ADVANCED ACADEMIC CENTER

A CENTER FOR INTER-DISCIPLINARY RESEARCH

2021-22

TITLE

ANNUAL AVERAGE RAINFALL PREDICTION

SUPERVISED BY

PADMINI KOUSALYA M , V BALASRI NITIN REDDY

GOKARAJU RANGARAJU
INSTITUTE OF ENGINEERING AND TECHNOLOGY
AUTONOMOUS

Advanced Academic Center

(A Center For Inter-Disciplinary Research)

This is to certify that the project titled

ANNUAL AVERAGE RAINFALL PREDICTION

is a bonafide work carried out by the following students in partial fulfilment of the requirements for Advanced Academic Center intern, submitted to the chair, AAC during the academic year 2021 - 22.

NAME	ROLL NO.	BRANCH
BANDI SRIVANI	21241A6611	AIML
BURLA AKASH	21241A04E5	ECE
DODDA ABHIRAM	21241A0511	CSE

NAME	ROLL NO.	BRANCH
GUBBA VENKATA SAI NITHIN	21241A05T4	CSE
GUGLOTH BINDU BHARGAVI	21241A66F3	AIML
PISHANGAL VENKATA RAMANA	21241A6652	AIML

This work was not submitted or published earlier for any study

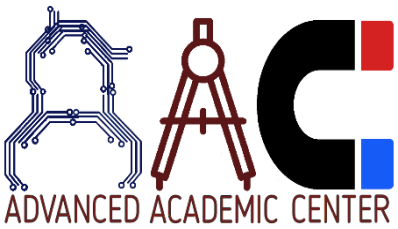
Dr/Ms./Mr.

Project Supervisor

Dr.B.R.K.Reddy
Program Coordinator

Dr.Ramamurthy Suri
Associate Dean,AAC

Index



ACKNOWLEDGEMENTS

We express our deep sense of gratitude to our respected Director, Gokaraju Rangaraju Institute of Engineering and Technology, for the valuable guidance and for permitting us to carry out this project.

With immense pleasure, we extend our appreciation to our respected Principal, for permitting us to carry out this project.

We are thankful to the Associate Dean, Advanced Academic Centre, for providing us an appropriate environment required for the project completion.

We are grateful to our project supervisor who spared valuable time to influence us with their novel insights.

We are indebted to all the above mentioned people without whom we would not have concluded the project.

Abstract

Rainfall is a major source of fresh water in many parts of our country. From watering the crops to hydroelectricity rainfall judges the yield. We store water in reservoirs and make a wise use of it. But heavy rainfall cause damage. The main problem that we are not ready to face challenges of rainfall is that rainfall patterns are unknown. If we can predict the patterns of rainfall we can optimize the storage of water and take precautions of potential problems that may arise in the future. How can we predict rainfall? Answer can be provided by modern computing techniques and data of rainfall in the past. One of the methodologies that we can use to input data and predict values is Machine Learning.

Introduction

Machine learning is a technology that is built upon computer science, statistical methods and probability. Machine learning techniques are coded into computer as algorithms using programming languages like Python, R and libraries like Skikitlearn, Tensorflow are used. The algorithms trained over a dataset are called Machine learning models [1].
Machine learning models are chosen according to the type of data available. The data available is termed as dataset. The dataset is used for analysis and should be understandable by machine learning models [2].To make data clean and uniform it is put through data preprocessing stage. In this stage missing values are filled, encoded, labelled if required [3]. After preprocessing data is analyzed by the machine learning model and is ready to output predicted values for given inputs. The methods of analyzation differ for model applied.

Dataset

The aim of the machine learning model is to predict annual average rainfall after training on the rainfall dataset. We require the model to take the labelled data which has more than one independent variables and give output.
Dataset contains rainfall data of Telangana state from 1901 to 2015. The independent variables are year, average rainfall in each month for the respective year and average temperature. The dependent variable, the factor to be predicted is annual rainfall in mm.

