**RAINFALL PREDICTION SYSTEM**

**Submitted by**

Magilan S - 16BLC1005

Sri Charan BVS - 16BLC1093

Ashish Kumar Sahu - 16BLC1077

**J Component - Report**

**ECM2002 – Machine Learning Algorithms**

**BACHELOR OF TECHNOLOGY**

in

**ELECTRONICS AND COMPUTER ENGINEERING**



**October 2018**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter** | **Title** | **Page** |
|  |  |  |
| 1 | ABSTRACT | 3 |
|  |  |  |
| 2 | SECTION I - Data Set Description | 3 |
|  | 2.1 Predictors | 3 |
|  | 2.2 Target | 4 |
| 3 | SECTION II - Data Preprocessing | 4 |
| 4 | SECTION III - Basic Models | 6 |
|  | 3.1 Logistic Regression | 6 |
|  | 3.2 KNN  3.3 Decision Tree | 7  8 |
|  |  |  |
| 5 | SECTION IV – Alternate/Advanced Models | 11 |
|  | 4.1 Random Forest | 11 |
|  | 4.2 SVM  4.3 Principal Components Analysis  4.4 Cross Validation | 12  14  17 |
|  |  |  |
| 6 | SECTION V – Comparison & Conclusion | 19 |
|  | 5.1 Comparison | 19 |
|  | 5.2 Inference | 20 |

1. **ABSTRACT:**

Rainfall prediction is helpful to avoid flood which save lives and properties of humans. Moreover, it helps in managing resources of water. Information of rainfall in prior helps farmers to manage their crops better which result in growth of country’s economy. Fluctuation in rainfall timing and its quantity makes rainfall prediction a challenging task for

meteorological scientists.

In all the services provided by meteorological department, Weather forecasting stands out on top for all the countries across the globe. The task is very complex as it requires numbers of specialized and also all calls are made without any certainty.

Objective of this project is to predict the rainfall of a localised area by observing the weather parameters such as temperature, humidity and wind speed of the particular area.

The observed data will be subjected to a machine learning model to infer the weather pattern of the place and predict the chance of rain in the near future.

1. **DATASET DESCRIPTION:**

* Daily weather data for the Austin KATT station from 21-2-2013 to 31-7-2017.
* To find whether it rains or not on a particular day based on the subsequent weather information.
* The dataset has 1100 records of data and 20 fields.
* It is a binary classifier problem.

**PREDICTORS:**

* Date (YYYY-MM-DD)
* TempHighF (High temperature, in Fahrenheit)
* TempAvgF (Average temperature, in Fahrenheit)
* TempLowF (Low temperature, in Fahrenheit)
* DewPointHighF (High dew point, in Fahrenheit)
* DewPointAvgF (Average dew point, in Fahrenheit)
* DewPointLowF (Low dew point, in Fahrenheit)
* HumidityHighPercent (High humidity, as a percentage)
* HumidityAvgPercent (Average humidity, as a percentage)
* HumidityLowPercent (Low humidity, as a percentage)
* SeaLevelPressureHighInches (High sea level pressure, in inches)
* SeaLevelPressureAvgInches (Average sea level pressure, in inches)
* SeaLevelPressureLowInches (Low sea level pressure, in inches)
* VisibilityHighMiles (High visibility, in miles)
* VisibilityAvgMiles (Average visibility, in miles)
* VisibilityLowMiles (Low visibility, in miles)
* WindHighMPH (High wind speed, in miles per hour)

**TARGET:**

* Rainfall Status - Whether it rained or not

**SOURCE:**

Kaggle - <https://www.kaggle.com/grubenm/austin-weather>

1. **PRE - PROCESSING:**

After the data is selected, it has to be transformed into the form the user can use it. This preprocessing step is about getting the selected data into a form that you can work.

