A Machine Learning Approach to Weather Prediction with Dimensionality Reduction

Fathima Nazimudeen
fathimanazimudeen@gmail.com

***Abstract*— Weather Prediction is an essential area of analysis in everyday life. The proposed work predicts weather using classification and regression models. PCA is used here as a dimensionality reduction technique. The regression model predicts the value of relative humidity by using multiple linear as well as polynomial regression. The classification model predicts the atmospheric condition as dry or not by using K-Nearest Neighbor (KNN) and Logistic Regression classification. Finally a comparison is made on these algorithms. Experimental results with the application of the principal component analysis method at the stage of pre-processing of the input data are also presented. The experimental results shows that in the case of classification, KNN outperforms Logistic Regression by providing an accuracy of 79.4% without PCA. Also with PCA the same accuracy is obtained with only 8 components. In Regression Analysis, Polynomial regression(degree=2) provides best results than simple linear regression. Also PCA reduces the no: of dimensions from 16 to 11.**

I. INTRODUCTION

Weather forecasting means predicting the weather conditions (conditions of atmosphere) of a particular given area or location. More importantly, accurate weather prediction is very important to pursue day-to-day activities. Weather forecasting has traditionally been done by physical models of the atmosphere, which are unstable to perturbations. Machine learning, on the contrary, is relatively robust to perturbations and doesn't require a complete understanding of the physical processes that govern the atmosphere. Therefore, machine learning may represent a viable alternative to physical models in weather forecasting[1]. This paper explores their application to weather forecasting to potentially generate more accurate weather prediction.

Here in this system we used parameters like temperature, Pressure, Windspeed, etc to predict relative humidity and atmospheric condition whether dry or not through regression and classification tasks.

The objective of my work is to 1) Predict the value of relative humidity by minimizing the cost function and error which is the difference between desired and calculated values. 2) Predict the atmospheric condition as Dry or Not dry using classification models and to find the best models by comparing the result of various algorithms This paper uses the techniques of machine learning algorithms namely Linear Regression and polynomial regression for regression task and K Nearest Neighbor and Logistic Regression for classification task. PCA is used for dimensionality reduction. Results with the application of the principal component analysis method at the stage of pre-processing of the input data are also compared.

This paper is organised as follows: section II describes the data analysis part. Section III describes about the methods used in this work. The experimental results and analysis is shown in section iv and section v explains my conclusions.

II. DATA ANALYSIS

The dataset used in this project have been downloaded from the website http://rp5.ru/. The data are measurements collected by the weather station 2978 in Helsinki from September 2006 to May 2019. The dataset consists of 4486 observations of 16 features which include air temperature in degrees Celsius, 2 meters above the earth's. surface. (T), atmospheric pressure at weather station level, in millimeters of mercury(Po), atmospheric pressure reduced to mean sea level, in millimeters of mercury(P), mean wind speed at a height of 10-12

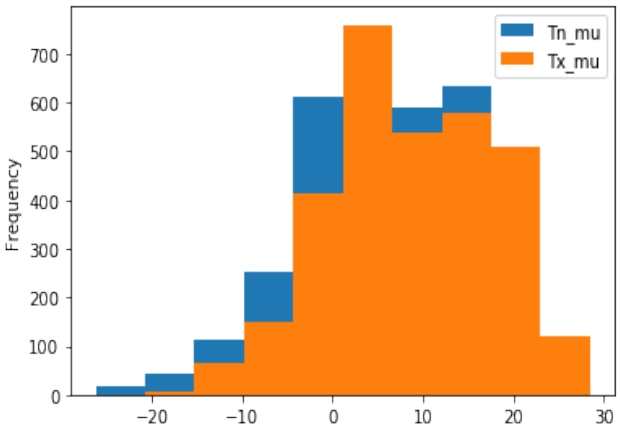


Figure 1: Histogram of Tn_mu and Tx_mu

meters above the earth's surface, in meters per second (Ff), minimum air temperature, in degrees Celsius, over the past day (Tn), maximum air temperature, in degrees Celsius, over the past day(Tx), horizontal visibility, in km(W), dewpoint temperature at a height of 2 meters above the earth's surface, in degrees Celsius (Td), relative humidity, in percentage, 2 meters above the earth's surface.(U),"OBSERVED" is a categorical variable, where 0 (not dry) indicates that the amount of precipitation was more than 0.3 millimeters, and 1 (dry) indicates that there was little or no precipitation.U_mu which represents relative humidity is the dependent variable which is to be predicted using regression algorithms and 'OBSERVED" column is the target categorial column which is to classified using classification algorithms.