STATE	YEAR	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBE	OCTOBER	NOVEMBE	DECEMBE	AVG. TEM	AVG. RAINFALL
TELANGAI	1901	6.9	41.8	7.8	45.2	22	123.6	237.8	177.2	77.7	75.5	12.2	0	29.26	827.7
TELANGAI	1902	0	0	0.2	10.7	7.3	52.4	146.3	142.8	190.5	41.7	31.2	7.3	28.89	630.4
TELANGAI	1903	12.9	4.6	0	9.9	40.7	99.2	505.2	246.7	191.9	155.8	15.5	1.1	29.27	1283.4
TELANGAI	1904	0	0	10.8	0.8	14.7	104.2	139.5	50	162.3	44.4	0	0	29.41	526.7
TELANGAI	1905	0	4.3	12.8	27.6	32.2	129.5	82.4	237.3	179.1	19.6	0	0	29.23	724.9
TELANGAI	1906	22.5	1.2	13.4	2.4	0.7	211.1	210.8	226.7	96.3	20.5	14.9	34.8	29.63	855.2
TELANGAI	1907	1	3.3	10.2	61.9	0.2	217.5	160.5	263.3	116.8	0.3	3.6	5	29.58	843.7
TELANGAI	1908	35.6	2.6	5.2	0.3	6.5	107.4	254.9	168.3	401.2	0.1	0	0.7	29.32	982.8
TELANGAI	1909	0.5	5.9	0.5	26.4	2.2	133.8	288.3	168.6	138.5	4.6	0	0.2	29.12	769.5
TELANGAI	1910	0	0	0	4.2	25	220.9	198.2	150.3	230.5	101.4	45.3	0	29.11	975.9
TELANGAI	1911	0	0	7.9	0.7	7.8	133.9	122.1	176.3	174.4	23	6	9.9	29.28	662
TELANGAI	1912	0	37.5	0	20.4	6.6	28.5	263.3	196.8	149.5	7.8	33.4	0	29.61	743.8
TELANGAI	1913	0	13.4	0	8	34.6	83.5	337.2	108.8	101.7	35.2	0	9.7	29.33	732.2
TELANGAI	1914	0	0	1.2	34.1	41.5	233.1	271.3	195.1	278.6	16.2	10.3	2	29.72	1083.5
TELANGAI	1915	15.4	8.6	77.8	18.8	26.7	165.6	140.4	236.2	186.8	122	23.6	0.1	29.55	1022

