Using Data Mining for Weather Prediction

CS-C3120 Data Science, Autumn ‘19

Data mining is a computer assisted process of digging through large sets of data and analyzing it. Data mining techniques enables us to predict behaviours and trends which can be used by companies to make proactive decisions.[1] Data mining can help us to predict meteorological data that is weather prediction. Weather prediction has a vital application in meteorology and it has been a challenging problem since the past century. Weather forecasting is the use of science and technology to predict the condition of weather. This project aims to estimate the weather by using predictive analysis. For this it is necessary to use different data-mining techniques to get the most accurate estimate. This report introduces a classifier approach for prediction of weather condition and shows how techniques like PCA, linear regression and binary classification model can be utilized for the prediction purpose.

1. INTRODUCTION

Weather forecasting has been a major troublesome issue all over the world in this century. Environmental changes has caught a great deal of attention among scientists because of the sudden changes that happen. [2] There are several limitations upon developing a reliable and accurate weather model thus it ends up hard predicting weather here and now with effectiveness. The difficulties of weather forecasting, among others, are the enormous weather datasets, many affecting parameters, etc.[2] For this purpose, data mining techniques are performed which enables users to analyze data from a wide range of dimensions or angles, classify it, and condense the connections recognized. Some fundamental terms related to Data Mining are: Classification, Learning and Prediction. Classification is a data mining (machine learning) method used to predict aggregate participation for information cases. Learning refers to training and mapping contribution to yield information. Prediction identifies with modeling and the logical relationship of the model sooner or later. Finding patterns and data may prompt sensible predictions.

In this project we try to use different data mining techniques to develop a model which can predict the relative humidity and classify the weather conditions as “dry” or “not dry”. With this project, we aim to understand the different factors which affect the weather and their net contribution in it. We then will try to apply different techniques to pre-process the data to bring them in the format we plan to work upon. This includes merging and combining different related data together. After preprocessing is done, we try to look for the possible relationships between different aspects of the data by plotting correlation matrix, histograms, etc. We also try to develop an accurate regression model to predict the relative humidity on the test data set. After this, we develop a binary classifier which considers different parameters and classifies the test data. From this project I hope to gain confidence and experience in the data science field. This project will help me learn the usage of different libraries and pre-processing methods which helps during machine learning.

2. DATA ANALYSIS

The dataset consisted of 4 CSV files out of which two were for training purposes and the rest were for test purposes. The names of the files along with the number of columns and rows each of the consisted is given below [Table 1].

File	Rows	Col.
weather_data_train.csv	3140	16
weather_data_test.csv	3140	16
weather_data_train_labels.csv	1346	2
weather_data_test_labels.csv	1346	2

Table 1

The files containing the training and testing data (first two files in above table) consisted of all the different parameters which affects the relative humidity and decides the classification of weather. The rest of the two files consisted of the labels which were either 0 for “not dry” and 1 for “dry”, and relative humidity associated for each row entry in the training data. Initial analysis of training data was done by extracting features such as mean, standard deviation, minimum, maximum, etc. as shown in Fig. 1

	T_var	Po_var	P_var	Ff_var	Tn_var	Tx_var	VV_var	Td_var
count	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000
mean	4.466326	4.062051	4.058249	1.345066	3.662163	4.132288	98.008192	2.768010
std	5.254302	7.278152	7.274098	1.255878	7.034635	5.929607	106.710901	4.497696
min	0.016429	0.005000	0.005714	0.000000	0.000000	0.000000	0.000000	0.008095
25%	0.928571	0.492604	0.500000	0.571429	0.125000	0.382500	20.408036	0.582143
50%	2.568839	1.516964	1.519107	0.982143	0.980000	2.000000	55.928571	1.313125
75%	6.169955	4.395759	4.418750	1.696429	3.920000	5.780000	140.008884	3.089107
max	66.571250	108.117143	108.515536	17.571429	93.845000	105.125000	645.760000	77.449821

