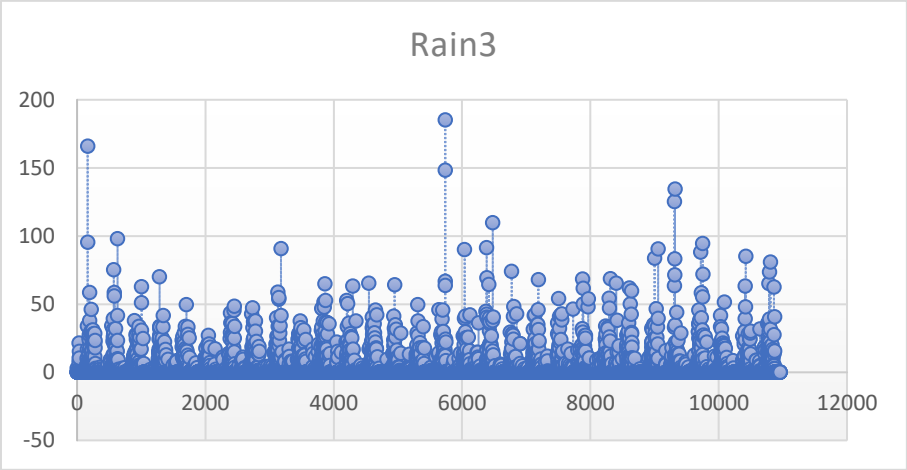



*Figure 14: Day-wise Rainfall graph for place 3 in duration of 1961-1990*

II. Data Pre-processing:

Similar way of place 1, the data going to be transformed in lesser dimension and according to the image attached in the left image if any number is considered in between 4 to 6, it will be reasonably good. The point to be noted that till 4$^{th}$ component, more than 95% data are covered and till 6$^{th}$ approximately 98%. So, it would be a good assumption if we take 4 or 5 components and for small amount of data (information), increasing dimensions unnecessarily would not be a very wise decision.

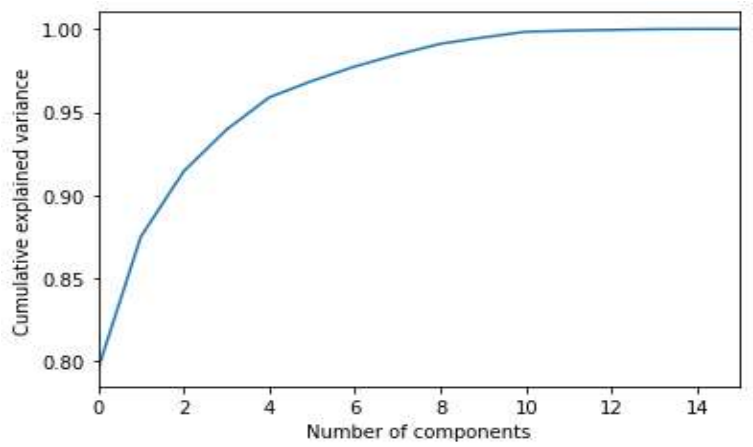



*Figure 15: PCA for place 3*

Feature scaling:

Min-Max scaling done.

Splitting of data:

Only 20% data is kept for model testing.