TELANGAN	1916	0	4.1	0	15.9	16.1	233.2	216.4	137	291.8	153.1	95.1	0	29.18	1162.6
TELANGAN	1917	0	65.5	33	34.1	52.7	184.1	267.5	180.1	275.6	106.7	6.6	0	29.14	1206
TELANGAN	1918	20.2	0	32.6	8.3	55.5	97.8	106.1	89.6	96.9	6.4	7.5	33.9	29.32	554.7
TELANGAN	1919	5.5	39.2	31.3	35	23.6	158.8	139.4	115.3	129.9	130.6	78.5	7.4	29.23	894.6
TELANGAN	1920	7.4	0	1.7	24.4	31.4	62.8	138.8	66.5	78.9	24.9	0	0	29.55	437
TELANGAN	1921	7.1	0	0.1	5.4	5.5	200.9	296.7	161.1	240.7	68	17.9	0	29.46	1003.3
TELANGAN	1922	55.3	0	0	6.6	50.1	69.2	184.3	158.3	170.2	23	50.7	0	30.18	767.5
TELANGAN	1923	3	5.3	23.3	10.7	18	43.5	222.8	80.3	334.4	44	1	0	29.58	786.2
TELANGAN	1924	37	0	0	9.2	34.6	62.8	124.7	220.2	288.2	51.6	116.1	0	29.05	944.5
TELANGAN	1925	0	0	0	49.8	97.8	168.1	266.2	239.2	148.1	187.5	8.1	3.1	29.7	1167.9
TELANGAN	1926	46.1	2	24	48.8	24	74.7	227.3	205.1	70.5	41	0	0	29.81	763.5
TELANGAN	1927	9.6	11	4.6	2	14.3	248.3	266	158.5	142.4	59.9	89	0	29.75	1005.5
TELANGAN	1928	0	15.8	47.3	4.5	27.9	143.9	248.6	191.6	215.2	104.1	0	7.4	29.99	1006.4
TELANGAN	1929	5.3	64.8	0	21.7	2.8	146	115	142.2	232.5	35.9	0	24.5	30.23	790.5
TELANGAN	1930	0	14.7	5.4	13.5	1.7	184.9	131.1	144.7	215.6	60	64.3	0	29.75	835.8
TELANGAN	1931	0.3	5.1	4	13.3	25.9	187.3	368.6	155.9	288.1	107.4	25.5	1.6	29.79	1183.1
TELANGAN	1932	0	40.3	1.3	17.6	25.8	119.7	300.4	179	172.9	32	60.9	0	29.6	949.9
TELANGAN	1933	12.9	17.4	15.7	25	63.4	332	243.4	226.7	262.2	128.9	24	44.5	30.06	1396.3
TELANGAN	1934	0.1	0	0.1	18.3	1.2	111.7	296.9	280.2	162.3	15.2	42	0	29.84	927.9
TELANGAN	1935	16.2	3.4	1.4	42.2	2.7	115.1	309.7	162.9	198.3	43.5	2.6	0.5	29.64	898.5
TELANGAN	1936	2.5	79.1	15.3	8.4	73.7	200	268.9	220.9	128.4	96.4	48.2	15.1	30.3	1157
TELANGAN	1937	0	33.2	28.6	105.6	3	77.8	285.8	102.4	153.6	72.1	0.9	3	30.13	866.1
TELANGAN	1938	0	41.3	14.2	16.6	27.8	226.3	331.6	256.5	201.9	59.2	0	0	29.82	1175.3
TELANGAN	1939	0	0	22	16.7	3	73.5	173.2	202.5	107.1	130.3	11.8	0	28.76	740
TELANGAN	1940	0	0.5	11.4	33.6	91.4	128.5	345.8	281.5	94.7	64.9	22.2	8.1	29.81	1082.6
TELANGAN	1941	16.6	3.3	4.8	4.8	17.5	121.4	87.9	149.9	137.5	49.6	0.7	0.4	29.72	594.4
TELANGAN	1942	0	23	0.2	31.7	30.9	160	250.1	277.3	107.4	20.9	3.9	1.6	29.9	906.9
TELANGAN	1943	15.2	0.6	1.7	34.4	26.9	180.4	179.3	133.4	262.2	62.7	2.6	0	31.63	899.3
TELANGAN	1944	0	13.7	82.3	5.4	15.7	114.5	224.1	142.4	207.7	131.9	22.7	0	31.42	960.5
TELANGAN	1945	4.3	0	0	39	11.2	144.5	332.1	238.4	159.1	94.1	4.8	0	30.12	1027.4
TELANGAN	1946	0	15.9	5.2	21.7	14.5	152.1	287.2	149.5	85	30.7	57.5	3.6	28.96	822.8
TELANGAN	1947	13.4	18	3.3	1.8	10.8	59.9	273.3	315.5	248.3	35.3	16.2	9.7	29.22	1005.5
TELANGAN	1948	3.7	4.6	5.1	20	15	71.4	239.3	165.4	242.5	25	123	0	28.47	915
TELANGAN	1949	0.5	0	2.4	13.8	67.9	152	335.5	173.6	198.1	94.6	0.2	0	28.49	1038.6
TELANGAN	1950	0	34.7	14.2	0.2	11.6	78.4	222.7	118.5	254.1	21.8	4.7	0.5	28.3	761.3
TELANGAN	1951	0	0	56.1	20.1	31.1	116.8	356	151.7	103.9	64.4	0.7	0	28.73	900.8
TELANGAN	1952	0.5	16.4	0.8	12.4	32.5	80.1	205.1	148.9	137.7	88.3	0.2	6.9	28.65	729.7
TELANGAN	1953	5.1	0	3	26.8	0.7	224	200.9	364.5	269.5	175.8	1.7	0	28.83	1272.1
TELANGAN	1954	0	0	13.2	4.3	14.4	105.3	300.4	192.5	252.9	44.2	0	1.2	28.38	928.4
TELANGAN	1955	1	0	2.9	9.8	27.8	230.3	259.9	300.1	249.7	116.2	5.5	0	28.53	1203.2
TELANGAN	1956	0.7	1.5	1.9	5.9	52.9	220.1	393.8	133.3	156.9	94.5	59.1	0	28.62	1120.5
TELANGAN	1957	0	7.7	23.7	33.4	18.6	157.9	206.9	358.5	100.9	75.6	0	0	28.95	983.2
TELANGAN	1958	2.1	4.9	10	24.7	14.1	72.9	407.5	351.7	135.2	69.8	20.9	0	28.67	1113.9
TELANGAN	1959	1	0.4	0	4	16.3	163.7	403.2	348.3	212.8	68.6	4.2	1.7	28.66	1224.1
TELANGAN	1960	0.7	0	32.9	4.1	21.9	223.9	203.6	82.1	197.5	59.4	12	1.4	28.94	839.5
TELANGAN	1961	0.8	6.1	2.7	11.8	46.1	142.2	368.3	209.2	144.5	188.3	26.4	0.2	28.82	1146.5
TELANGAN	1962	0.2	25.1	2.1	47	20.3	84.5	270.5	264.1	289.6	62.5	38.6	34.2	28.11	1138.7
TELANGAN	1963	0	0.3	4	35.5	9	179.5	214.7	372.7	103.3	91.2	0	0	28.66	1010.1
TELANGAN	1964	0	0.2	5.5	2.1	0.7	109.1	186.3	232.3	302.4	41.7	7.2	0	28.66	887.5
TELANGAN	1965	2.7	0.8	4.4	6.8	2.7	124.3	318.6	152.2	166.2	0.1	0	0	28.76	778.9

