Rain Fall prediction

A Project Report in partial fulfillment of the degree

Bachelor of Technology

in

Computer Science &Engineering

By

2103A51453	CH.SOUMIKA
2103A51363	K.SRILEKHA
2103A51575	V.ASHWINI

Under the Guidance of
D. Ramesh

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
S R UNIVERSITY, ANANTHASAGAR, WARANGAL

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled “MACHINE FAILURE PREDICTION” is a record of Bonafide work carried out by **ch.soumika, k.srilekha, v.ashwini** bearing Roll No(s) **2103A51453, 2103A51363, 2103A51575** during the academic year 2022-2023 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science Engineering** by the SR UNIVERSITY, WARANGAL.

Supervisor

Mr. D. Ramesh
SR University

Head of the Department

Dr. M. Sheshikala
SR University

External Examiner

ACKNOWLEDGEMENT

We express our thanks to course coordinator Mr. D. Ramesh, Asst. prof. for guiding us from the beginning through the end of the course project. We express our gratitude to head of the department CS&AI, Dr. M. Sheshikala, Associate Professor for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, Dr C. V. Guru Rao, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thank to all teaching and non-teaching staff of the department for their suggestions and timely support.

2103A51453-ch.soumika

2103A51363-k.srilekha

2103A51575-v.ashwini

ABSTRACT

Rainfall prediction is an important area of research , as it has significant implication for agriculture, water resources management, and disaster preparedness. The prediction of rainfall involves the use of various techniques and models ,ranging from statical methods to machine learning algorithms.

These models take into account a variety of variable such as temperature, humidity wind speed and atmospheric pressure, which are all known to affect rainfall patterns.

In recent years, there has been significant progress in the development of rainfall prediction models, thank to advancement in computing power and data collection. These models have been used to provide accurate prediction of rainfall patterns in different regions, helping farmers and policy makers make informed decision about crop planning And water resource management.

However, despite these advancement ,prediction rainfall patterns remain a complex and challenging task, a weather patterns can be highly unpredictable and influenced by a range of factor. Therefore , continued research in this field is needed to further improve the accuracy and reliability of rainfall prediction models, and to develop new approaches that can better account for the complexity of weather pattern.

Table of Contents

1.INTRODUCTION:

Our project is about rainfall prediction it predict the rainfall in each state for all the months form Jan to Dec. As global warming is increasing earths temperature due to which our local regions yearly rainfall patterns have been affecting This harms the farmers and other people who depend on rainfall Proper water supply keeps farm land in a good condition For this rainfall prediction there are many researches conducted using data mining and machine learning proper rainfall should be there to in correct way to prevent flooding, drought, landslides, mass movements and avalanches Timely and accurate forecasting can help reduce human and financial loss The main theme of this project is to study and identify atmosphere that cause rainfall and to the intensityIt describes the relationship between atmospheric variables that affect the rainfall Rainfall is a climate factor that affects many human activities like agricultural production construction,power generation,foresty and tourism A study is conducted and identified solar radiation,perceptible water vapor are important variables for daily rainfall prediction This is using data driven machine learning algorithm but it is better to use simple linear regression which has only one independent feature.