III. Data analysis:

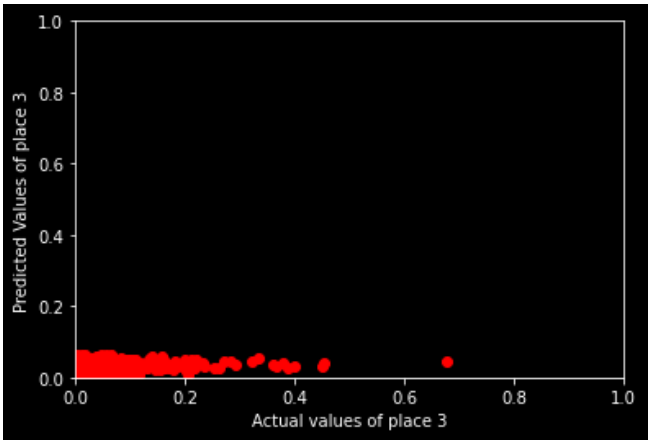



*Figure A: Prediction using LR*

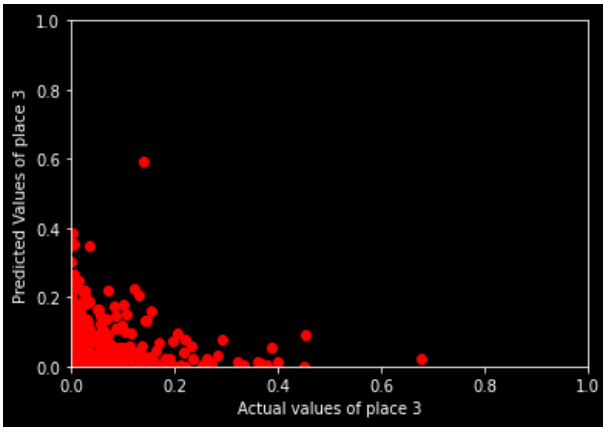



*Figure B: Prediction using DT*

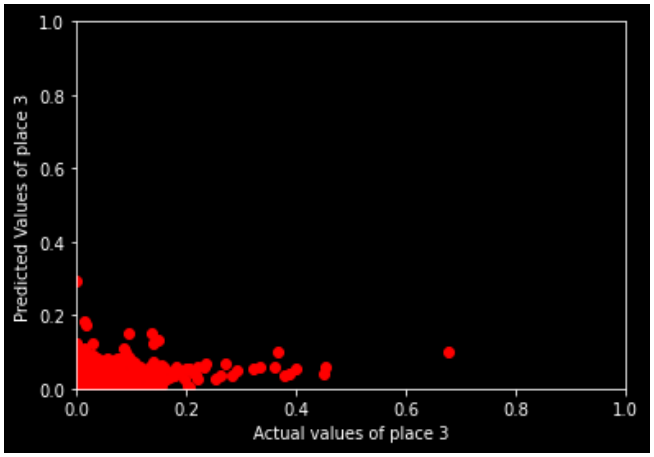



*Figure C: Prediction using RF*

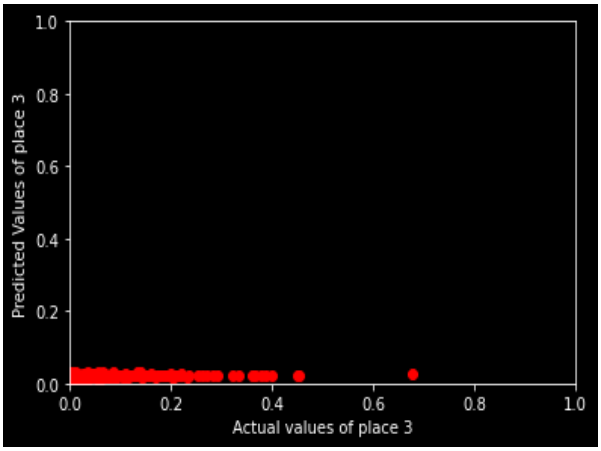



*Figure D: Prediction using SVM*

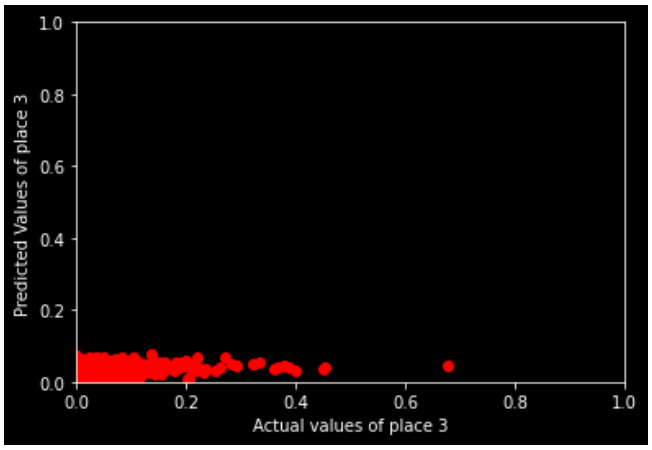



*Figure E: Prediction using ANN*

*Figure 16: Plots of predictions for place 3 using different Algorithms (Actual value and predicted value).*