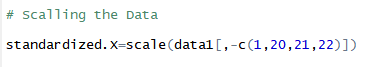
Three common data preprocessing steps are formatting, cleaning and sampling:

* **Formatting**: The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.
* **Cleaning**: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely
* **Sampling**: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

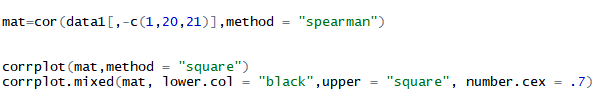
To remove the empty fields,

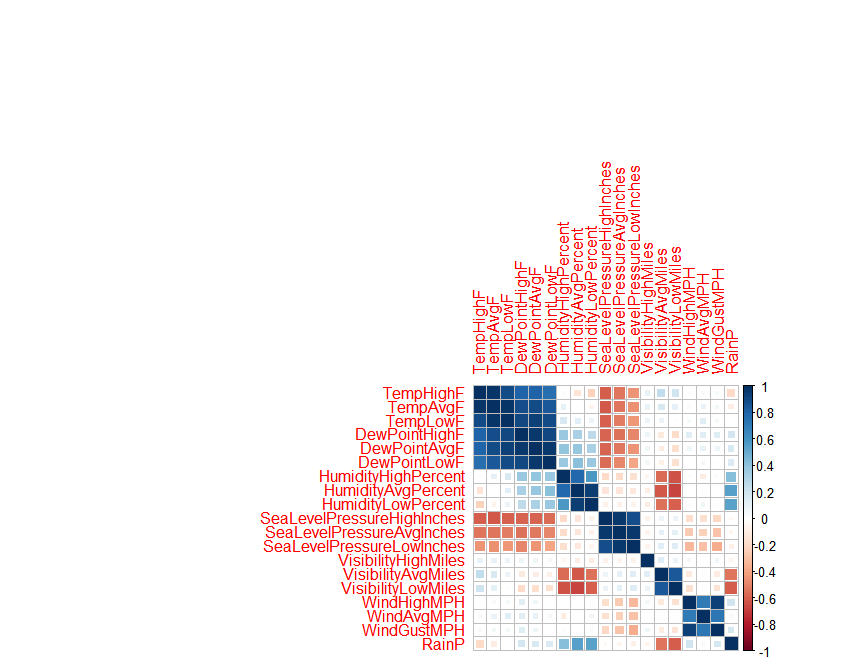
****

To standardise the data,



To find the correlation among the fields,





1. **BASIC MODELS:**
2. **Logistic Regression:**

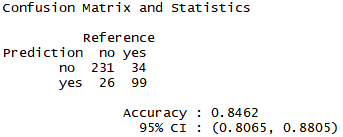
[Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

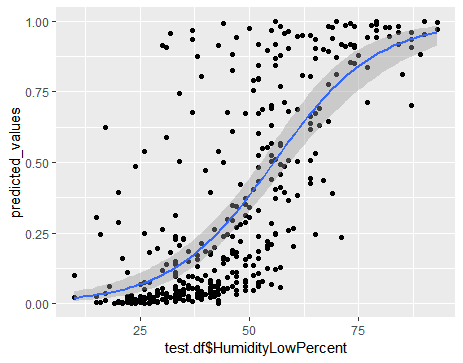
**Assumptions:**

* + The dependent variable should be dichotomous in nature (e.g., YES vs NO).
  + There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores.
  + There should be no high correlations (multicollinearity) among the predictors.  This can be assessed by a correlation matrix among the predictors. Tabachnick and Fidell (2013) suggest that as long correlation coefficients among independent variables are less than 0.90 the assumption is met.
  + At the center of the logistic regression analysis is the task estimating the log odds of an event.  Mathematically, logistic regression estimates a multiple linear regression function defined as:



**Accuracy:**

****

****

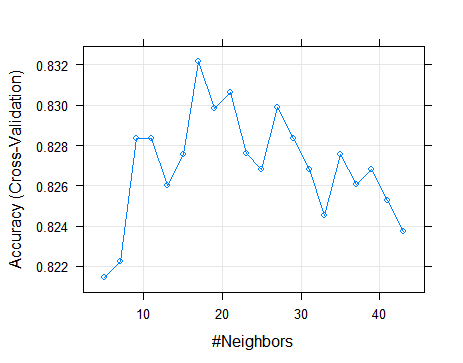
1. **K – Nearest Neighbours:**

The k-NN is a nonparametric method that bases its prediction on the target outputs of the k-nearest neighbors of the given query point (Hastie et al., 2009). Specifically, given a data point, we compute the Euclidean distance between that point and all points in the training set. We then pick the closest k training data points and set the prediction as the average of the target output values for these k points