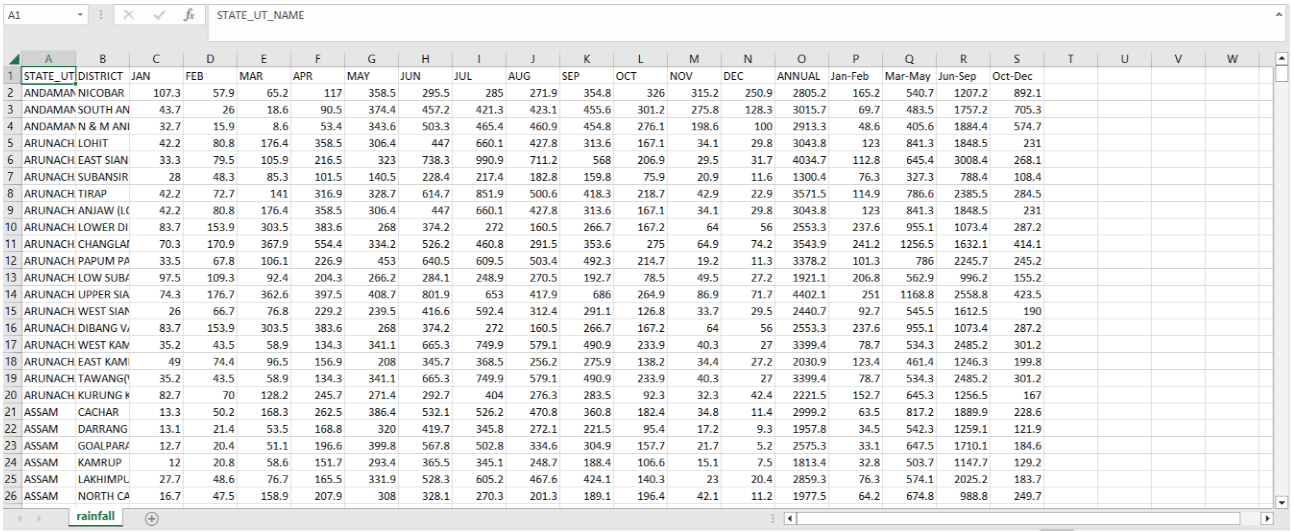
2.LITERATURE REVIEW

In previous research papers, we have observed that different machine learning algorithms have been used. Few papers are based on deep learning also. The field of Artificial Intelligence has been the suitable area to carry out all types of predictions on the dataset by extracting and data preprocessing. Logistic Regression, Support Vector Machine, Naïve Bayes Classification, Linear regression and ridge regression etc. are the various machine learning algorithms the have been used. We have observed that the algorithms work together by generating the pattern among the available dataset and proceeding with prediction. Mid Infrared Spectroscopy combined with few machine learning algorithms. Deep learning is something that works by generating biases and weights in the layers, rule based takes the bulk values and signifies a rule in it. SVM are used with algorithms especially which follows a close correlation among the variables taken into consideration. Artificial Neural Network inspired by the structure and function of the human brain. PLS regression stands for Partial Square regression, which is a statistical technique used for modelling the relationship between the two sets of variables. In PLS regression, both the predictor variables and the response variables are transformed into new sets of variables called latent variables, which are linear combination of the original variables. PLS regression is useful for predicting a response variable from a large number of predictor variables, even when these variables are highly correlated. It is commonly used in fields such as chemistry, biology, and engineering, where there are many variables to consider in modelling complex systems. It is also used in data analysis and machine learning to identify important variables and reduce dimensionality of the data

Authors	**Region**	**Techniques**	**Rainfall predicting attribute**	**Accuracy measure**
M.Kannan et al.	*Global*	*Regression*	*Rainfall,humidity*	*MSE*
S. Chattopadhyay	*Global*	*ANN*	*Rainfall*	*MSE*
P. Dutta, H. Tahbilder	*Global*	*Regression*	*Rainfall*	*MSE*
P. Goswami, Srividya	*Global*	*ANN*	*Mean rainfall*	*Relative percentage error*
S. Kannan, S. Ghosh	*Local (river)*	*Decision tree, CART, K-mean*	*Rainfall,humidity*	*MSE*
A. Naik	*Global*	*Monthly*	*Wind,speed,temperature,humidity*	*RMSE*
S.nanda	*Global*	*Yearly*	*Min_max temperature*	*MSE*
R.Deshpande	*Local*	*Monthly*	*Rainfall*	*MSE*
G.shrivastava	*Local*	*Yearly*	*Humidity,dew point,pressure*	*MSE*
P.dutta,H.Tahbilder	*Global*	*Monthly*	*Min-Max,temperature,wind direction,humidity,rainfall*	*RMSE*

3.DATA SET DESCRIPTION

- This dataset contains 641number of rows and 19 columns each row has Indian state name ,district ,rainfall amount in cm/inches of each and every month from January to December
- As this dataset considered as supervised because it is labelled dataset to train algorithms that to classify data or predict outcomes accurately
- JAN
- FEB
- MAR
- APR
- MAY
- JUNE
- JUL
- AUG
- SEP
- OCT
- NOV
- DEC
- JAN-FEB
- MAR-MAY
- JUN-SEP

