WEATHER FORECASTING

Btech.AG - SRIMAYA MOHAPATRA- 203001170028

MAHESH JENA - 203001170027

RUKMINIBALLABHA BEHERA – 203001170047

Abstract:

As we know that in ancient time, people usually don't get about weather news and condition for which people suffer many losses due to sudden change in weather conditions. In modern technology, we usually get all reports related to weather forecasting, for ex- we can know about weather information on the temperature, climatic conditions, humidity, pressure etc. in a particular place. Also we can get the weather conditions of any locations simply by using device. Basically we can get this information's with the help of machine learning where it takes inputs of the first experience and predict the output.

Introduction

So in our project we want to show about which weather condition, we can predict rainfall, i.e – at how much pressure, temperature or humidity we can predict rainfall through good accuracy

For our Project, we took 7 columns and 100 Rows. These are our columns

Temperature(C) Humidity Visibility(km) Precip WindSpeed(km/h) LoadCover Apparent Temperature (C)

For splitting arrays or matrices into random train and test subsets we use Sklearn model.

We Use Simple Linear, Multilinear, Polynomial Linear, Decision tree algorithms in our project.

We use Seaborne for plotting the graph and pair plot to find the relationship between different types of columns. Seaborne give the correlation between the value to identify the columns needed for prediction.

We have generated heat map to find the correlation to determine the accuracy easily. For accuracy in Simple Linear, Multilinear, Polynomial Linear we took x – axis as "Temperature(C)" and y axis as "Apparent Temperature(C)"

We have converted string values into integer value, since machine understand only numerical value.

For testing and training we use Random State = 0, for which 75% of testing Data will be fixed and 25% of training will be fixed, so the data will not change between the testing and training.

In Simple linear and Multilinear we use r2_score for finding the accuracy

We use StandardScaler to standardize the value, set the range of Dataset, we do graph to check whether training or testing are same or not (to know whether training is having more accuracy or testing having more accuracy), in graph intercept means where the slope is started and coefficients means to find out the parameters of the value.

In Multilinear Regression when we convert multiple number of string value into numerical value we use label encoder and one Hot Encoder, we use label encoder to convert string values to the

numerical values and we use One Hot Encoder for dividing these values into column.
In Polynomial regression we use scatterplot and lines to plot the Linear Regression results and polynomial regression results.
Simple Linear, Multilinear, Polynomial Linear are supervised learning Algorithms

We used impurity (to measure and compare) to calculate each feature with output, we know impurity by using sini as criterion for yes and no condition. Here we calculate root node by using sini impurity. Decision tree follow CART algorithm (Classification and Regression)

We do iris and knn, SVM, Decision Tree Classification, Random Forest in Classification, for training and testing we use append function. So Basically we used Supervised Algorithm for our Project.

We do K Means Clustering which is unsupervised learning Algorithm. We used 5 cluster in K Means Clustering.

Literature

We are getting our Dataset from https://www.kaggle.com/salmamaamouri/weather-prediction-regression-neural-model. Our Dataset have 7 Columns and 100 Rows. Our Dataset is based on weather forecasting where we get the information about temperature, humidity, pressure. Our project is about historical weather around Szeged, Hungary, from 2006 to 2016.So the main objectives of our project is on analysis of various factors affecting Weather, using of different prediction and classification of algorithms. We have taken first 100 records from the Dataset. We have used Temperature(C), Humidity, Visibility(km), Precip, WindSpeed(km/h), LoudCover, Apparent Temperature (C) in columns. We used Numpy, Pandas, Matplotlib, sklearn in our library . We used Supervised and Unsupervised Learning for our Dataset.

	Temperature (C)	Humidity	Visibility (km)	Precip	Wind Speed (km/h)	Loud Cover	Apparent Temperature (C)
0	9.472222	0.89	15.8263	rain	14.1197	0	7.388889
1	9.355556	0.86	15.8263	rain	14.2646	0	7.227778
2	9.377778	0.89	14.9569	rain	3.9284	0	9.377778
3	8.288889	0.83	15.8263	rain	14.1036	0	5.944444
4	8.755556	0.83	15.8263	rain	11.0446	0	6.977778
...	...	...	...	...	...	...	...
94	7.827778	0.72	15.8263	rain	13.8943	0	5.405556
95	7.855556	0.72	15.0052	rain	9.8049	0	6.122222
96	7.316667	0.75	15.8746	rain	6.6654	0	6.211111
97	7.244444	0.75	15.8746	rain	7.1162	0	6.005556
98	5.438889	0.88	9.9820	rain	3.7191	0	5.438889