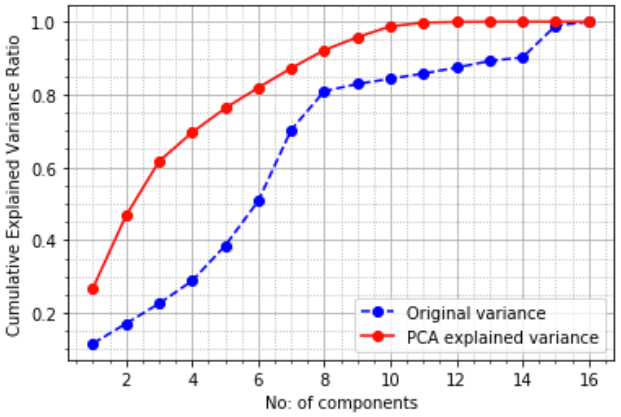


Figure 3: Cumulative Explained variance ratio vs no: of components

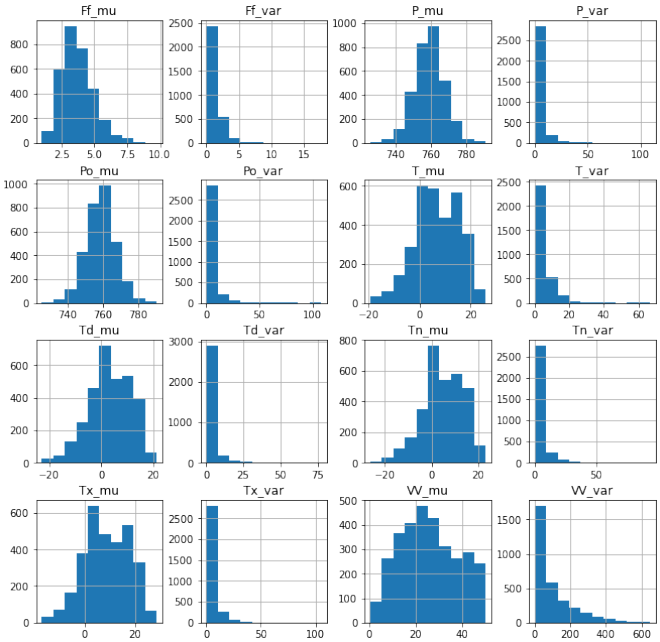


Figure 2: Histogram of Features

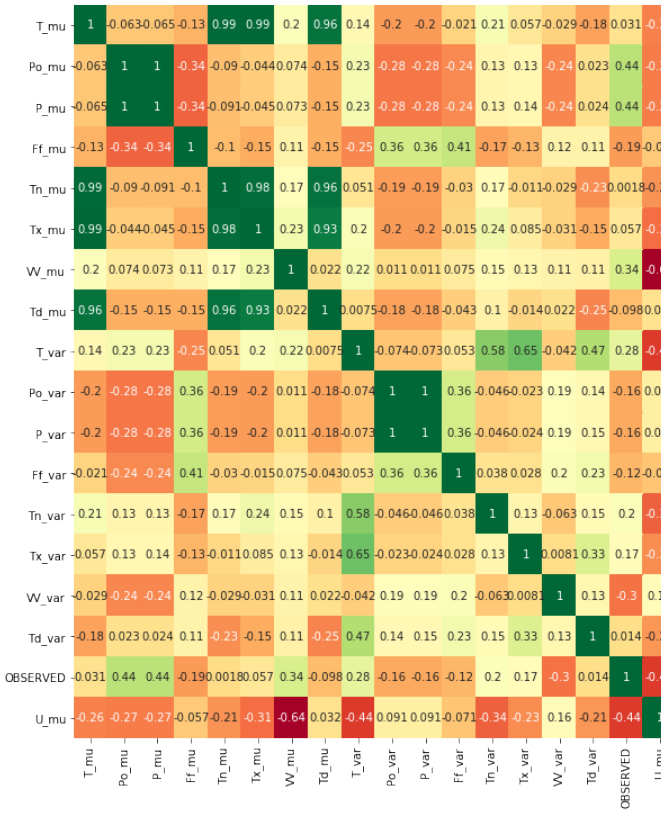


Figure 4: Correlation Plot

For the numerical attributes, the mean and the variance values for each day are provided. For example, “T mu" indicates the mean of the air temperature, and “T var" indicates the variance of the air temperature.

The data is splited into train and test (approximately 70% of the data are the training set, and 30% are the test set). Some visualisations obtained as a while data analysis is shown below. Figure 1 shows the histogram of two features Tn_mu and Tx_mu. The plot shows that both these variables are having a normal distribution. Figure 2 shows the histogram of 16 features. This shows that almost all mean features have normal distribution while variance features have less variance. The feature ‘VV_mu’ has right skewed distribution.

Figure 4 shows the correlation plot of the features 'T_mu', 'P_mu', 'Td_mu', 'Ff_mu', 'VV_mu' and 'U_mu' and Figure 5 shows the corresponding pairplots. In Correlation plot, the value in box indicates the correlation coefficient which measures the strength and direction of the linear relationship between two variables. But these coefficients cannot capture nonlinear relationship between dependent and independent variable.

From the pairplot we can infer that some of the features have linear dependence with the target variable.

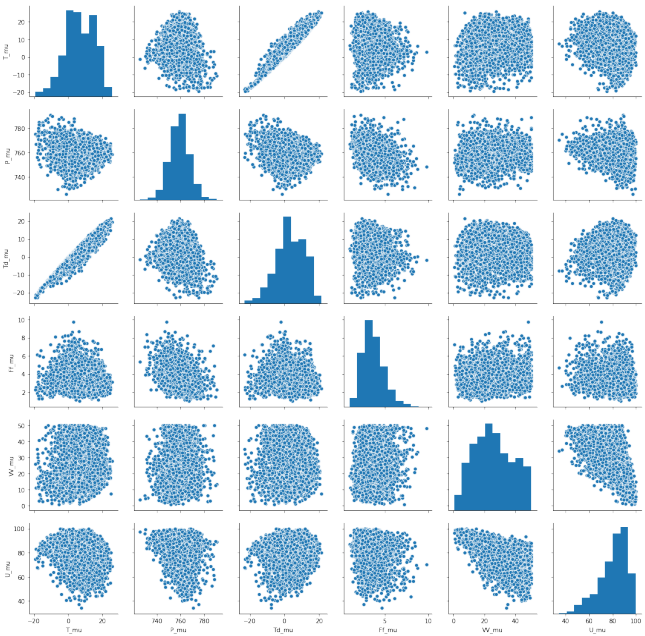


Figure 5: PairPlot