*Figure A: Plot of predictions using Linear Regression*
*Figure B: Plot of predictions using Decision Tree Regression*
*Figure C: Plot of predictions using Random Forest Regression*
*Figure D: Plot of predictions using Support Vector Regression*
*Figure E: Plot of predictions using multiple Layered Perception Regression.*

Mean Squared Error Values:

| Algorithm | MSE |
|---|---|
| Linear Regression | 0.00162372 |
| Decision Tree Regression | 0.00299174 |
| Random Forest Regression | 0.00169688 |
| Support Vector Regression | 0.00182767 |
| Multiple Layered Perception (ANN) | 0.00159693 |

*Table 5: Mean Squared Error Values of Place 3 for different Algorithms*

Place 4:

I. Visualisation of Data:

Rainfall of Place 4 has been provided here in a graph to get an easy understanding by visual effect. Here, a period is similar to a year. And seasonal changes of Rainfall are also observed.

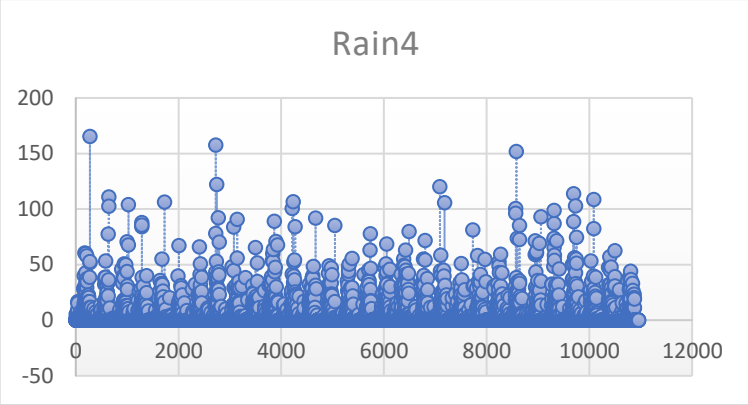



*Figure 17: Day-wise Rainfall graph for place 4 in duration of 1961-1990*

II. Data Pre-processing:

Similar as place 3, because both of them are in box B.