	T_mu	Po_mu	P_mu	Ff_mu	Tn_mu	Tx_mu	VV_mu	Td_mu
count	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000	3140.000000
mean	6.780096	758.805939	759.148643	3.674008	4.888025	8.661810	25.952053	3.373812
std	8.595021	8.521047	8.525485	1.222216	8.595641	8.910328	12.230215	8.083700
min	-19.312500	725.525000	725.875000	1.000000	-26.200000	-16.950000	0.625000	-22.912500
25%	0.987500	753.471875	753.821875	2.750000	-0.250000	2.250000	16.421875	-1.612500
50%	6.537500	758.850000	759.187500	3.500000	4.750000	8.125000	24.875000	3.118750
75%	13.953125	764.178125	764.525000	4.375000	11.950000	16.200000	35.375000	10.100000
max	25.787500	790.425000	790.812500	9.750000	23.150000	28.500000	50.000000	21.462500

Fig. 1

2.1 Histogram

A histogram is the most commonly used graph to show frequency distribution. Histogram can prove to be extremely useful when the data is numerical and we want to analyze the shape of data's distribution.[3]

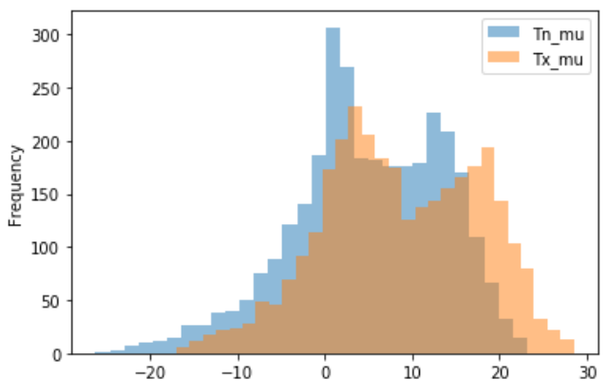


Fig. 2

Above plot [Fig. 2] is of the frequency distribution of Tn_mu and Tx_mu. We can see that the distributions have two maximums i.e Bimodal (double-peaked) Distribution which represents the two extreme seasons - summer and winter. Also, the distribution of both variables follow a similar trend. This is expected because both of them have a high correlation which means that if one rises/falls, the other rises/falls too. We can also see that the frequency of first maxima in both graphs is greater than frequency of second maximum which tells us that the number of cold days are more than hot ones.

2.2 Pair Plots

A pair plot allows us to see both distribution of single variables and relations between two variables together. Pair plots are a great method to identify trends for follow-up analysis. [4]

Pair plots for **T_mu, P_mu, Td_mu, Ff_mu, VV_mu, U_mu** is given below.

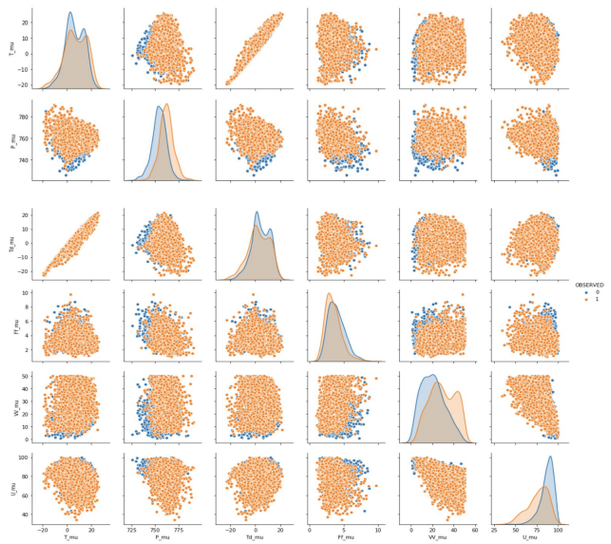


Fig. 3

From the above pair plots [Fig. 3] we can see that most of the data is randomly yet closely distributed without any significant correlation with each other. The only pair plot with any visible correlation is of **T_mu** and **Td_mu** which is expected because of their strong linear relationship in the real world.