Workflow

1. Required libraries are imported.
2. The rainfall dataset is imported using pandas.
3. The dataset has less nan values which cannot be processed by the machine learning model.
4. The dataset is divided into X and Y.
5. Using simple imputer nan values are replaced with median values. The strategy to be used is set to be median as accuracy is found to be better when median is used in hit and trail method.
6. The dataset has Strings as ‘Telangana’ and machine learning models do not accept any strings. So using OneHot encoder the strings are encoded.
7. Now the dataset is divided into training and testing sets. The datavalues to be included in either of the sets are chosen in random manner and 20% of the values are in test set while training set has 80% of the values.
8. Using the training set the model is trained. The model for rainfall dataset is Multiple Linear regression as the dataset has more than one labelled variables.
9. To predict the temperature another model is used in the same program.
10. Year is taken as a input from the user and using the year temperature is predicted.
11. The predicted temperature value is used as an input value for the rain fall prediction.
12. The temperature dataset has only one dependent variable the year and one independent variable, average temperature of the respective year.
13. The temperature dataset is imported using Pandas, divided into X and Y. There are no strings and no nan values in the dataset, so imputing and encoding parts of data preprocessing are skipped.
14. The model used for temperature prediction is Linear Regression.
15. The model is trained and the output value, the predicted temperature of the given year is returned to the rainfall_prediction function.
16. Gradio is used to create user interface to take the input and display prediction.

Input

1. First value of input is name of state which is a string and not accepted. So the name of the state is mapped to an integer and used instead.
2. Second value of input is year which is to be provided by the user.
3. The next 12 values occupy the average rainfall values of 12 months of the year. It is not possible for the user to provide these values so the values are replaced by mean values of all the years for the respective months as constants. It increased the error range, the range is -100 to 100.
4. The last value of the input is average temperature of the year which is predicted and replaced by the program itself.