A. Principal Component Analysis (PCA)

Principal Component Analysis(PCA) is one of the most popular linear dimensionality reduction technique[5]. The PCA method defines transformation matrix from the set of input vectors composed of (possibly) correlated components to another set of vectors composed of orthogonal and uncorrelated components. PCA saves the most relevant information from the multidimensional dataset and at the same time reduces its dimension. Some information that is statistically irrelevant is therefore discarded. PCA works by finding the eigenvectors and eigenvalues of the covariance matrix of the dataset[3]. The Eigenvectors are called the “Principal Components” of the dataset.

In this paper, PCA is used to reduce the no of dimensions and no of components for the pca is selected using the information of explained variance. Figure 3 shows the plot of cumulative explained

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Explained Variance	4.30	3.18	2.39	1.27	1.06	0.90	0.86	0.79	0.57	0.48	0.16	0.03	0.01	0.0	0.0	0.0
Explained Variance Ratio	0.27	0.20	0.15	0.08	0.07	0.06	0.05	0.05	0.04	0.03	0.01	0.00	0.00	0.0	0.0	0.0

variance ratio vs no: of components. From this plot we can find that about 11 components can preserve 99% of the variance in the dataset. Thus PCA reduces the dimensionality from 16 to 11. Figure 6 shows the PCA plot of first 2 components where yellow and blue dots represents two categories of observed variable. From the plot we can infer that PCA is not providing a good separation of the 2 classes. The obtained explained variance and its ratio is shown below.