2.3 Correlation

The next step would be to see the correlation of different parameters with each other. The correlation tells us the relation between two variables. Positive correlation means that both the variables move in the same direction whereas negative correlation means that they move in opposite directions.[5] We calculate correlation coefficient (**r**) using formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

As we can see from the correlation matrix [Fig 4], some variables have low correlation with each other while some variables hold high correlation with each other. This high correlation between variables brings about a redundancy in the information that can be gathered by the data set. This redundancy can be removed by dimensionality reduction which is described in the next section.

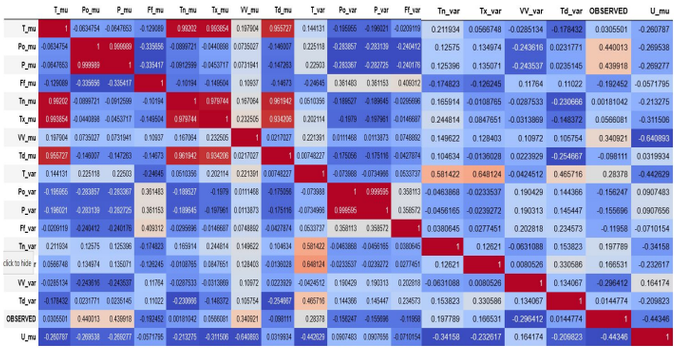


Fig. 4

2.4 Principal Component Analysis (PCA)

The idea behind PCA is simply to find a low-dimension set of axes that summarizes data. It is used when we need to tackle the curse of dimensionality among data with linear relationships. It helps to reduce the computational and cost complexities by transforming the original variables to the linear combination of these variables which are independent. [6]

PCA with 3 components is applied to reduce the dimensionality of the training data. We can see that the 3 components (Figure no. 5) explains around 61% of the original data. Can this value be higher for 3 components? Yes, if the correlation between the variables is higher.

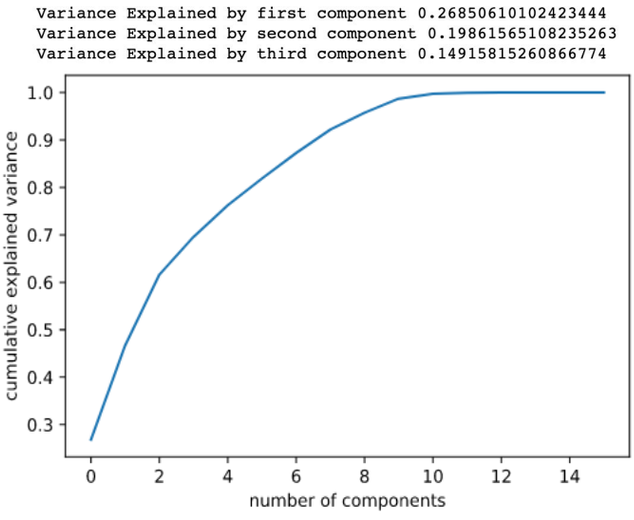
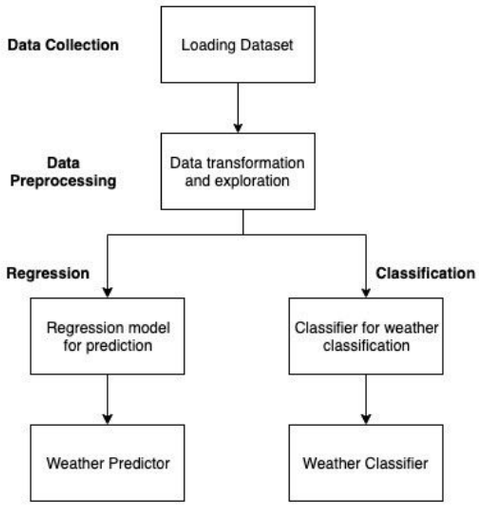


Fig. 5

3. METHODS

In this project, the system predicts the relative humidity based on current weather data which consists of various parameters. The system also classifies the weather based on the weather data into two classes. The different data mining techniques used are PCA, linear regression and K- nearest neighbours. The system methodology is shown in the block diagram.