Accuracy

Mean absolute error of 0.08 and R2 score of 0.9999996684

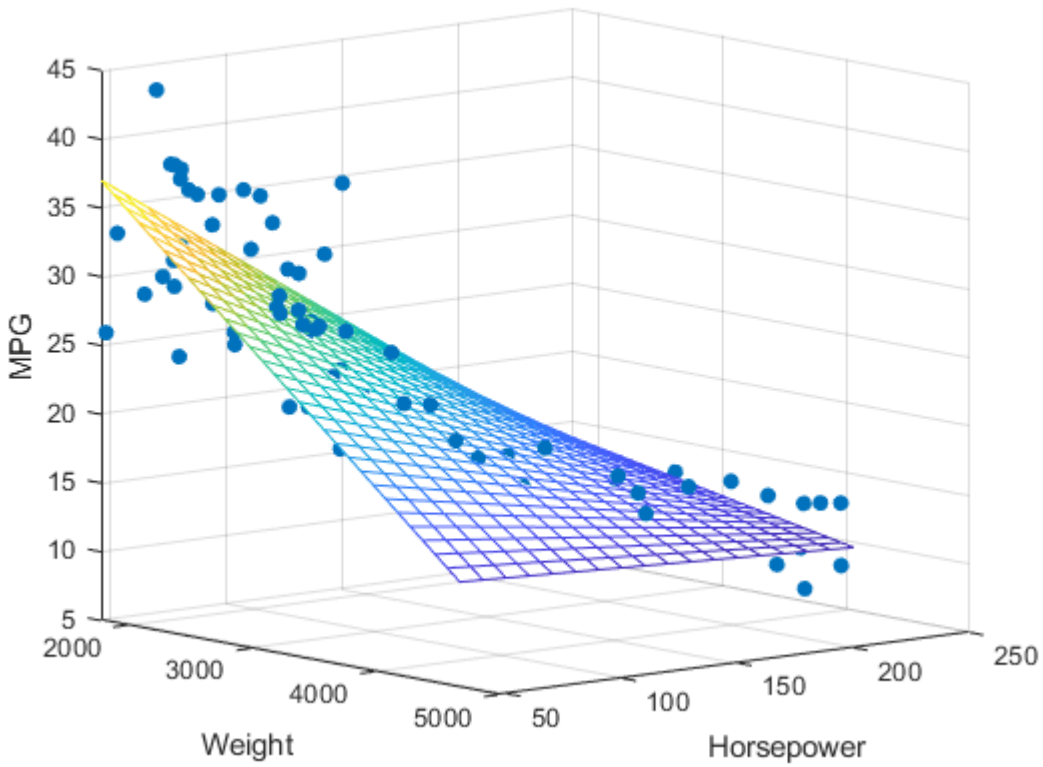
MULTIPLE LINEAR REGRESSION [5][6]

In a linear regression an independent variable and a dependent variable exist. A line is graphed such that the sum of distances of the line from all the points is minimum. The line is used to predict new values for any given input. In a Multiple Linear Regression we deal with more than one independent variables. A graph can be visualized for two independent variables and for more it is beyond human imagination.
A Multiple Linear regression model is designed by the equation

$$yi=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_p x_{ip}+\epsilon$$

where, for $i=n$ observations:
Yi = dependent variable
xi = explanatory variables
$B0$ = y-intercept (constant term)
Bp = slope coefficients for each explanatory variable
ϵ = the model's error term (also known as the residuals)

Multiple Linear Regression takes linear relationship between independent and dependent variables into consideration. The relationship is measured based on statistical methods. The independent variables should be independent of each other. Sometimes there may be some hidden relationships between the independent variables themselves. But they must not be too correlated to get accurate predictions.
To measure the accuracy of Linear Regression models coefficient of determination used is R2 score. R2 score lies between 0 and 1, the closer to one the higher is the accuracy of predictions. R2 score is measured based on correlation coefficients between independent and dependent variables. To test the efficiency of the model R2 score is measured for test values and predicted values.



An example graph depicting data-points and plane for multiple linear regression [7]

Code

1. Importing libraries

```
import numpy as nm
import pandas as pd
import matplotlib.pyplot as plot
import gradio as gr
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn import linear_model
```

Libraries used:

1. Numpy for array and matrix operations
2. Matplotlib for graphing
3. Gradio for user interface creation
4. Sklearn for linear and multi linear models.
5. Pandas for data importing.