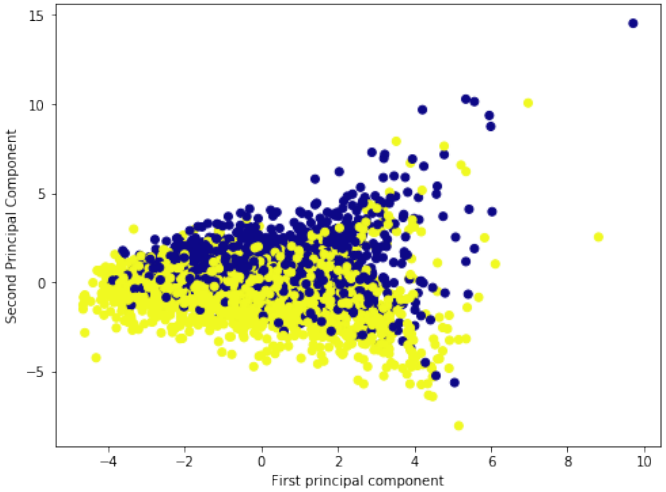


Figure 6: PCA plot of first 2 components

III. METHOD

The flowchart of the whole process in shown in Figure 7 . The main algorithms used for classification and regression analysis is explained below.

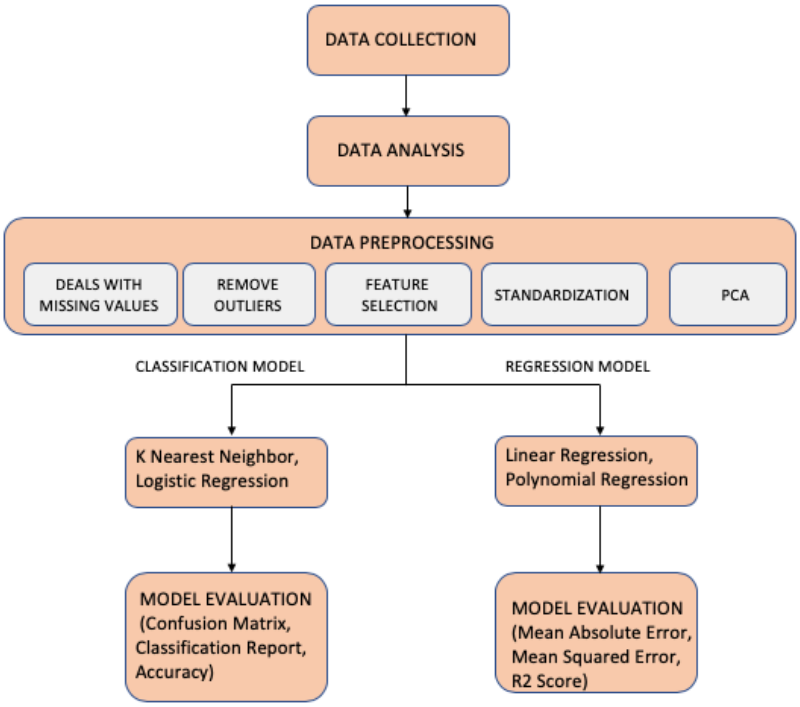


Figure 7: Flowchart of the proposed work

1. Regression Model

The target variable for regression analysis in our study is the relative humidity. The first algorithm that was used was linear regression, which seeks to predict the relative humidity as a linear combination of the features[6]. The second method used linear regression with polynomial features. Then I implement these two algorithms after applying PCA to the input dataset.