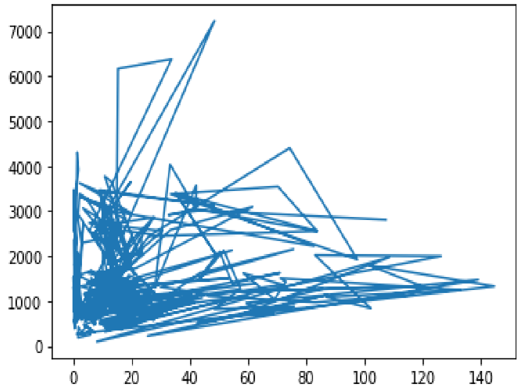


STATE_UT_NAME

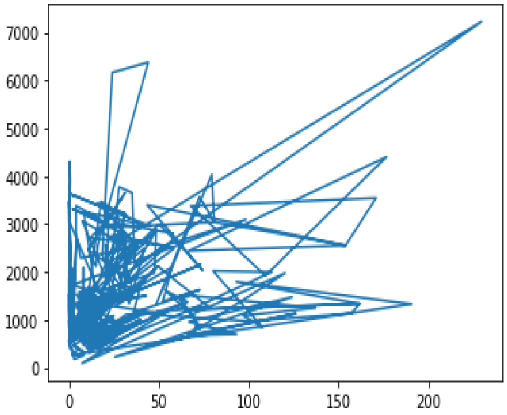
STATE_UT	DISTRICT	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
ANDAMAN	NICOBAR	107.3	57.9	65.2	117	358.5	295.5	285	271.9	354.8	326	315.2	250.9	2805.2	165.2	540.7	1207.2	892.1
ANDAMAN	SOUTH AN	43.7	26	18.6	90.5	374.4	457.2	421.3	423.1	455.6	301.2	275.8	128.3	3015.7	69.7	483.5	1757.2	705.3
ANDAMAN	N & M ANI	32.7	15.9	8.6	53.4	343.6	503.3	465.4	460.9	454.8	276.1	198.6	100	2913.3	48.6	405.6	1884.4	574.7
ARUNACH	LOHIT	42.2	80.8	176.4	358.5	306.4	447	660.1	427.8	313.6	167.1	34.1	29.8	3043.8	123	841.3	1848.5	231
ARUNACH	EAST SIAN	33.3	79.5	105.9	216.5	323	738.3	990.9	711.2	568	206.9	29.5	31.7	4034.7	112.8	645.4	3008.4	268.1
ARUNACH	SUBANSIR	28	48.3	85.3	101.5	140.5	228.4	217.4	182.8	159.8	75.9	20.9	11.6	1300.4	76.3	327.3	788.4	108.4
ARUNACH	TIRAP	42.2	72.7	141	316.9	328.7	614.7	851.9	500.6	418.3	218.7	42.9	22.9	3571.5	114.9	786.6	2385.5	284.5
ARUNACH	ANJAW (LC	42.2	80.8	176.4	358.5	306.4	447	660.1	427.8	313.6	167.1	34.1	29.8	3043.8	123	841.3	1848.5	231
ARUNACH	LOWER DI	83.7	153.9	303.5	383.6	268	374.2	272	160.5	266.7	167.2	64	56	2553.3	237.6	955.1	1073.4	287.2
ARUNACH	CHANGLAI	70.3	170.9	367.9	554.4	334.2	526.2	460.8	291.5	353.6	275	64.9	74.2	3543.9	241.2	1256.5	1632.1	414.1
ARUNACH	PAPUM PA	33.5	67.8	106.1	226.9	453	640.5	609.5	503.4	492.3	214.7	19.2	11.3	3378.2	101.3	786	2245.7	245.2
ARUNACH	LOW SUBA	97.5	109.3	92.4	204.3	266.2	284.1	248.9	270.5	192.7	78.5	49.5	27.2	1921.1	206.8	562.9	996.2	155.2
ARUNACH	UPPER SIA	74.3	176.7	362.6	397.5	408.7	801.9	653	417.9	686	264.9	86.9	71.7	4402.1	251	1168.8	2558.8	423.5
ARUNACH	WEST SIAN	26	66.7	76.8	229.2	239.5	416.6	592.4	312.4	291.1	126.8	33.7	29.5	2440.7	92.7	545.5	1612.5	190
ARUNACH	DIBANG VA	83.7	153.9	303.5	383.6	268	374.2	272	160.5	266.7	167.2	64	56	2553.3	237.6	955.1	1073.4	287.2
ARUNACH	WEST KAM	35.2	43.5	58.9	134.3	341.1	665.3	749.9	579.1	490.9	233.9	40.3	27	3399.4	78.7	534.3	2485.2	301.2
ARUNACH	EAST KAMI	49	74.4	96.5	156.9	208	345.7	368.5	256.2	275.9	138.2	34.4	27.2	2030.9	123.4	461.4	1246.3	199.8
ARUNACH	TAWANG(	35.2	43.5	58.9	134.3	341.1	665.3	749.9	579.1	490.9	233.9	40.3	27	3399.4	78.7	534.3	2485.2	301.2
ARUNACH	KURUNG K	82.7	70	128.2	245.7	271.4	292.7	404	276.3	283.5	92.3	32.3	42.4	2221.5	152.7	645.3	1256.5	167
ASSAM	CACHAR	13.3	50.2	168.3	262.5	386.4	532.1	526.2	470.8	360.8	182.4	34.8	11.4	2999.2	63.5	817.2	1889.9	228.6
ASSAM	DARRANG	13.1	21.4	53.5	168.8	320	419.7	345.8	272.1	221.5	95.4	17.2	9.3	1957.8	34.5	542.3	1259.1	121.9
ASSAM	GOALPARA	12.7	20.4	51.1	196.6	399.8	567.8	502.8	334.6	304.9	157.7	21.7	5.2	2575.3	33.1	647.5	1710.1	184.6
ASSAM	KAMRUP	12	20.8	58.6	151.7	293.4	365.5	345.1	248.7	188.4	106.6	15.1	7.5	1813.4	32.8	503.7	1147.7	129.2
ASSAM	LAKHIMPU	27.7	48.6	76.7	165.5	331.9	528.3	605.2	467.6	424.1	140.3	23	20.4	2859.3	76.3	574.1	2025.2	183.7
ASSAM	NORTH CA	16.7	47.5	158.9	207.9	308	328.1	270.3	201.3	189.1	196.4	42.1	11.2	1977.5	64.2	674.8	988.8	249.7

4.DATA VISULAZATION:

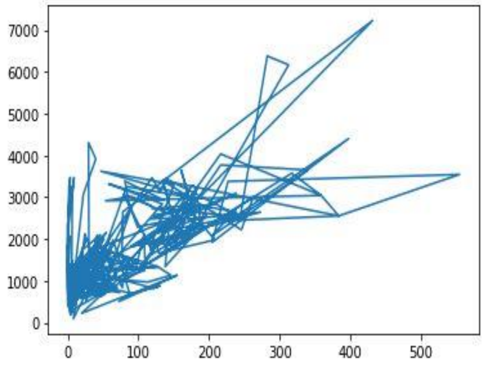
The following are plotting of each feature against the target.