99 rows × 7 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Temperature (C)          99 non-null     float64
 1   Humidity                 99 non-null     float64
 2   Visibility (km)          99 non-null     float64
 3   Precip                   99 non-null     object
 4   Wind Speed (km/h)        99 non-null     float64
 5   Loud Cover               99 non-null     int64
 6   Apparent Temperature (C) 99 non-null     float64
dtypes: float64(5), int64(1), object(1)
memory usage: 5.5+ KB
```

Existing Work –

Simple Linear Regression

Polynomial Regression

Multi-Linear Regression

Logistic Regression

KNN (KNearest Neighbour)

SVM (Support Vector Machine)

Decision Tree Classification

Decision Tree Regression

Random Forest

KMeans Cluster

Proposed Work -

In Regression Multi Linear Regression gives best accuracy.

Proposed Modelling

Multilinear Algorithm is a Supervised learning comes under regression. We used Label encoder and One Hot Encoder to convert multiple number of string value to numerical value. We used sklearn.model_selection to split arrays into testing and training.

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```

	rain	snow
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0
...	...	...
94	1	0
95	1	0
96	1	0
97	1	0
98	1	0

99 rows × 2 columns

Proposed Evaluation

For training we use regressor.fit to fit linear model and train the data. We get accuracy by cost function and gradient descent.
The Cost function is inverse proportional to accuracy. If gradient descent increases ,then cost function is local minima

Cost Function and Gradient Descent