III. Data Analysis:

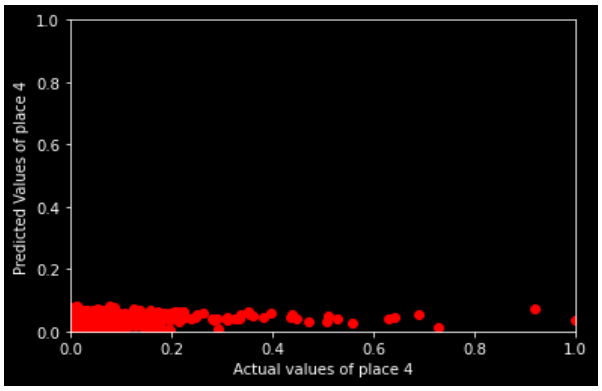



*Figure A: Prediction using LR*

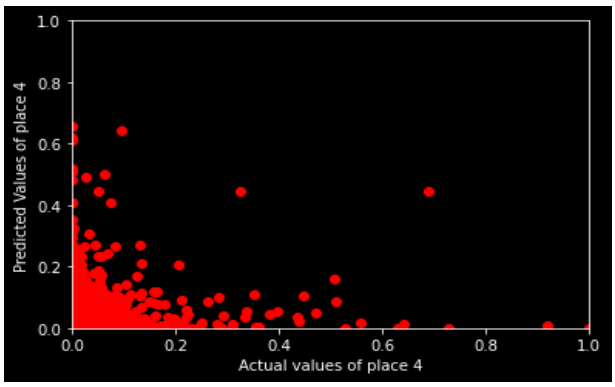



*Figure B: Prediction using DT*

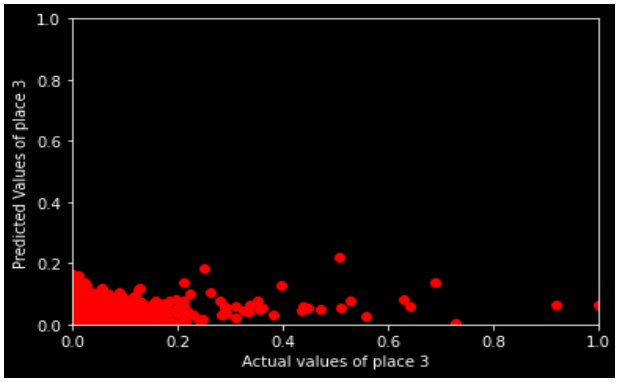



*Figure C: Prediction using RF*

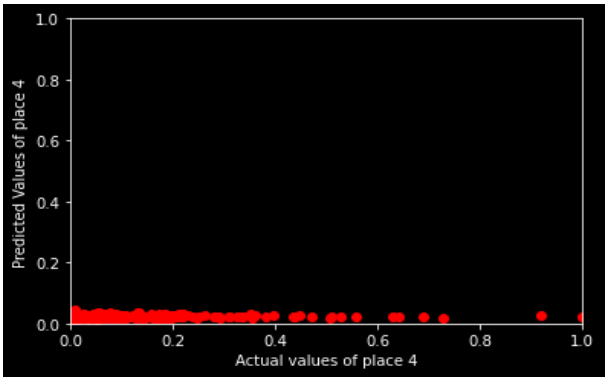



*Figure D: Prediction using SVR*

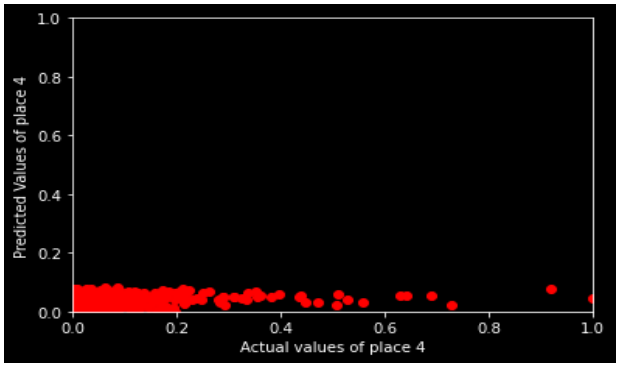



*Figure E: Prediction using ANN*

*Figure 18: Plots of predictions for place 4 using different Algorithms (Actual and predicted value).*
*Figure A: Plot of predictions using Linear Regression*
*Figure B: Plot of predictions using Decision Tree Regression*
*Figure C: Plot of predictions using Random Forest Regression*
*Figure D: Plot of predictions using Support Vector Regression*
*Figure E: Plot of predictions using multiple Layered Perception Regression.*

Mean Squared Error:

| Algorithm | MSE |
|---|---|
| Linear Regression | 0.00392106 |
| Decision Tree Regression | 0.00693513 |
| Random Forest Regression | 0.00396593 |
| Support Vector Regression | 0.00426360 |
| Multiple Layered Perception (ANN) | 0.00385146 |