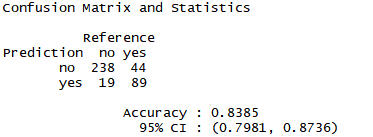
In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

**K – Selection:**

****

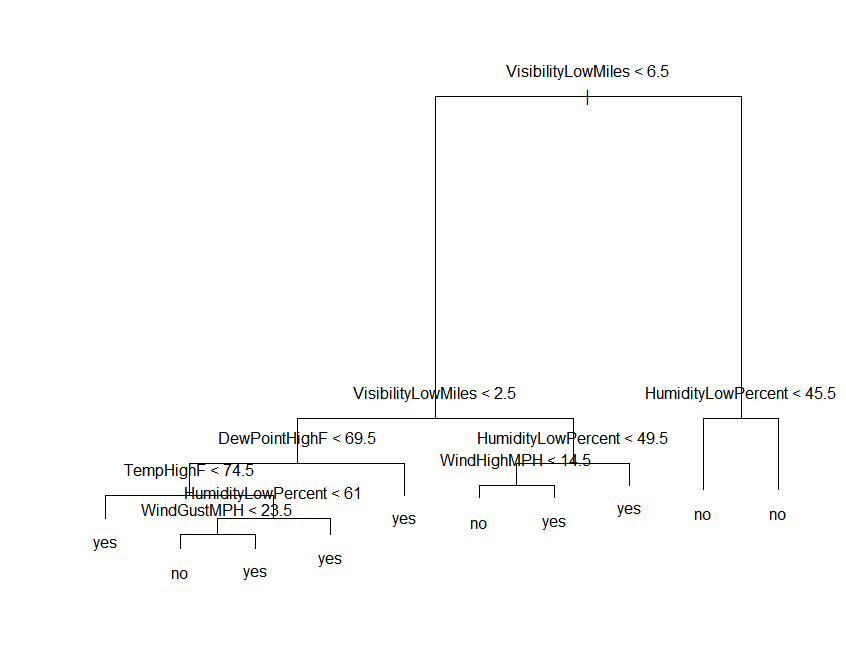
The accuracy is maximum for the model when K = 19.

**Accuracy:**

****

1. **Decision Tree:**

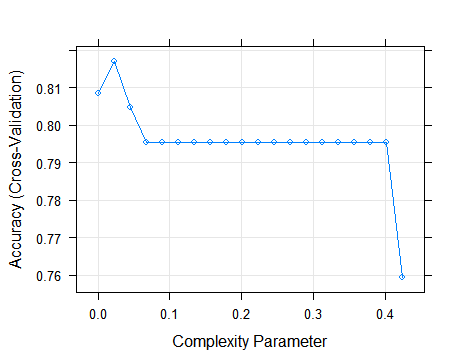
Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

****

Decision tree with all the fields included.