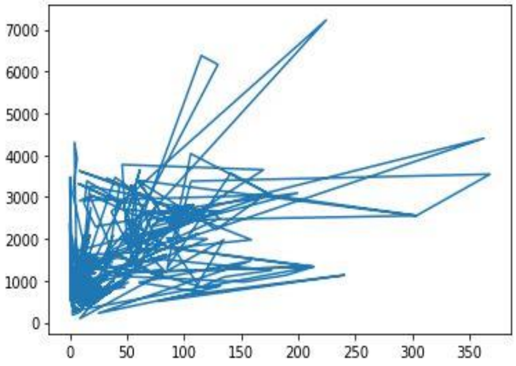
Jan vs annual



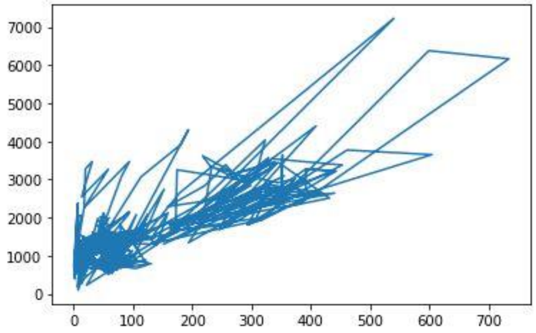
feb



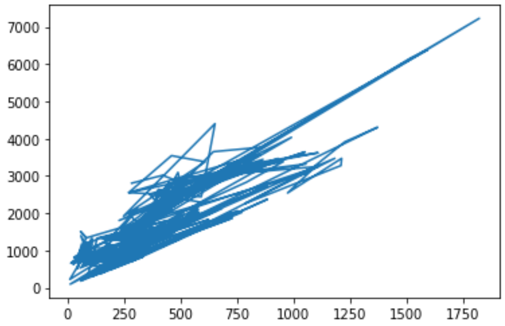
mar



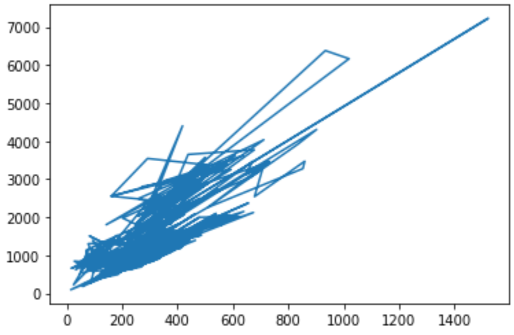
apr



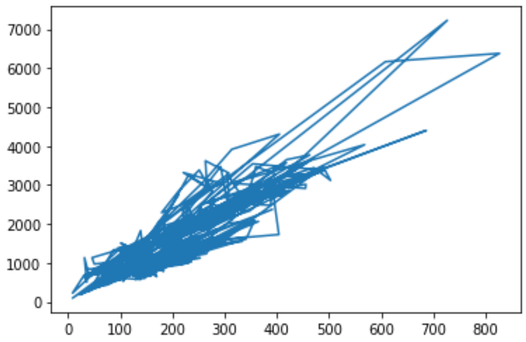
may



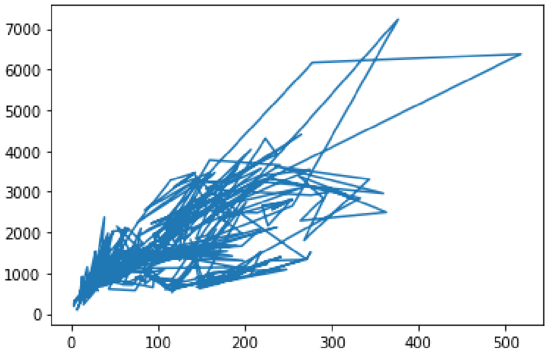
june



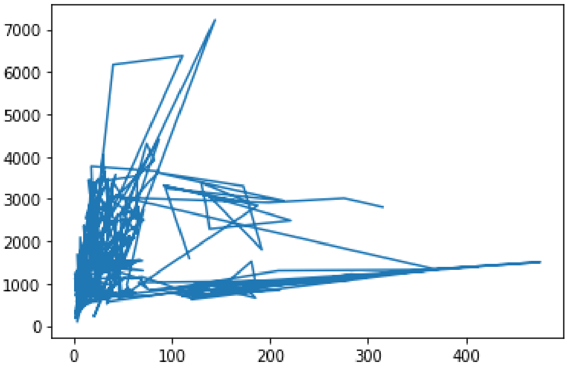
july



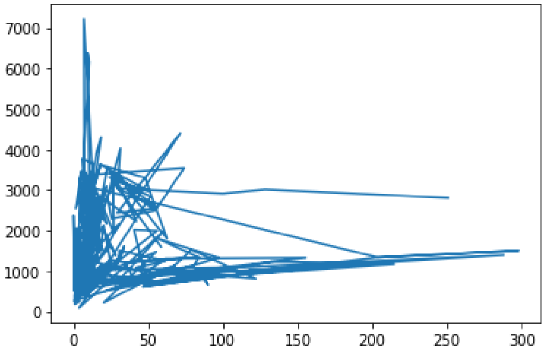
august



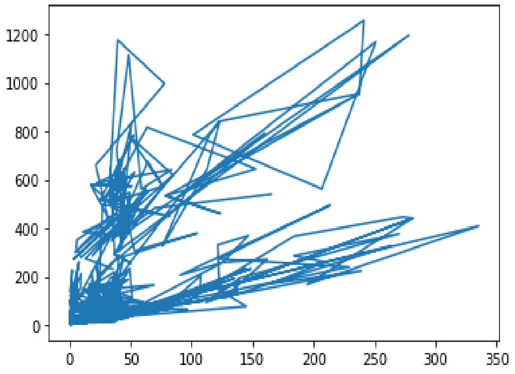
September



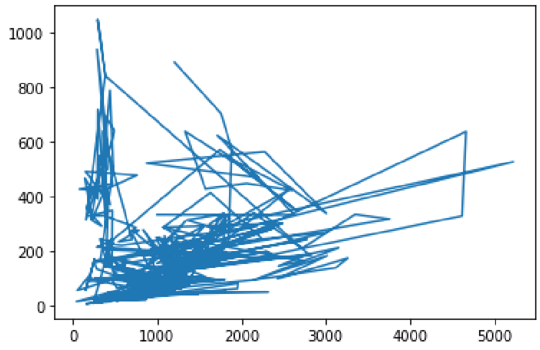
october



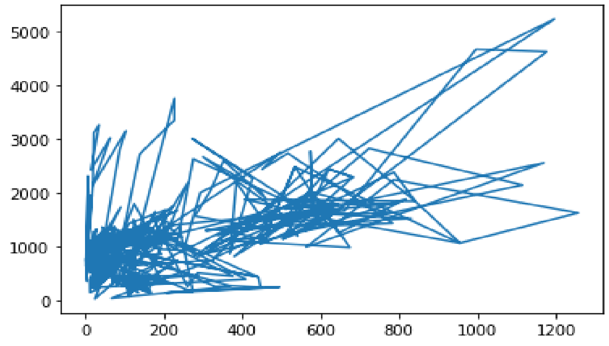
November



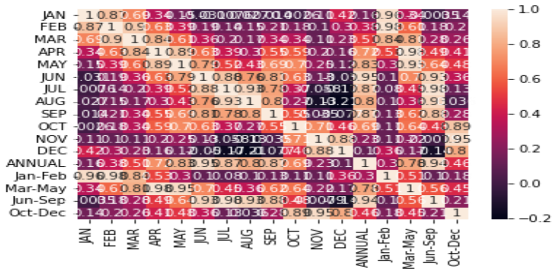
December



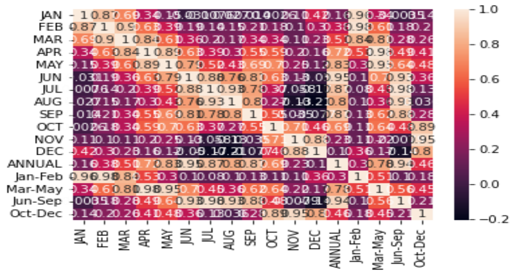
Jan-feb vs mar-may



mar-may vs jun-sep



Co variance



co relation

5. METHODOLOGY

PROCEDURE TO SOLVE THE GIVEN PROBLEM

In this project Dogecoin price prediction and prediction, we use three approaches:

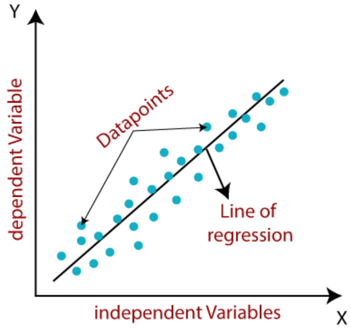
- Linear regression
- K-Nearest Neighbour
- Support Vector Machine
- Decision Tree

Linear regression:

Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimiz the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

Advantages of linear regression algorithm:

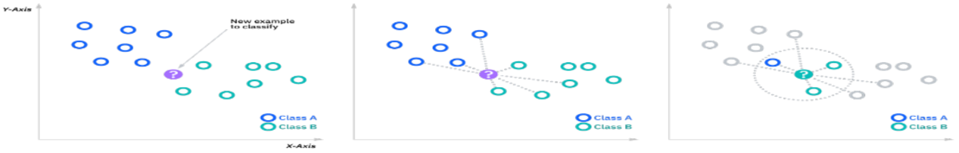
- Linear regression performs exceptionally well for linearly separable data
- Easier to implement, interpret and efficient to train
- It handles overfitting pretty well using dimensionally reduction techniques, regularization, and cross-validation
- One more advantage is the extrapolation beyond a specific data set



K-Nearest Neighbour:

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.



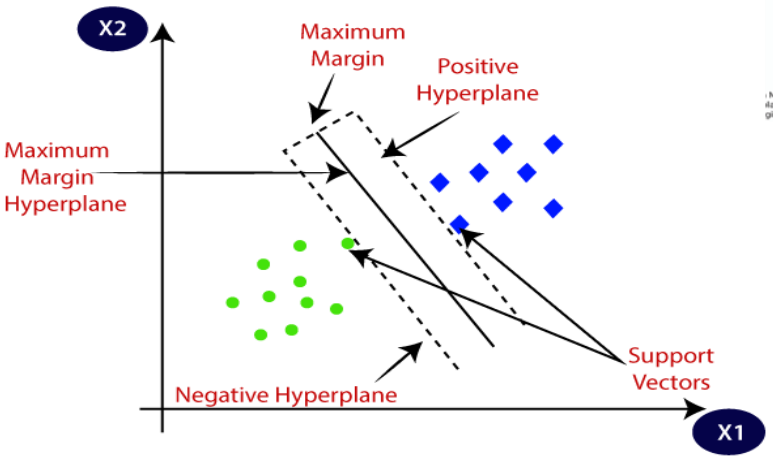
KNN Formula:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