2.Importing rainfall dataset

```
def rain_prediction(Year):

    dataset = pd.read_csv('raindata3.csv')
    X = dataset.iloc[:, :-1].values
    Y = dataset.iloc[:, -1].values
```

3. Imputing and encoding

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values=nm.nan, strategy='median')
imputer = imputer.fit(X[:, 1:14]) # fit is a method
X[:, 1:14] = imputer.transform(X[:, 1:14])

from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), [0])], remainder='passthrough')
X = nm.array(ct.fit_transform(X))
```

Imputing is used fill the null spots in the dataset with respect to the data. In this model null spots are filled with median of the data values.
The strings in the dataset are to be encoded so that we can train the model. Onehot encoder is used to encode the data.

4. Splitting data

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X ,Y, test_size=0.2, random_state = 1)
```

Dataset is divided into training and testing sets

5. Training

```
from sklearn import linear_model
reg = linear_model.LinearRegression()
reg.fit(X_train, Y_train)

area_map = {1:'TELANGANA'}
X[:, 0] = map(X[: ,0],area_map)

input_list = nm.array([[1,year,9.58,11.68,15.6,20.18,27.37,145.12,249.59,218.05,177.50,77.22,21.85,9.14,temperature]])
y1_pred = reg.predict(input_list[:1])
```

Training set is used to test the data. The model used is Multiple linear regression for rainfall prediction.

6. Temperature prediction using Linear regression

```
def temp_predict(year):
    year = int(year)
    dataset = pd.read_csv('tempdata.csv')
    X = dataset.iloc[:, :-1].values
    Y = dataset.iloc[:, -1].values

    from sklearn.model_selection import train_test_split
    X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, random_state = 0)


    from sklearn.linear_model import LinearRegression
    regressor = LinearRegression()
    regressor.fit(X_train, Y_train)
    y_pred = regressor.predict([[year]])
```

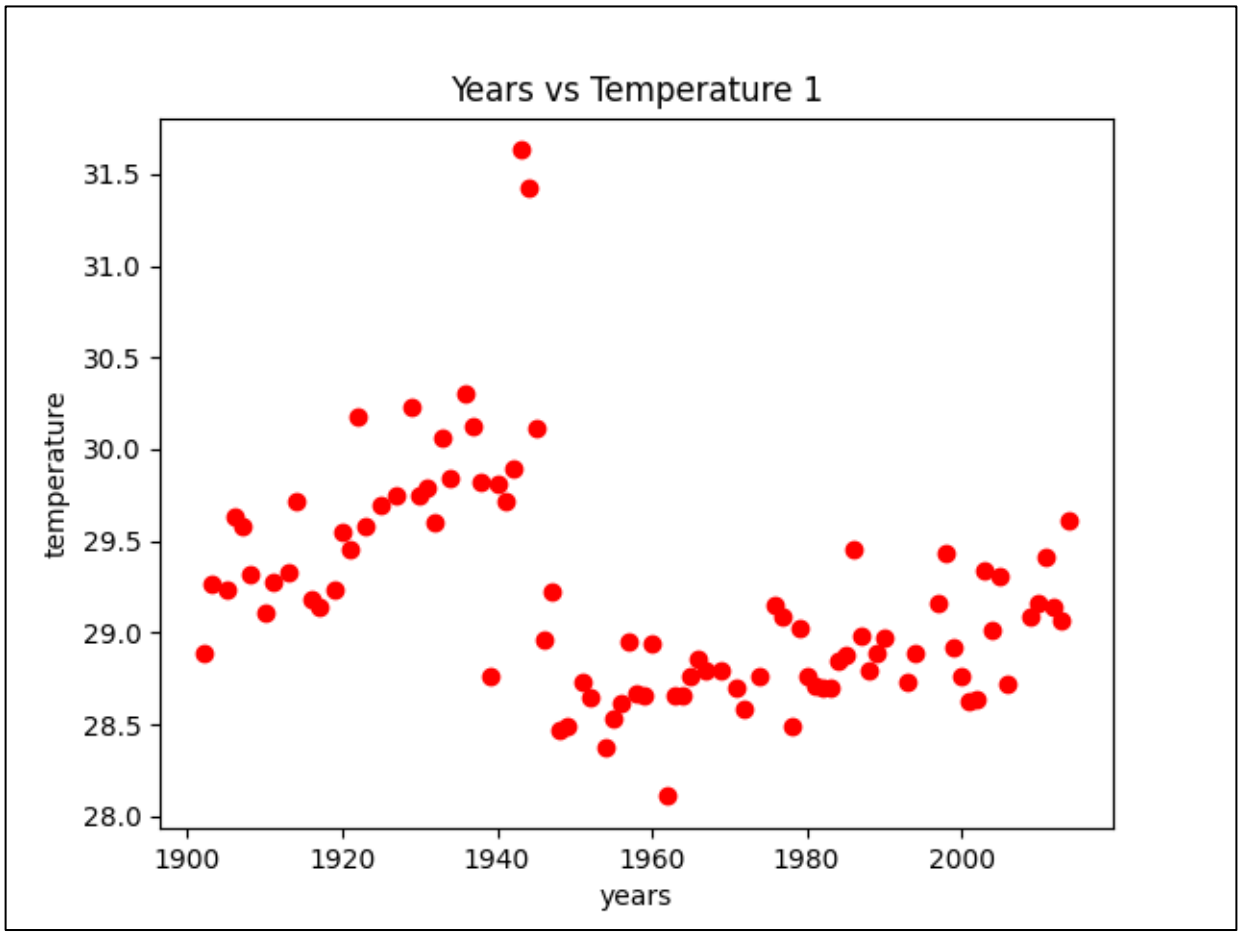
Temperature for the given year is predicted using temperature dataset which has year and corresponding average temperature of the year.