Methodology Blck Diagram

The first step towards successful completion for this project would be to import important libraries into the system. The several libraries used in this project are pandas, numpy, seaborn, scikit learn and matplotlib.

3.1. Data Collection and Preprocessing

The initial stage of data-mining is data collection and preprocessing. The crucial stage is data preprocessing because valid data will only yield accurate output.

3.1.1 Principal Component Analysis

We start preprocessing by standardizing the data using below given formula. We should standardize the variables before applying PCA because if one component (eg T_mu) varies less than another(eg. W) because of their respective scales, PCA might determine that the direction of maximal variance more closely corresponds with the 'W' axis, if those features are not scaled.[7]

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

After standardizing the resultant data has a mean of zero and standard deviation of one. After this, we will be projecting the 16 dimensional weather data to three-dimensional principal components. The variance explained by each component along with cumulative explained variance has been described in Figure 5.

3.2 Regression

Regression analysis is a powerful statistical method that allows you to examine the relation between two or more variables of interest. The process of performing a regression allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other. [8]

3.2.1 Linear Regression

A linear regression is where the relationships between the variables can be described with a straight line. The equation has the form **Y = aX + b**, where Y is the dependent variable and X is an independent variable.[9]

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

We will now try to train a linear model with the provided training data. The train model is then tested on a test model and the analysis of error is presented below [Fig. 6 & Table 2].

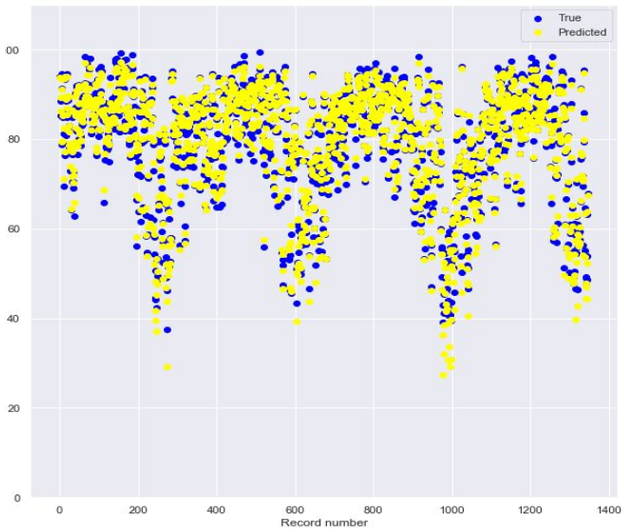


Fig. 6

Result	
RMSE	1.3717
R-Square	0.99

Table 2

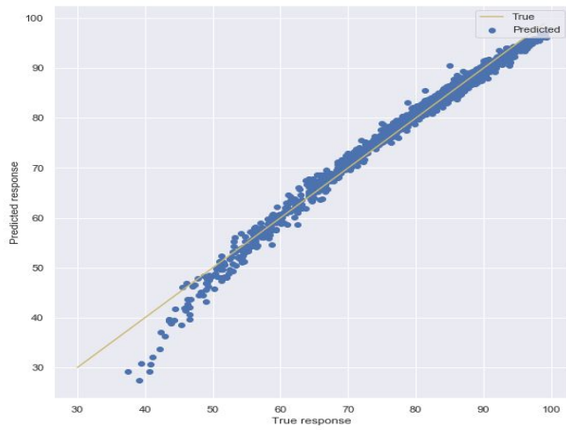


Fig. 7

From the above table [Table 2] we can see that the model performs well. But why? Well, the reason is given in the pair plot. If you see the pair plot closely, we can see the distribution of various parameters with relative humidity is randomly yet closely distributed without any particular direction. In such cases fitting a polynomial function through the data actually proves to work better because it reduces the overall error. Increasing the degree of the polynomial may seem to decrease the error on the training data but in reality it just overfits and thus work poorly on test data.