*Table 6: Mean Squared Error Values of Place 4 for different Algorithms*

Place 5

I. Visualisation:

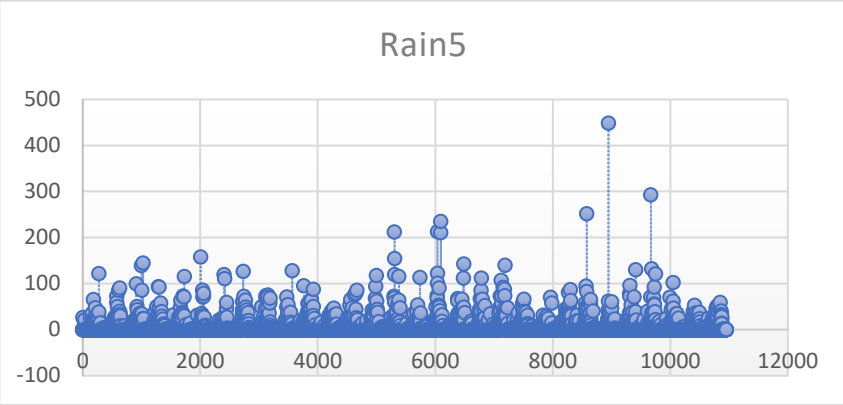



*Figure 19: Day-wise Rainfall graph for place 5 in duration of 1961-1990*

II. Data pre-processing:

PCA:

Here also done by similarly to fix the number of dimensions can be used for the analysis.