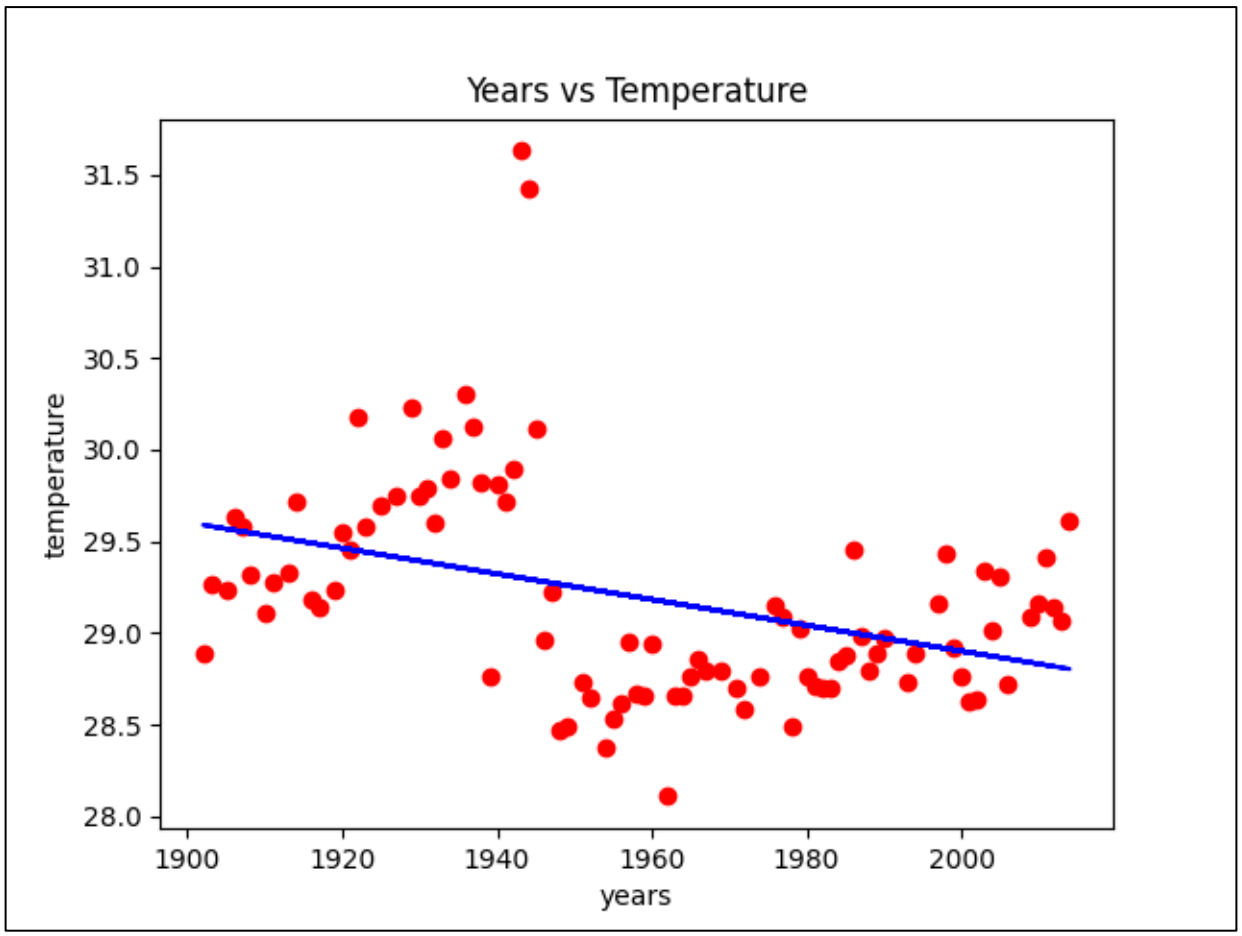
7. Plotting temperature data

```
plot.scatter(X_train, Y_train, color = 'red')
plot.title('Years vs Temperature 1')# title for our graph
plot.xlabel('years')
plot.ylabel('temperature')
plot.show()# shows graphic

plot.scatter(X_train, Y_train, color = 'red')
plot.plot(X_train, regressor.predict(X_train), color = 'blue')
plot.title('Years vs Temperature')# title for our graph
plot.xlabel('years')
plot.ylabel('temperature')
plot.show()# shows graphic
```