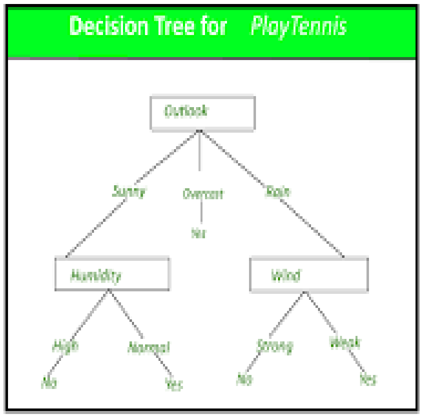


Decision tree

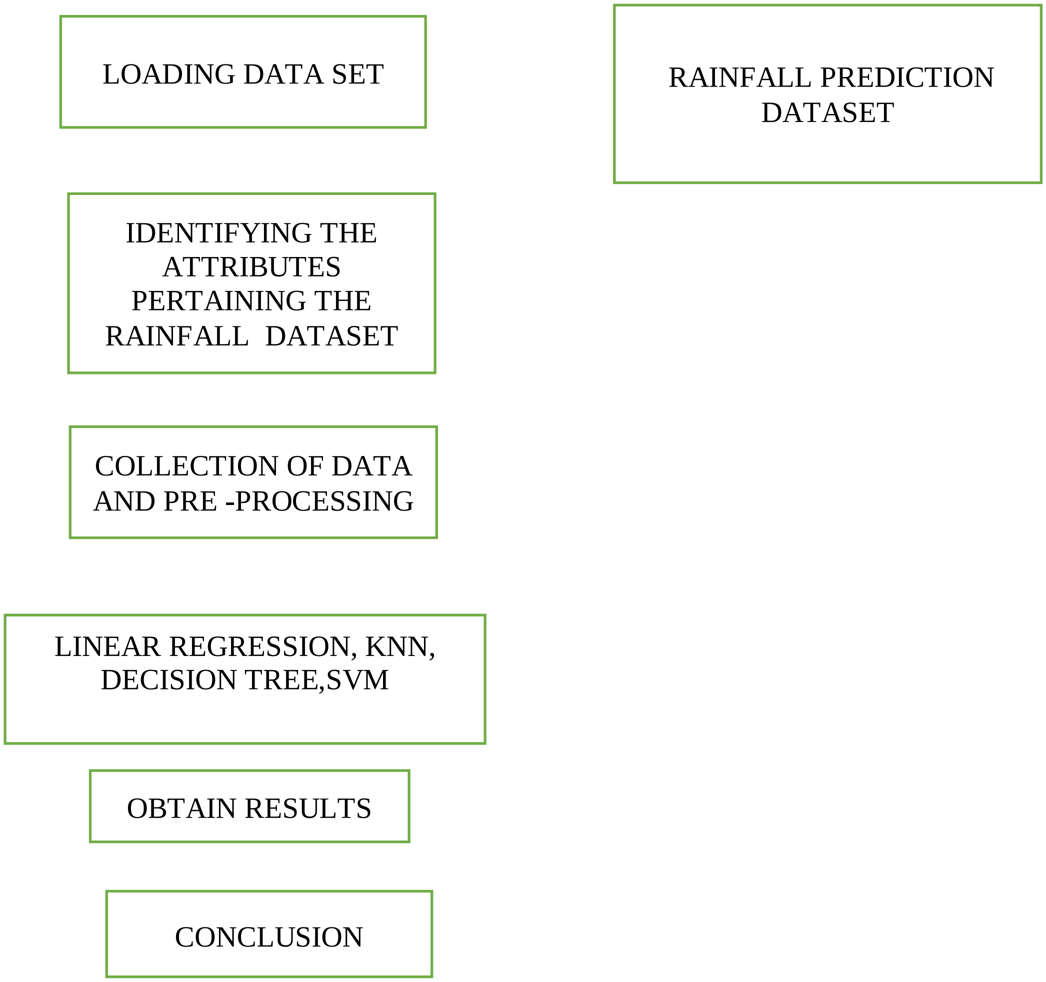
Decision trees are a nonparametric supervised learning method used for classification and regression. The deeper the tree, the more complex the decision rules and the fitter the model. Decision tree uses the tree representation to solve the problem. In which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. The primary challenge in the decision tree implementation is to identify the attributes. There are two popular attribute selection measures they are Entropy and Gini index. Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information content

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$IG(T,A) = Entropy(T) - \sum_{v \epsilon A} \frac{|T_v|}{T} \cdot Entropy(T_v)$$



6. MODEL ARCHITECTURE


LOADING DATA SET
RAINFALL PREDICTION DATASET
IDENTIFYING THE ATTRIBUTES PERTAINING THE RAINFALL DATASET
COLLECTION OF DATA AND PRE -PROCESSING
LINEAR REGRESSION, KNN, DECISION TREE,SVM
OBTAIN RESULTS
CONCLUSION

-

SOFTWARE DESCRIPTION

Software requirements:

Operating system: Windows

Platform: google Collab

Programing language: python

7. RESULTS

CODE

Dataset:

```
import pandas as pd
d=pd.read_csv('/content/rainfall prediction.csv')
print(d)
```

output:

```
             STATE_UT_NAME      DISTRICT   JAN   FEB    MAR   APR  \
0    ANDAMAN And NICOBAR ISLANDS       NICOBAR  107.3  57.9   65.2  117.0
1    ANDAMAN And NICOBAR ISLANDS   SOUTH ANDAMAN   43.7  26.0   18.6   90.5
2    ANDAMAN And NICOBAR ISLANDS   N & M ANDAMAN   32.7  15.9    8.6   53.4
3            ARUNACHAL PRADESH         LOHIT   42.2  80.8  176.4  358.5
4            ARUNACHAL PRADESH     EAST SIANG   33.3  79.5  105.9  216.5
..                       ...           ...   ...   ...    ...    ...
636                   KERALA        IDUKKI   13.4  22.1   43.6  150.4
637                   KERALA      KASARGOD    2.3   1.0    8.4   46.9
638                   KERALA  PATHANAMTHITTA   19.8  45.2   73.9  184.9
639                   KERALA       WAYANAD    4.8   8.3   17.5   83.3
640              LAKSHADWEEP    LAKSHADWEEP   20.8  14.7   11.8   48.9

      MAY    JUN     JUL    AUG    SEP    OCT    NOV    DEC  ANNUAL  Jan-Feb  \
0    358.5  295.5   285.0  271.9  354.8  326.0  315.2  250.9  2805.2    165.2
1    374.4  457.2   421.3  423.1  455.6  301.2  275.8  128.3  3015.7     69.7
2    343.6  503.3   465.4  460.9  454.8  276.1  198.6  100.0  2913.3     48.6
3    306.4  447.0   660.1  427.8  313.6  167.1   34.1   29.8  3043.8    123.0
4    323.0  738.3   990.9  711.2  568.0  206.9   29.5   31.7  4034.7    112.8
..     ...    ...     ...    ...    ...    ...    ...    ...     ...      ...
636  232.6  651.6   788.9  527.3  308.4  343.2  172.9   48.1  3302.5     35.5
637  217.6  999.6  1108.5  636.3  263.1  234.9   84.6   18.4  3621.6      3.3
638  294.7  556.9   539.9  352.7  266.2  359.4  213.5   51.3  2958.4     65.0
639  174.6  698.1  1110.4  592.9  230.7  213.1   93.6   25.8  3253.1     13.1
640  171.7  330.2   287.7  217.5  163.1  157.1  117.7   58.8  1600.0     35.5

     Mar-May  Jun-Sep  Oct-Dec
0      540.7   1207.2    892.1
1      483.5   1757.2    705.3
2      405.6   1884.4    574.7
3      841.3   1848.5    231.0
4      645.4   3008.4    268.1
..       ...      ...      ...
636    426.6   2276.2    564.2
637    272.9   3007.5    337.9
638    553.5   1715.7    624.2
639    275.4   2632.1    332.5
```

```
640   232.4   998.5   333.6

[641 rows x 19 columns]
```

Linear regression:

```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
mm=lr.fit(x_train,y_train)
yp=mm.predict(x_test)
print(yp)
```

output:

```
[1233.9 1223.4 1327.9 1057.6 2641.8  646.5  961.1 1070.6  485.7 1122.9
 1029.6 3470.6 1209.3  308.1 2958.4  498.  2814.4 1796.5 1068.5  646.1
 2440.7 1973.9 1081.4 2859.3 1293.1 3468.3  898.2  992.9 1235.7 1535.5
 3094.5  966.7  793.4  449.2  747.1  544.  1803.2  818.   508.1 3218.7
  746.9 2480.6  839.2 1336.5  460.6 1533.5 6379.9 1003.3  837.  1087.7
 2127.5  622.8 1123.6  685.6 1366.2 1680.7 1481.6  788.4  777.  2512.6
  992.2  747.1 1336.5  388.8  863.6 2805.2 1416.2  708.4 1293.3  902.6
  974.9  747.1 1474.3  613.9  449.4  700.4 2731.1 1921.1  807.8 2123.9
 1528.2  655.  1091.6 1618.3 3302.5  572.  1146.8 1385.5 1148.6 1109.9
 2374.1  886.1 2116.9  818.7  897.4 2098.  1005.6  419.5  714.4 1363.3
 1448.3  936.2 1155.4 1062.7  871.5  720.  1008.4  455.6 1192.2 1191.5
 2814.4  986.3  963.9  252.9  850.1 1229.  1104.7  301.6 1474.1 3399.4
 1010.8 1504.4 1530.9 1392.7 1584.9 1462.1  692.7 2556.6 1206.7]
```

```
from sklearn.metrics import mean_squared_error
print(mean_squared_error(yp,y_test))
```

output:

```
2.862281872213113e-25
```

```
from sklearn.metrics import mean_absolute_error
print(mean_absolute_error(yp,y_test))
```

output:

```
3.615505827542091e-13
```

```
mse = mean_squared_error(y_test, yp)
print("Mean Squared Error:", mse)
```

output:

```
Mean Squared Error: 2.862281872213113e-25
```

```
mae = mean_absolute_error(y_test, yp)
print("Mean Absolute Error:", mae)
```

output:

```
Mean Absolute Error: 3.615505827542091e-13
```

K-Nearest Neighbour:

```
from sklearn.neighbors import KneighborsRegressor
knn = KNeighborsRegressor(n_neighbors=3)
```

```
knn.fit(x_train, y_train)
y_pred = knn.predict(x_test)

mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error:", mae)
```