We now try to see which of the 16 parameters affect the weather significantly. For this, we can look at the correlation matrix and try to rule out some parameters with low correlation with relative humidity. Why would this work? Well, correlation between two variables measures the strength of the linear relationship between them. Since we are training a linear model it makes sense to remove those variables with lesser correlation with relative humidity. From the correlation matrix we can see that variables **VV_mu, T_mu, Td_mu** have the highest correlation with relative humidity. Now we will train the regression model with these parameters and see the RMSE and R-Square values. After that we will marginally add other variables and repeat the analysis to see how much does their addition contribute towards increasing the accuracy.

A. Using VV_mu, T_mu, Td_mu

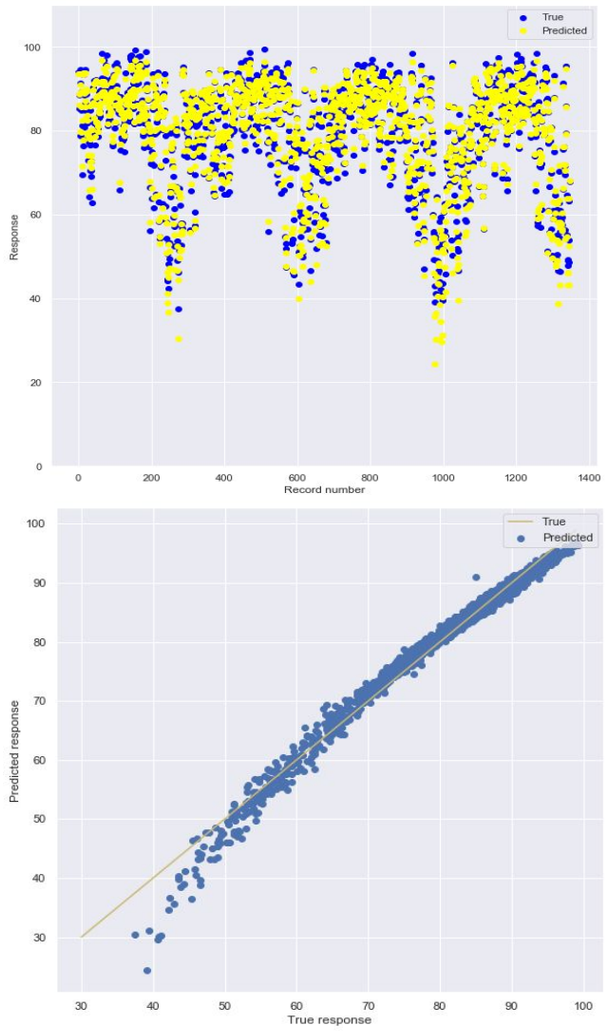


Fig. 8

Result	
RMSE	1.4728
R-Square	0.98

Table 3

B. Adding other variables

Addition of other variables did not contribute much towards reducing the RMSE or increasing the R-square value. This was expected because the correlation of other variables with relative humidity wasn't so significant. With the above analysis [Fig. 8, Table 3] we can say that variables **VV_mu, T_mu, Td_mu** are the strongest predictor of relative humidity.

We saw how we can use correlation between two variables to reduce the dimensionality of our data. In linear regression without reducing dimensionality we achieved RMSE 1.3717 whereas after dimensionality reduction we got a RMSE of 1.4728.