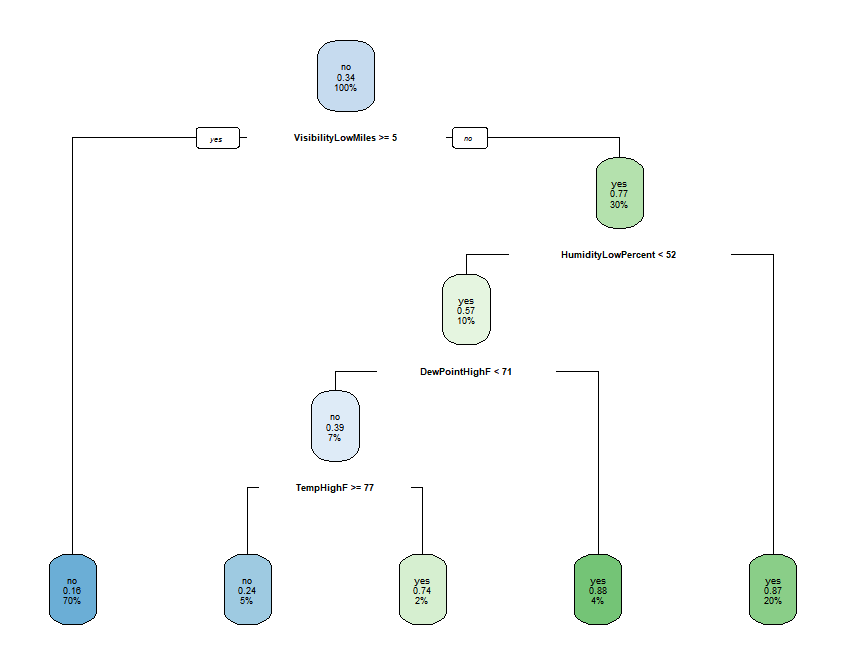
**Prunning:**

To find the complexity parameter at which the accuracy in maximum.



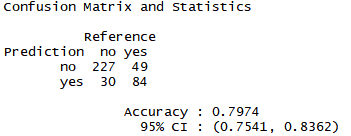
The accuracy is maximum when the when the complexity parameter is 0.0202.

Prunning at that weight will give the max. accuracy.

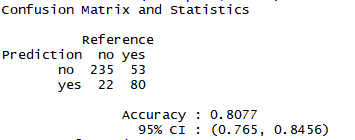


**Accuracy:**

**Before Prunning,**



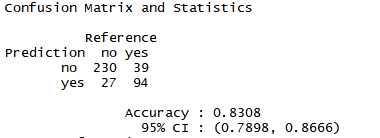
**After Prunning,**

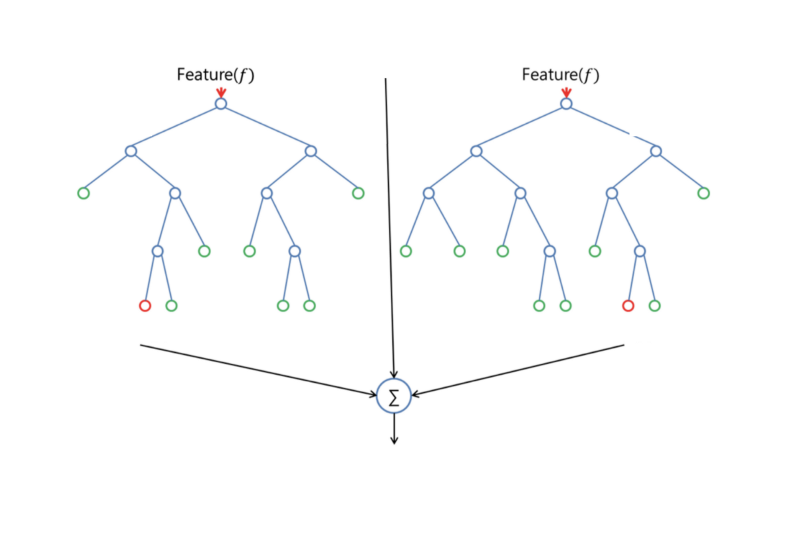


1. **ADVANCED MODELS:**
2. **Random Forest:**

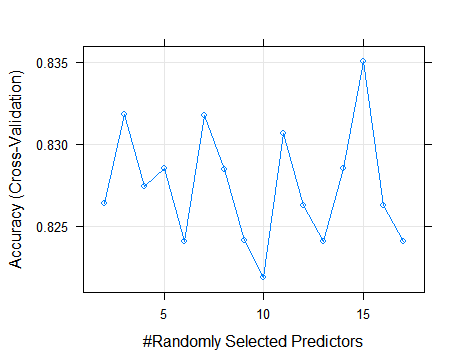
**Random forests** or **random decision forests** are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks, that operate by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting" \o "Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