Cost Function shows how our model is predicting compared to original Dataset

Cost=$\frac{1}{m}\sum_{1=0}^{m}$ $|Ypredict - |$ (Linear Regression)

Cost=$\frac{1}{m}\sum_{1=0}^{m}$ $[ylog(Ypredict) + (1-Y)log\ (1-Ypredict)$

$Y_{predict}$ =σ($w^2\ x$ + b), Y=0 $|_{error}$=-log($Y_{predict}$)

By using the Gradient descent algorithm to change w and c value

Cost= $\frac{-1}{m}\sum$ $[y\ log\ y + (1-y)log\ (1-\hat{y})]$

$\hat{y} = \sigma(w^2 x+b)$ $\sigma(x) = \frac{1}{1+e^x}$

$w = w - x[\frac{\alpha coscos\ t}{\alpha coscos\ s}]$

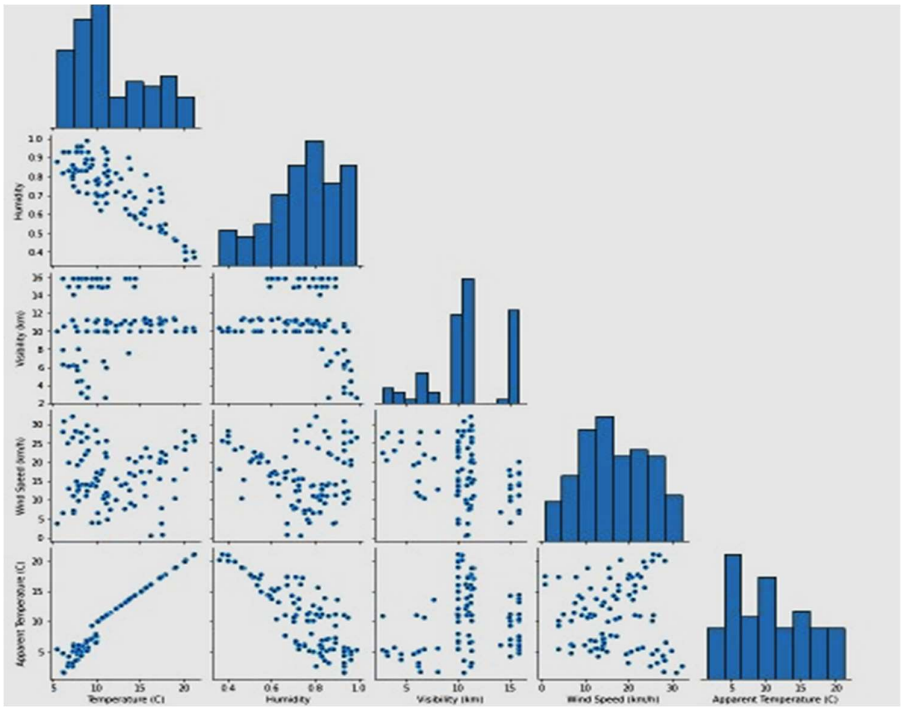
$\frac{d\ cost}{dw}$ ($\hat{y}$-y) x $\frac{d\ cost}{db}$=($\hat{y}$-y)

Proposed Accuracy

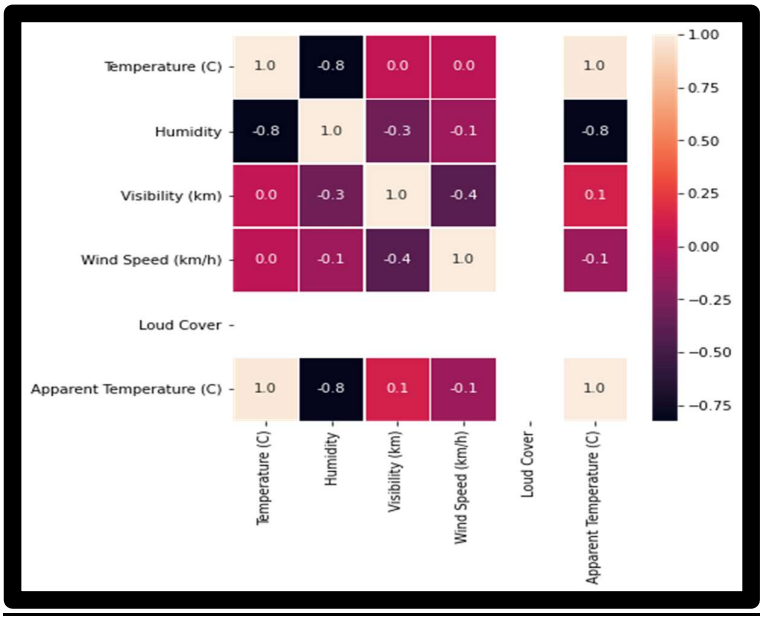
Here Root Square, r2_score () is used to find the accuracy, from sklearn.metrics package. We get an accuracy of 0.9569110317047951

Plot Result and Analysis

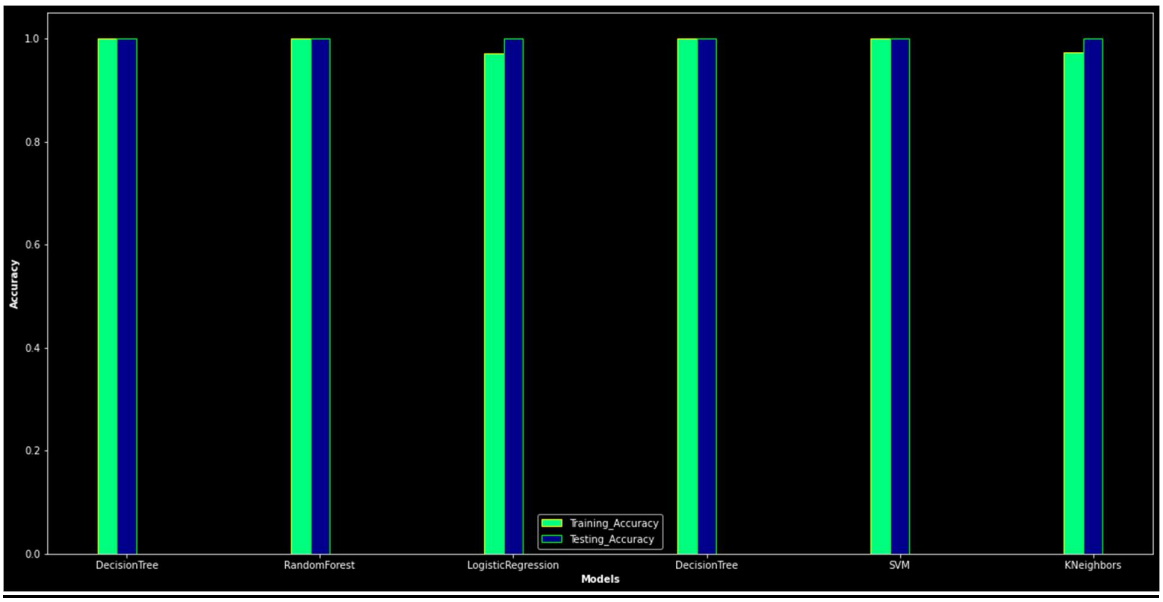
Relationships between different types of column



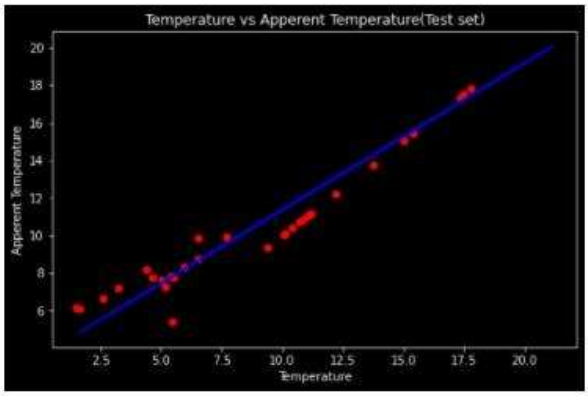
On Observing heat map, we see that for correlated we can't take the relationship whenever light values are there, so we take the thicker values for the relationship to identify the accuracy easily. So we take x – axis as "Temperature(C)" and y axis as "Apparent Temperature(C)"

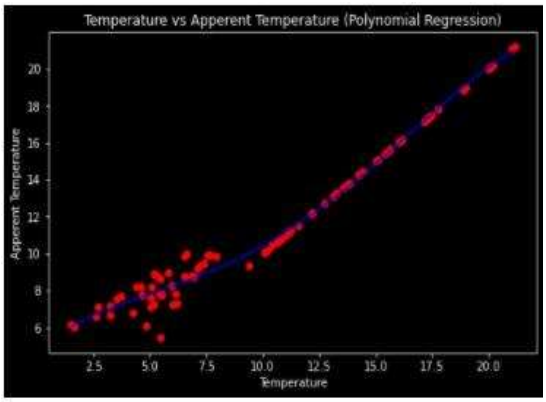


We also do bar graphs of each columns by rc parameters .For Classification we have two columns for classifications value, we convert the precip column of string value to numerical value, since due to lack of classifications value in our Datasets, we don't get better accuracy in every classification for many our accuracy value is overfitting. Here Green Color is Training accuracy and Blue color is testing accuracy.