3.2.2 Linear Regression with PCA

In the previous subsection we reduced the dimensionality by handpicking some of the features we thought affected the relative humidity significantly. Now we try to apply PCA and see how our model works. For this we train the model with different number of components and note the optimal number with low RMSE. As seen in given figure [Fig. 9] there is a substantial decrease in RMSE when PCA equals 12. Therefore optimal number of components for linear regression is **12.**

The RMSE with 12 components is around 2, which is greater than what we got in linear regression. Why is that? It is mainly because when we apply PCA on data with low correlation, we lose most of the information.

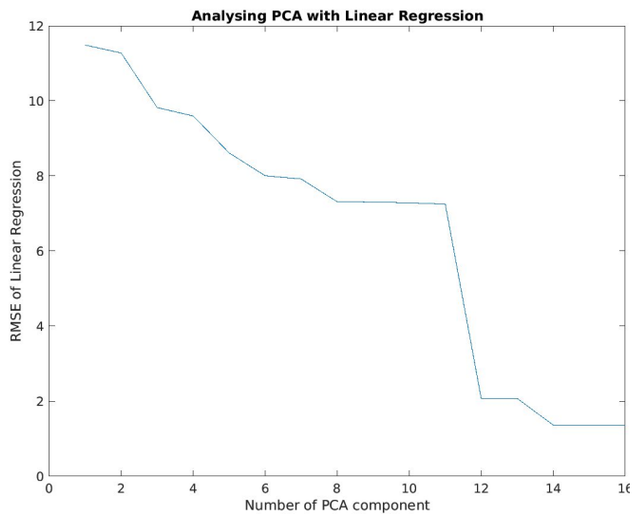


Fig. 9

3.3 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.[10]

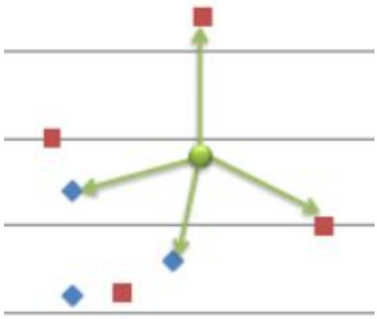
In this project, we aim to create a classifier which classifies weather into 0(not dry) or 1(dry). This kind of classification with two categories is called binary classification. We tested various classification algorithms and the analysis of their accuracy is given below [Table 4].

Classifier	Accuracy
KNN	78.6%
Logistic Regression	79.9%

Table 4

3.3.1 K-nearest neighbour

The K-nearest neighbour is a simple to implement supervised learning algorithm which is used to solve classification problems. The KNN assumes that similar objects exists in close proximity and that is how this algorithm classifies.[11]



The data points neighbouring testing data votes upon its class and the majority wins. K in the name KNN represents the number of neighbouring points who votes. K should be kept an odd number so we can clearly define majority in cases of binary classification like ours. With higher values of K, we get a smoother and well defined boundaries across different classes.

Choosing the right value of K.

To select the optimal value of K, we trained and tested the algorithm several times with different values of K and chose that value for which the error was the least while making prediction on data it hasn't seen earlier. Presented below [Table 5] is the analysis with different values of K.

K	Accuracy (max 100%)
2	72.21%
3	75.55%
4	74.96%
5	76.52%
6	75.78%
7	77.78%
8	76.96%
9	78.60%
10	77.36%

Table 5

We can see from the above table that the maximum accuracy is achieved when K = 9. Also, we can notice that the value of accuracy for even value of K is always smaller than the neighbouring odd values just as we expected.

3.3.2 Logistic Regression

Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. Logistic Regression is primarily based on estimating the parameters of a logistic model.[12] A typical logistic function is represented as:

$$\frac{1}{1+e^{-t}}$$

Where **t** represents a vector of input features. Logistic Regression internally uses a loss function to optimize itself for better classification accuracy, which is defined as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$
$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log (1-h_\theta(x^{(i)}))]$$

The model tries to minimize this convex loss function to reach a global minima and identify a threshold plane to classify the independent variables.