**Accuracy:**

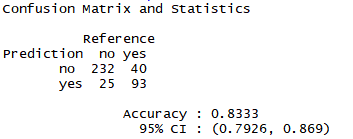
****



**Randomly selected predictors:**

****

Max. accuracy when mtry = 15



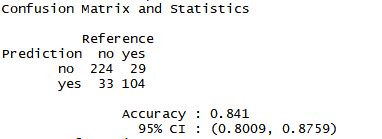
1. **SVM (Support Vector Machine):**

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However,  it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes

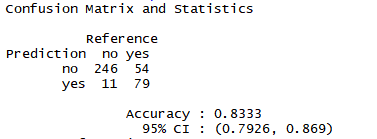
Different SVM Kernels,

* Linear
* Polynomial
* Radial

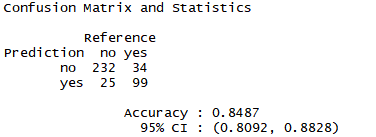
**SVM with Linear Kernel Accuracy,**

****

**SVM with Polynomial Kernel with degree = 3,**

****

**SVM with Radial Kernel Accuracy,**

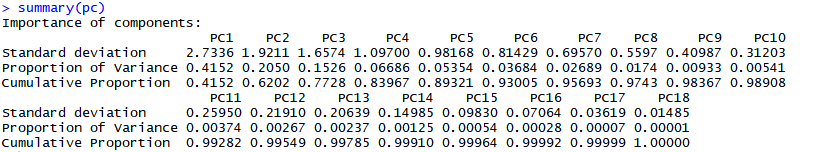
****

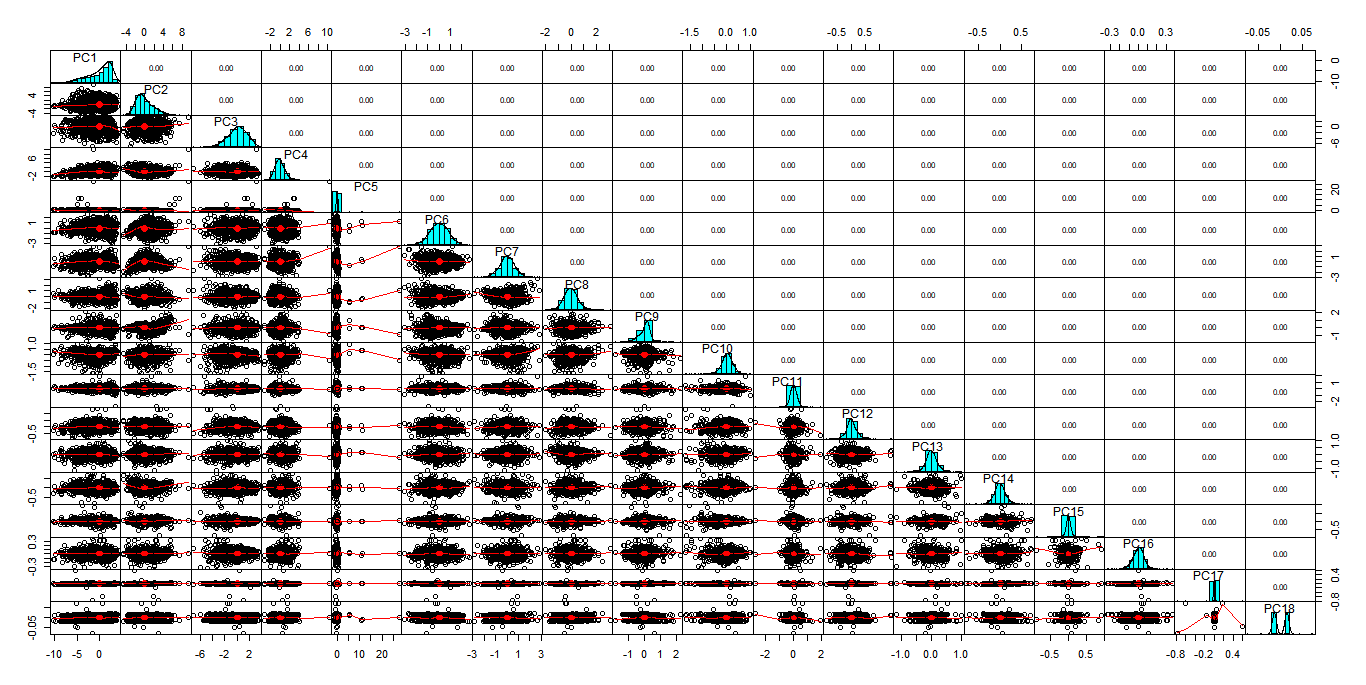
**PRINCIPAL COMPONENT ANALYSIS:**

Principal component analysis (PCA) is a statistical procedure that uses an [orthogonal transformation](https://en.wikipedia.org/wiki/Orthogonal_transformation) to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of [linearly uncorrelated](https://en.wikipedia.org/wiki/Correlation_and_dependence) variables called principal components.