output:

```
Mean Absolute Error: 50.94470284237724
```

```
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
```

output:

```
Mean Squared Error: 7804.861584840647
```

```
from sklearn.metrics import mean_squared_error
print(mean_squared_error(yp,y_test))
```

output:

```
70357.10162790696
```

```
from sklearn.metrics import mean_absolute_error
print(mean_absolute_error(yp,y_test))
```

output:

```
101.26434108527131
```

Support Vector Machine:

```
from sklearn.svm import SVR
model = SVR(kernel='linear')
model.fit(x_train,y_train)
y_pred=model.predict(x_test)

mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error:", mae)
```

output:

```
Mean Absolute Error: 0.047535886200007877
```

```
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
```

output:

```
Mean Squared Error: 0.003450724484773055
```

```
from sklearn.metrics import mean_squared_error
print(mean_squared_error(yp,y_test))
```

output:

```
70357.10162790696
```

```
from sklearn.metrics import mean_absolute_error
print(mean_absolute_error(yp,y_test))
```

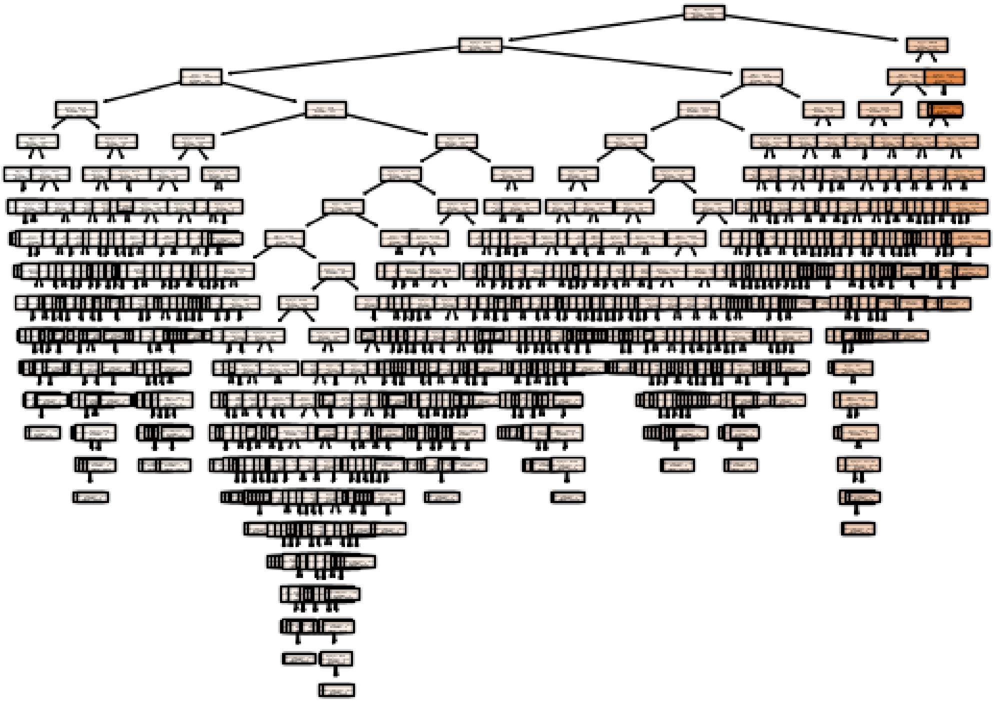
output:

```
101.26434108527131
```

Decision Tree:

```
from sklearn.tree import DecisionTreeRegressor
model=DecisionTreeRegressor()
model.fit(x_train,y_train)
yp=model.predict(x_test)
print(yp)
from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_test,yp))
from sklearn import tree
tree.plot_tree(model,filled=True)
```

output:



sno	Machine Learning Model	Mean square error
1	Linear regression	2.862281872213113e-25
2	k-nearest neighbour	7804.861584840647
3	Decision tree	24676.522558139535
4	Support vector machine	0.003450724484773055

8. CONCLUSION AND FUTURE SCOPE

- There are some specific problems in the world that pushes the capability of data science and the technology available in this field to their edge among them one is rainfall predicition

- We can easily conclude that for rainfall prediction this is the best way to use it by forming a range of highest and lowest predicted values by adding bias in the model

- Rainfall prediction main objective is prediction of amount of rain in a specific well or division by using various techinques and finding out which one is best

- Future scope of rainfall prediction

The future scope of rainfall prediction is very promising, with advancements in technology and data analysis techniques. Some of the potential developments in this field include:

- Improvements in Data Collection

- Integration of Big Data

- Advances in Cloud Computing

- Development of Early Warning Systems

- In summary, the future of rainfall prediction looks bright, and with continued research and innovation, we can expect more accurate and reliable predictions that can help people and communities prepare for extreme weather events.

9.REFERENCES

http://repository.wit.ie/3326/1/InfomationScience_postprint.pdf

https://www.sciencedirect.com/science/article/pii/S0022030215004932

https://www.kaggle.com/code/darsh79/starter-rainfall-in-india-99bfc809-4

https://www.tandfonline.com/doi/abs/10.4081/ijas.2009.s2.399

https://orbi.uliege.be/handle/2268/224000

https://www.sciencedirect.com/science/article/pii/S0022030221005099