After training, logistic regression gave classification accuracy of 79.9 % which is better than k-NN classifier.

4. RESULTS

This section presents our findings while completing the project. Table presents a summary of those findings.

4.1 Regression

- *Linear regression*
  We found out fitting a linear model on the training data without any dimensionality reduction gave RMSE of 1.3717 with

R-square of 0.99. After reducing the dimensionality to 3 by handpicking some parameters [Table 6] by looking at the correlation matrix, we achieved a RMSE of 1.4728 with R-square of .98. Adding anymore parameters did not affect these readings significantly.

- *Linear Regression with PCA*

  After applying PCA we found out that 12 components (Fig 9) were optimal with RMSE of around 2.

Parameters	RMSE
VV_mu, T_mu, Td_mu	1.6047
VV_mu, Td_mu	10.669
T_mu, Td_mu	10.41
VV_mu, T_mu	8.99
Any other combination of parameters	1.3708 - 1.6047

Table 6

4.2 Classification

- *KNN*

  We found out that the minimum error came for K = 9 with accuracy of 78.60%. The confusion matrix is represented in Fig 12.

```
Confusion Matrix:
[[320 108]
 [195 723]]
```

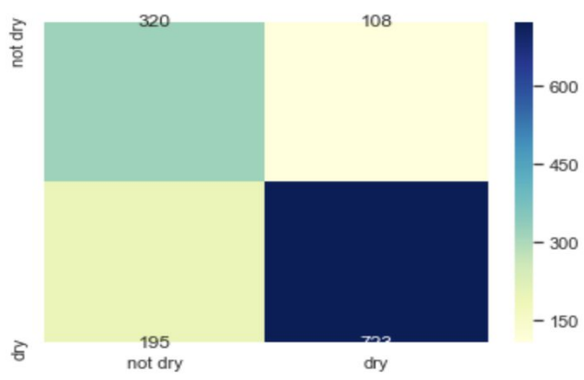


Fig. 12

- *KNN (k =9) with PCA*

  After dimensionality reduction using PCA, we found that 8 components yield maximum accuracy [Fig 10].

- *Logistic Regression*

  Applying logistic regression gave an accuracy of 79.9% which was better than KNN.

- *Logistic Regression with PCA*

  After dimensionality reduction using PCA we found out that minimum 9 PCA components are needed for good accuracy [Fig 11]