This transformation is defined in such a way that the first principal component has the largest possible [variance](https://en.wikipedia.org/wiki/Variance) (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is [orthogonal](https://en.wikipedia.org/wiki/Orthogonal) to the preceding components.

The resulting vectors (each being a [linear combination](https://en.wikipedia.org/wiki/Linear_combination) of the variables and containing *n* observations) are an uncorrelated [orthogonal basis set](https://en.wikipedia.org/wiki/Orthogonal_basis_set). PCA is sensitive to the relative scaling of the original variables.





After performing PCA for the given fields the correlation between each principal components has become **zero.**

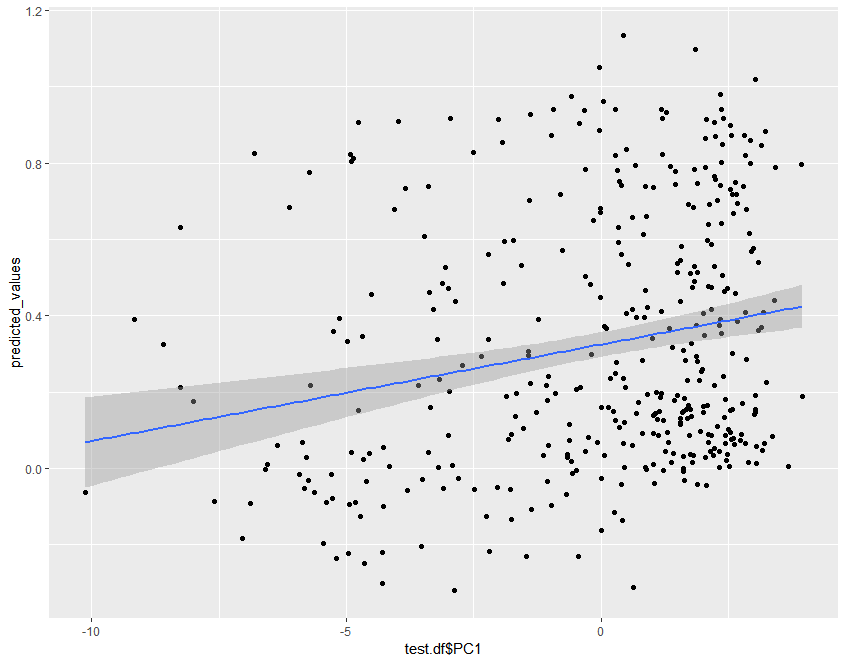
This avoids the multi-collinearity problem and helps us to find the fields with more weightage.

It also helps us to weed out outliers from our dataset.

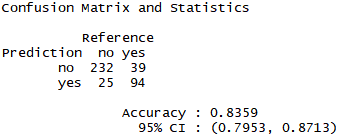
****

It shows the correlation of each predictor and data points which lie in PC 1 and PC2.