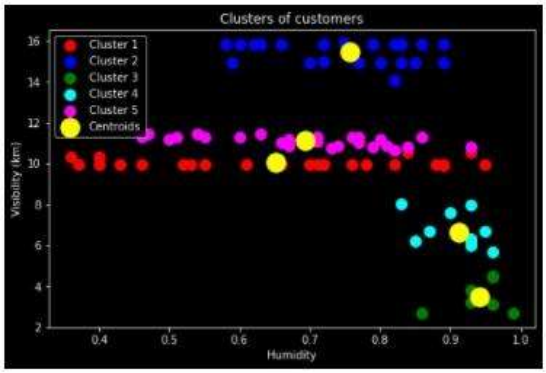
Simple Linear Regression **Polynomial (Non Linear Regression)**

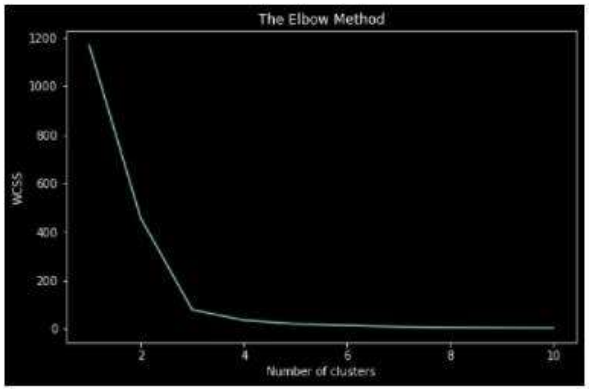




K means Clustering **Elbow Technique**

(Cluster formation with its centroid)





Conclusion

In summary, weather plays a very important role in our daily life and without the meteorologist and forecaster we would have difficult planning our daily activities. As we can see the weather is a simple subject like we may have been thinking. The use of precision agriculture in weather forecasting can help farmers make more accurate management decision as well as economic benefit. Meteorologist and forecasted predict the weather and its possible changes, but in reality weather is still unpredictable.

Future Enhancement:

As we can see due to less accuracy it's very difficult to predict weather, so in future we can update with more accuracy for accurate prediction

In our Project ,For better accuracy we can also take more number of rows or same number of Rows from original Dataset for which we may increase accuracy from 0.95-0.98.

Reference

We get the dataset of our project from kaggle.com and we do coding in Jupyter Notebook

We used 9 Supervised Algorithms where Simple Linear, Multilinear, Polynomial Linear, Decision Tree Regression comes under Regression and KNN, SVM, Logistics regression, Decision Tree Classification, Random Forest under classification.
We used 1 Unsupervised Algorithm which is Kmeans Cluster .
Decision Tree is a CART function.

Submitted to –
Mr. Avinash Seekoli
(Department of CSE)