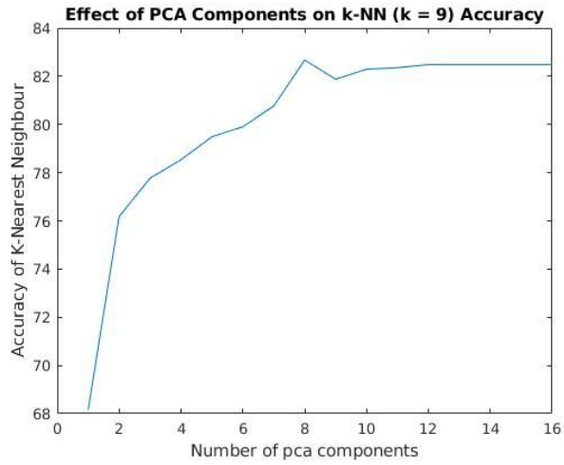


Fig. 10

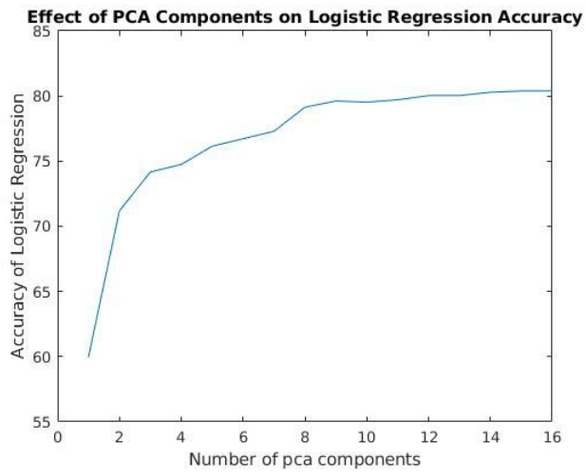


Fig. 11

5. CONCLUSION AND DISCUSSION

In this project we are predicting the relative humidity and classifying weather on parameters like temperature, pressure, air speed, etc. The most surprising, yet expected, finding in the project was the low correlation of variables with the overall weather. Although our model could predict relative humidity with a high accuracy, but in the real world it is difficult to build an accurate weather model. Why? It's because the atmosphere is so vast—it's impossible to observe and know everything about every bit of atmosphere and the planet. If we could

observe everything, then theoretically, we could predict the future with 100 percent accuracy - which we all know is against the rules of time.

We developed a linear regression model for predicting the relative humidity on any given day. While analyzing which parameters are the strongest predictors of relative humidity, we found out that only three parameters, **VV_mu, T_mu, Td_mu,** affected the relative humidity the most. Using these three parameters in a linear regression model gave RMSE of 1.4728 with R-square value of .98. Removing even one of these parameters shot up the RMSE while reducing the R-square value. Adding any new parameter did not significantly contribute in reducing the RMSE. We applied PCA to reduce dimensionality and found out that the optimal number of components to be used is 12 with RMSE of around 2.5.

As per project requirements we developed a weather classifier too. The first classier we used was KNN. We achieved the maximum accuracy with K = 9. The second classifier we used was logistic regression. Both of the models performed well.

REFERENCES

[1.] Alexander, D. Data mining, 1997 [online]. Available from: https://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/ [Accessed 1 December 2019]

[2.] The Gaurdian. The Guardian view on weather forecasts: we need the bigger picture, 2019 [online]. Available from:https://www.theguardian.com/commentisfree/2019/mar/19/the-guardian-view-on-weather-forecasts-we-need-the-bigger-picture [Accessed 1 December 2019]

[3.] ASQ. What is Histogram?[online] Available from: https://asq.org/quality-resources/histogram [cited 1 December 2019]

[4.] Koehrsen, W. Visualizing data with pair plots in python, 2018 [online].Available from: https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf52916 [accessed 2 December 2019].

[5.] Anon., Overview of correlation. [online] Available from: https://www.uv.es/visualstats/vista-frames/help/lecturenotes/lecture11/overview-ovrh.html [accessed 2 December 2019].

[6.] Li, L. Principal Component Analysis for Dimensionality Reduction, 2018 [online] . Available from: https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad [accessed 2 December 2019]

[7.] mixOmics vignette. Principal component analysis [online]. Available from: https://mixomicsteam.github.io/Bookdown/pca.html [accessed 2 December 2019]

[8.] Foley, B. What is Regression Analysis and Why Should I Use It 2018. [online]. Available from: https: //www.surveygizmo.com/resources/blog/regression-analysis/ [accessed 3 December 2019]

[9.] Lumen Learning, Chapter 7: Correlation and Simple Linear Regression [online] . Available from:https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-7-correlation-and-simple-linear-regression/ [accessed 1 December 2019]

[10.] Oracle, Classification [online]. Available from:https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm [accessed 2 December 2019]

[11.] Harrison, O. Machine Learning Basics with the K-Nearest Neighbors Algorithm, 2018 [online] . Available from: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761[accessed 2 December 2019]

[12.] Wikipedia, Logistic regression, 2019 [online] . Available from: https://en.wikipedia.org/wiki/Logistic_regression [accessed 3 December 2019]

Supporting Libraries

1. Scikit-learn : https://scikit-learn.org/stable/
2. Matplotlib: https://matplotlib.org/
3. Seaborn : https://seaborn.pydata.org/
4. Pandas: https://pandas.pydata.org/
5. Numpy: http://www.numpy.org/