Linear Regression:

It is a linear method used for defining the relation between a dependent variable (Y) and one or more independent variables or explanatory variables, denoted by (X). If there are multiple independent variables, the process is defined as Multiple Linear Regression (MLR). The general equation for a linear regression is given as

$$y = B_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n$$

Where y is the dependent variable, X_i's are the n independent variables and B_i's are the corresponding coefficients where i = (1, 2, … , n), B_0 is the intercept. In this work there are 16 features and the corresponding coefficients are shown in Figure 8. It shows that Td_mu has the largest coefficient and so it is a strong predictor of relative humidity.

Polynomial Regression:

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modelled as an *nth* degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y. In this work, polynomial regression with degree 2 (Quadratic)is used.

2. Classification Model

The Classification model is used to classify the atmospheric condition as dry or not dry depending on the input features. This paper uses 2 algorithms for classification namely K-nearest neighbors and Logistic Regression.

K-Nearest Neighbors:

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set[1]. To classify a data sample X, search is done for its K-nearest neighbors and then X is assigned to a class label to which

Variable	Coefficients
T_mu	-4.378
Po_mu	0.958
P_mu	-0.986
Ff_mu	-0.207
Tn_mu	0.278
Tx_mu	-0.337
VV_mu	-0.062
Td_mu	4.554
T_var	0.056
Po_var	0.159
P_var	-0.154
Ff_var	-0.026
Tn_var	0.044
Tx_var	-0.008
VV_var	0.0008
Td_var	0.046

Figure 8: Coefficients of Linear Regression

majority of its neighbors belong. In this method, the choice of k also affects the performance of k-nearest neighbor algorithm. The steps in this algorithm are:

1. Choose the value of K (no: of neighbors)
2. Calculate the distance between input sample and training samples
3. Sort the distances.
4. Take top K- nearest neighbors.
5. Apply simple majority.
6. Predict class label with more neighbors for input sample.

This paper uses the error method to find the value of K[2]. This method finds the error for all possible values of K and find the best K which gives the minimum error. Figure 9shows that the optimum value of K for this work is 36.

Logistic Regression:

It's a classification algorithm, that is used where the response variable is *categorical*. The idea of Logistic Regression [7] is to find a relationship between features and probability of particular outcome. In this work we have to predict atmospheric condition as dry or not sry. This type of a problem is referred to as Binomial Logistic Regression, where the response variable has two values 0(Dry) and 1 (Not dry). Multinomial Logistic Regression deals with situations where the response variable can have three or more possible values.

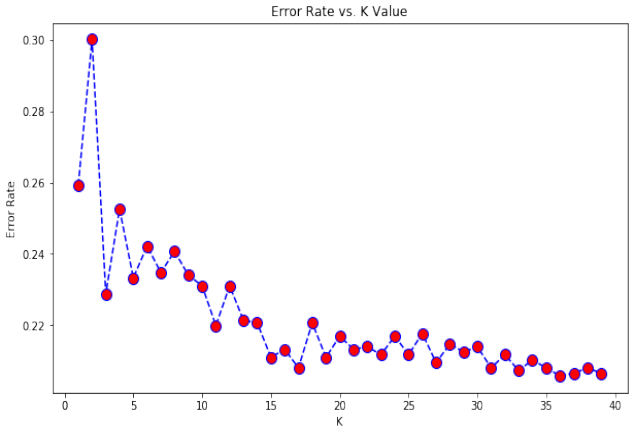


Figure 9: Error Vs K value

The logistic function can be expressed as

$$\log\left(\frac{p(X)}{1-p(X)}\right) = B_0 + B_1 X$$

where, the left hand side is called the logit or log-odds function. And its inverse is known as the *Sigmoid function* and it gives an S-shaped curve and is given by

$$p(X) = \frac{e^y}{1+e^y}$$

Where y=$B_0 + B_1 X$

IV. Experiments and Results

A. Regression Model

In regression analysis, performance metrics used to evaluate our model are

1. Mean Absolute Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(Predicted - Actual)$$

2. Mean Squared Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(Predicted - Actual)^2$$

3. R2 Score

$$R^2 = 1 - \frac{Explained\ Variation}{Total\ Variation}$$

Table 1 shows the results of regression analysis.