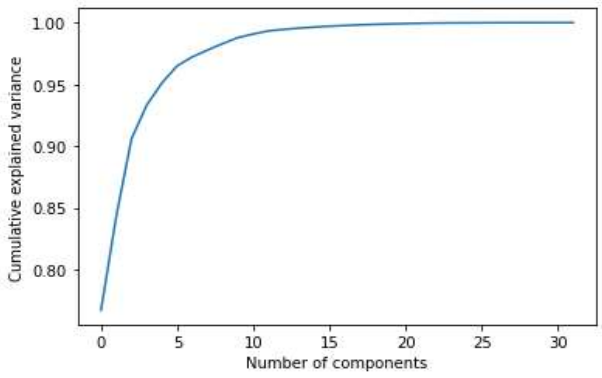



*Figure 20: PCA for place 5*

III. Data Analysis:

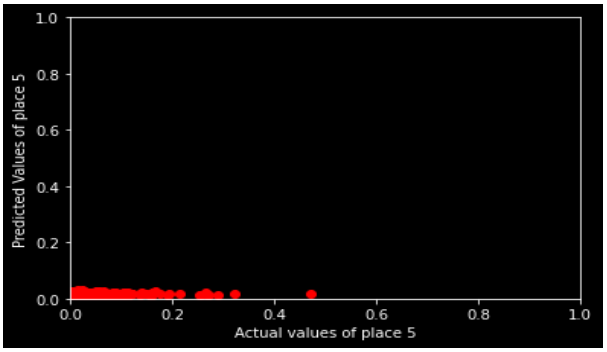



*Figure A: Prediction using LR*

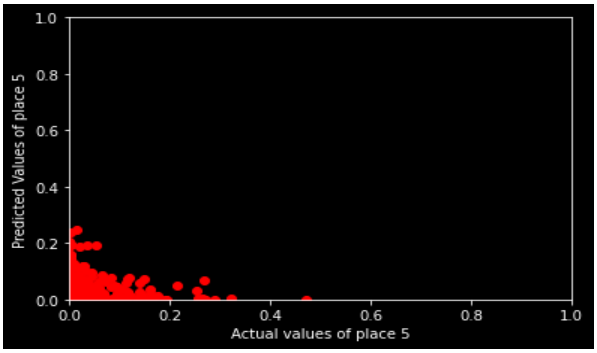



*Figure B: Prediction using DT*

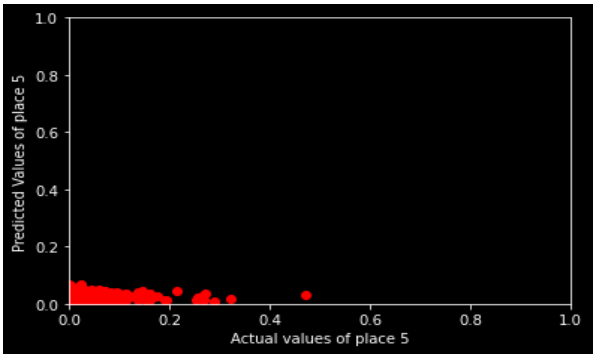



*Figure C: Prediction using RF*

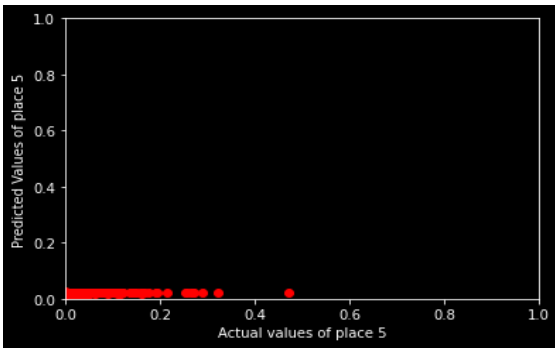



*Figure D: Prediction using SVR*

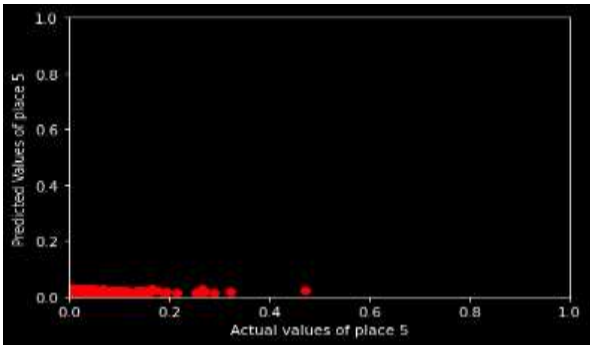



*Figure E: Prediction using ANN*

*Figure 21: Plots of predictions for place 1 using different Algorithms (Actual value and predicted value).*
*Figure A: Plot of predictions using Linear Regression*
*Figure B: Plot of predictions using Decision Tree Regression*
*Figure C: Plot of predictions using Random Forest Regression*
*Figure D: Plot of predictions using Support Vector Regression*
*Figure E: Plot of predictions using multiple Layered Perception Regression.*

Mean Squared Error:

| Algorithm | MSE |
|---|---|
| Linear Regression | 0.00068648 |
| Decision Tree Regression | 0.00117210 |
| Random Forest Regression | 0.00070355 |
| Support Vector Regression | 0.00089113 |
| Multiple Layered Perception (ANN) | 0.00067487 |

*Table 7: Mean Squared Error Values of Place 5 for different Algorithms*

## 6. Conclusion:

From the above analysis of rainfall prediction, the following points are concluded:

- Decrease in MSE, increase in accuracy.
- Neural Network provides a better result irrespective of situation.
- Dimensionality reduction provides an extra speed in simulation.
- Decision Tree cannot be used in any means as its MSE is too high than others.
- Linear regression provides better result than any other single layered algorithm.
- Multiple layered algorithms are better in terms of accuracy whereas Linear Regression is best in term of speed.
- Support vector regression is a highly time-consuming process which makes the system slower.
- In lower percentile data, over-estimation has been observed whereas for higher one (more than 80%) it is over-estimation.
- Input data must be scaled where there are some robust data points are present.
- Except Decision Tree algorithm (that is omitted in first), all other estimated vales (in general) are confined by a limit and it is 0.2 (scaled data).
- Irrespective of Places, MSE of any algorithm shows a general trend.

# 7. Limitations:

From the discussion of the previous sections along with conclusion, the following points need to be discussed.

- Over-estimation & Under-estimation: The each and every model, irrespective of places/zones, shows over-estimation in lower percentile data (under 25 percentile) and under-estimation in higher percentile data (over 90 percentile). The reason behind this can be suggested as Global Warming. In the diagram attached here, it is clear that in 1960 the 'Northern Hemisphere Annual Mean Temperature' was around 0ºC and over 30 years, in 1990, it became around 0.5ºC. So, mean temperature increased 0.5ºC. For annual maximum daily precipitation (Rx1day), the laws of thermodynamics suggest that the intensification would occur at a rate of about 6 to 7% per degree of warming roughly following the increase in atmospheric moisture content. Regionally, the thermodynamic response is expected to be substantially amplified or offset by a dynamic response leading to a higher intensification rates for instance over the tropics, and lower in the drier subtropics [12]. So, finally, it shows the tendency to make the rainfall similar in everyday throughout the year.