Using Matplotlib library scatter pots are graphed to know the data distribution.


Years vs Temperature 1
temperature
years
31.5
31.0
30.5
30.0
29.5
29.0
28.5
28.0
1900
1920
1940
1960
1980
2000


Years vs Temperature
temperature
years
31.5
31.0
30.5
30.0
29.5
29.0
28.5
28.0
1900
1920
1940
1960
1980
2000

Future developments

With already existing models that predict average rainfall in a period of time or predicts probability of rain on a particular day the models can also be trained to predict period of time over which there can be heavy or less rainfall, areas that maybe affected by heavy or no rainfall in future, if people of a particular area faces water problem due to less rainfall than required etc. But such models require more complex and accurate data and also machine learning models should be able to process large datasets, to process large datasets more computing power is also required. Rainfall prediction and weather prediction models can be used to warn farmers or in navigation systems of aerospace and naval systems which can help them in preventing damage.
Institutes like Indian Meteorological Department use data of previous year with rainfall, temperature, humidity, wind speed, elevation, and also data collected by weather monitoring satellites which can show the formation and accumulation of clouds over an area to predict rainfall in the area accurately. Such information is broadcasted prior to any weather alerts [8].
The future weather predicting systems may predict weather more accurately if they can access real time data of weather monitoring satellites.

References

1. https://docs.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model
2. https://labelyourdata.com/articles/what-is-dataset-in-machine-learning
3. https://www.javatpoint.com/data-preprocessing-machine-learning
4. Dataset from Kaggle
   https://www.kaggle.com/datasets/rajanand/rainfall-in-india
5. scribbr.com/statistics/multiple-linear-regression/
6. https://www.investopedia.com
7. https://medium.com/mlearning-ai/multiple-linear-regression-fundamentals-and-modeling-in-python-60db7095deff
8. https://mausam.imd.gov.in