Table 1:Regression Results

	Linear Regression		Polynomial Regression	
	Before PCA	After PCA	Before PCA	After PCA
MAE	1.1157	1.1159	0.335	0.333
MSE	2.429	2.431	0.276	0.274
R2 Score	0.984	0.984	0.998	0.998

The plots of the results are shown in figure 10 and 11. The results shows that as no: of components increases error decreases. Also the same error that obtained without PCA can be achieved with only 11 components. That is PCA reduced the no: of features from 16 to 11.

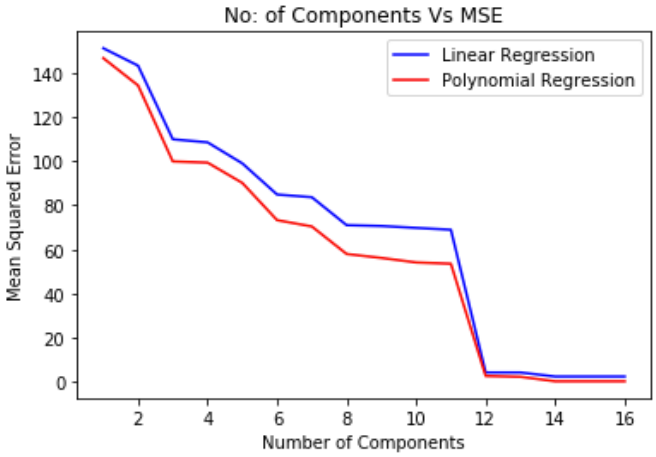


Figure 10: Mean Squared Error Vs No: of components in regression analysis

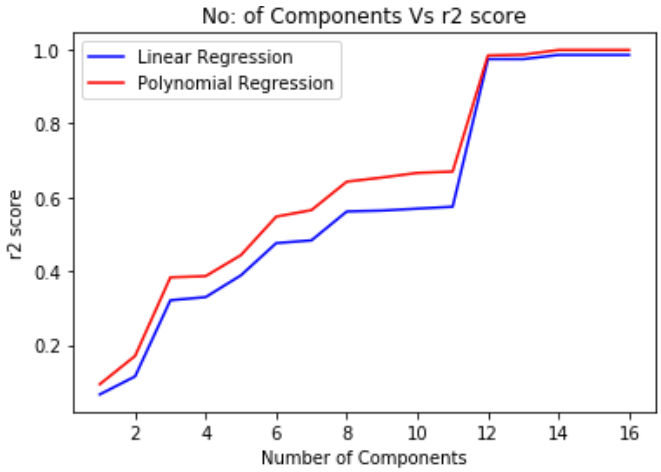
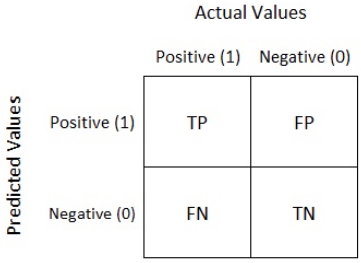


Figure 11: r2_score Vs No: of components in regression analysis

B. Classification Model

In Classification analysis, performance metrics used to evaluate our model are

1. Confusion matrix



Where TP= no: of true positives
TN = no: of true negatives
FP = no: of false positives
FN = no: of false negatives.

2. Classification report

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 * \frac{precision * recall}{precision + recall}$$

3. Accuracy

$$Accuracy = \frac{No: of\ correct\ Predictions}{Total\ no: of\ predictions}$$

Table 2 shows the results of classification analysis.

.

Table 2: Classification Results

	KNN		Logistic Regression	
	Before PCA (16 features)	After PCA (8 Components)	Before PCA (16 features)	After PCA (13 component)
Accuracy	0.794	0.7942	0.7949	0.7949
Precision	0.866	0.866	0.864	0.864
Recall	0.825	0.825	0.828	0.828
F1 score	0.845	0.845	0.846	0.846
rocauc	0.776	0.776	0.775	0.775
logloss	7.107	7.107	7.082	7.082

Table 3: Results with 11 PCA components

	KNN	Logistic Regression
Accuracy	0.797	0.789
Precision	0.861	0.867
Recall	0.837	0.816
F1 Score	0.849	0.841
rocauc	0.773	0.774
logloss	7.005	7.26