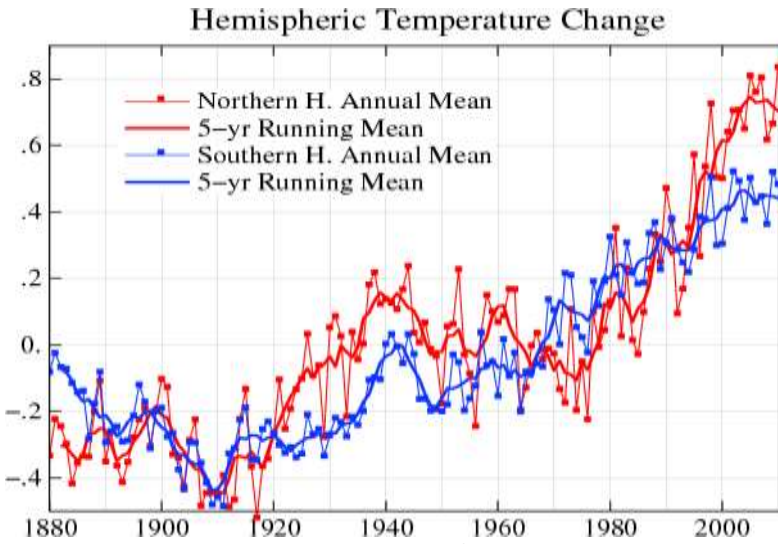


*Figure 22: Hemispheric temperature Change in last few decades*
*Source: data.giss.nasa.gov*

- Major Difficulty Inspection: On sub daily scales, the intensification rates were suggested to be larger due to highly dynamic nature of moisture-temperature interaction and latent heat release [12]. So, daily prediction is very difficult to reach the accurate state. And, due to the same reason, hourly prediction will be more difficult and on the other hand, seasonal prediction and monthly prediction will be more accurate than the previous two. So, in a nutshell, it can be concluded that smaller time period has a tendency of less accurate than larger time span.

## 8. Future Scope:

In this project, during prediction of Rainfall, some limitations have been faced and from those limitations there are some leads can be provided so that performance can be improved. Those scopes are enlisted below.

- For downscaling, in our study, where multiple Boxes should be considered, considered the effect with equal weight for each of them. In spite of equal weightage, further investigation can be done using the weight of the area covered by each Box for any places. This approach will look like more reasonable.

- Scaling was done using the algorithm of min-max scaler. As some peaks are found in some of the datasets where those points affecting the performance of the models, too much, those points can be omitted with Robust Scaling mechanism. So, very reasonable points can be filtered in this way.

# References:


1. Rainfall Prediction Using Machine Learning & Deep Learning Techniques, CMAK Zeelan Basha, Nagulla Bhavana, Ponduru Bhavya, Sowmya V

2. Aakash Parmar, Kinjal Mistree, Mithila Sompura Machine Learning Techniques for Rainfall prediction:A Review International Conference on Innovations in information Embedded and Communication Systems (ICIIECS).

3. Statistical downscaling of precipitation using long short-term memory recurrent neural networks, Saptarshi Misra · Sudeshna Sarkar · Pabitra Mitra

4. A REVIEW OF DOWNSCALING METHODS FOR CLIMATE CHANGE PROJECTIONS, by USAID, United States Agency International Development

5. Statistical downscaling of daily precipitation using support vector machines and multivariate analys, Shien-TsungChenPao-ShanYuYi-HsuanTang

6. Online Support Vector Regression, Francesco Parrella

7. C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

8. A.J. Smola, N. Murata, B. Schölkopf, K.-R. Müller, Asymptotically optimal choice of ε-loss for support vector machines, in: International Conference on Artificial Neural Networks, Springer, 1998, pp. 105–110.

9. A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.

10. https://gdcoder.com/decision-tree-regressor-explained-in-depth/

11. Rainfall Prediction Using Artificial Neural Networks, Sunyoung Lee, Sungzoon Cho, Patrick M. Wong

12. Models are likely to underestimate increase in heavy rainfall in the extratropical regions with high rainfall intensity, Aleksandra Borodina, Erich M. Fischer, Reto Knutti