**Logistic Regression with PCA: (All 18 principal components)**

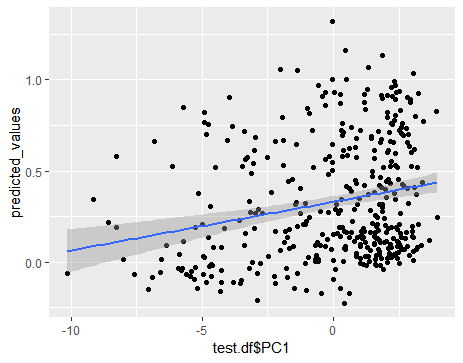
****

**Accuracy:**

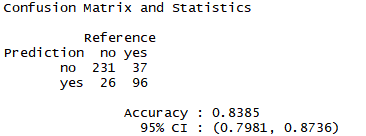
****

**Logistic Regression with PCA: (PC1 to PC10)**

Covering 98% of the variance,



**Accuracy:**

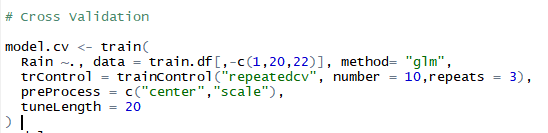
****

**CROSS – VALIDATION:**

Cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

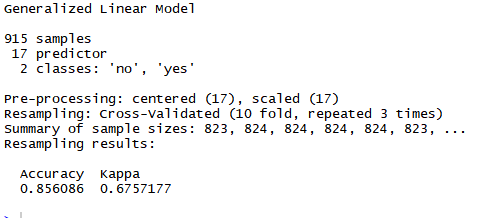
Suppose we have a [model](https://en.wikipedia.org/wiki/Statistical_model) with one or more unknown [parameters](https://en.wikipedia.org/wiki/Parameters), and a data set to which the model can be fit (the training data set). The fitting process [optimizes](https://en.wikipedia.org/wiki/Optimization_(mathematics)) the model parameters to make the model fit the training data as well as possible. If we then take an [independent](https://en.wikipedia.org/wiki/Independence_(probability_theory)) sample of validation data from the same [population](https://en.wikipedia.org/wiki/Statistical_population) as the training data, it will generally turn out that the model does not fit the validation data as well as it fits the training data. The size of this difference is likely to be large especially when the size of the training data set is small, or when the number of parameters in the model is large. Cross-validation is a way to estimate the size of this effect.

**Logistic Regression:**

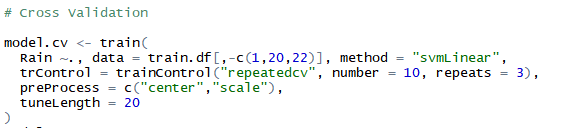
Performing K fold cross validation, ****

Where K = 10 and 3 repetition

**Accuracy after CV:**

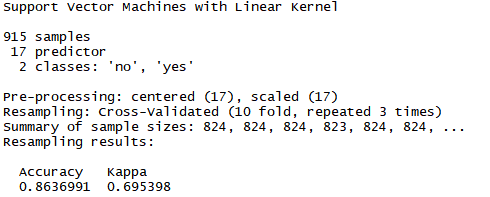


**SVM: (Linear kernel)**

****

Where K = 10 and 3 repetition

**Accuracy after CV:**

****

1. **COMPARISON & CONCLUSION**

After training various models to predict the rainfall status of a particular place, the models has to compared with each other to find model with the least RMSE or the model with maximum accuracy

|  |  |  |
| --- | --- | --- |
| **SL.No** | **Models** | **Accuracy (in %)** |
| 1 | Logistic Regression | 84.62 |
| 2 | KNN | 83.85 |
| 3 | Decision Tree | 80.77 |
| 4 | Random forest | 83.33 |
| 5 | SVM Linear Kernel | 84.10 |
| 6 | SVM Polynomial Kernel | 83.33 |
| 7 | SVM Radial Kernel | 84.87 |
| 8 | Logistic regression with PCA | 83.85 |
| 9 | Logistic Regression CV | 85.60 |
| 10 | SVM CV | 86.36 |

From the above table it is clear that the Logistic regression and SVM are best suited to predict the rainfall for the given dataset.

**Inference:**

* The accuracy of logistic regression with the actual predictor is greater than the accuracy of logistic regression with Principal components, this implies that there are very less outliers in the dataset.
* SVM with radial kernel gives the best accuracy for both training and testing.
* K – fold cross validation with k value equal to 10 and repetition equal to 3 is performed for logistic regression and SVM. The corresponding accuracies were 85.60 and 86.36 respectively.