The confusion matrix obtained are

KNN:

Before PCA: $\begin{bmatrix} 311 & 160 \\ 117 & 758 \end{bmatrix}$;

After PCA (with 8 components): $\begin{bmatrix} 296 & 144 \\ 132 & 774 \end{bmatrix}$

Logistic Regression:

Before PCA: $\begin{bmatrix} 309 & 157 \\ 119 & 761 \end{bmatrix}$;

After PCA (with 11 components): $\begin{bmatrix} 313 & 168 \\ 115 & 750 \end{bmatrix}$

Logistic Regression with C=7.753

Before PCA: $\begin{bmatrix} 309 & 152 \\ 119 & 766 \end{bmatrix}$;

After PCA (with 15 components): $\begin{bmatrix} 309 & 152 \\ 119 & 766 \end{bmatrix}$

Accuracy before PCA=0.7986

Accuracy after PCA=0.7986

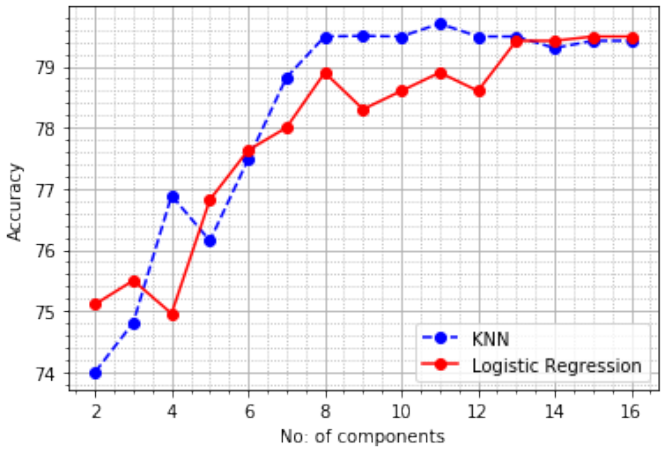


Figure 12: Accuracy Vs No: of components

Figure 12 shows the accuracy obtained vs no: of components. In the case of KNN classification, before applying PCA accuracy obtained was 79.42%. From Table 2 we can see that, After applying PCA the same accuracy can be attained with only 8 components. Also with 11components (Table 3), accuracy increased to 79.7%. In the case of Logistic Regression, before applying PCA accuracy was 79.49%. After applying PCA the same accuracy can be achieved with only 13 components. So we can said that PCA performs best in this classification process and also KNN outperforms Logistic regression. Also precision, recall, are all improved by using PCA.

V. Conclusion and Discussion

An approach to the weather prediction problem using machine learning algorithms and dimensionality reduction technique is presented. The experimental results shows that linear regression with polynomial features has better performance than simple linear regression. Also with the application of PCA, the same error that obtained without PCA can be achieved with only 11 components. In the case of classification, KNN outperforms Logistic Regression by providing an accuracy of 79.4% without PCA. Also with PCA the same accuracy is obtained with only 8 components. With PCA accuracy is also improved to 79.7% by using 11 components. PCA thus reduces the dimension from 16 to 11.

References

[1]. Nishchala C. Barde , Mrunalinee Patole, "Classification and Forecasting of Weather using ANN, k-NN and Naïve Bayes Algorithms", International Journal of Science and Research (IJSR) 2013

[2] Python for Data Science and Machine learning Bootcamp, www.udemy.com

[3] CC-C3160 data science, Computer exercises

[4]Mark Holmstrom, Dylan Liu, Christopher Vo, "Machine Learning Applied to Weather Forecasting"

[5]Meigarom Lopes, "Dimensionality reduction-Does PCA really improve dimensionality reduction", https://towardsdatascience.com/dimensionality-reduction-does-pca-really-improve-classification-outcome-6e9ba21f0a32

[6] George Seif,"5 types of Regression and their properties", Towardsdatascience,march 26,2018

[7] Saishruthi Swaminathan,"Logistic Regression-Detailed Overview", Towardsdatascience,march 15,2018