# Customer Churn Prediction

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. The main contribution of our work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely subject to churn.

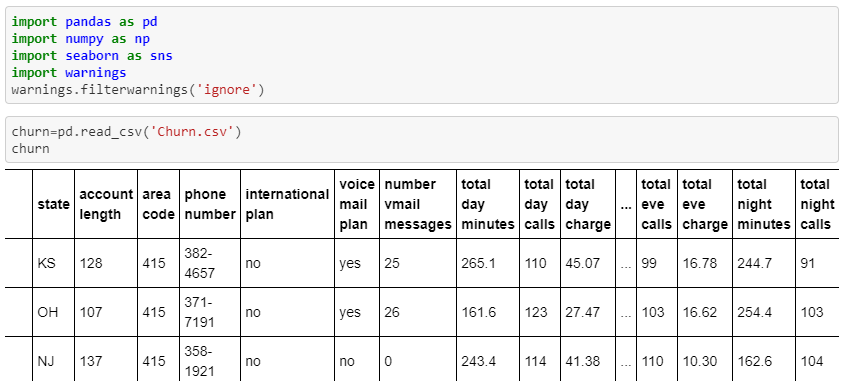
1. **Problem Definition:**

Customers’ churn is a considerable concern in service sectors with high competitive services. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase.

1. **Data Collection:**

The basic layer for predicting future customer churn is data from the past. We look at data from customers that already have churned response and their characteristics before the churn happened. By fitting a statistical model that relates the predictors to the response, we will try to predict the response for existing customers. This method belongs to the supervised learning category.

* Import relevant libraries in Python
* Read the dataset

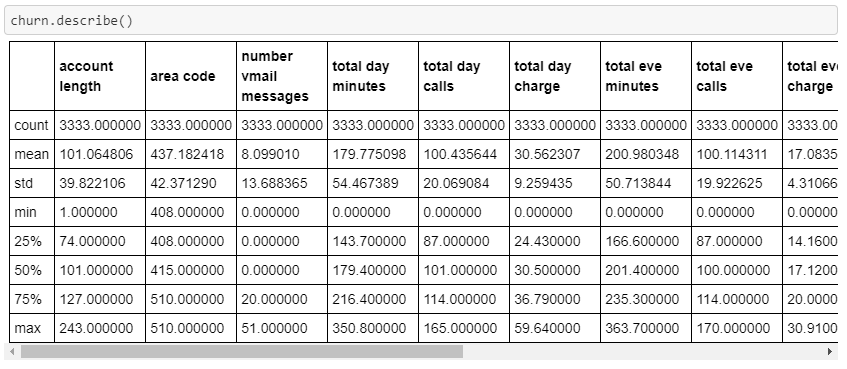


All entries have several features and of course a column stating if the customer has churned or not.

1. **Data Analysis:**

Data analysis is a process of obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analysed to obtain the required output.

* Descriptive analysis of the data set



We find characteristics of each column like min,max,mean,standard deviation etc..

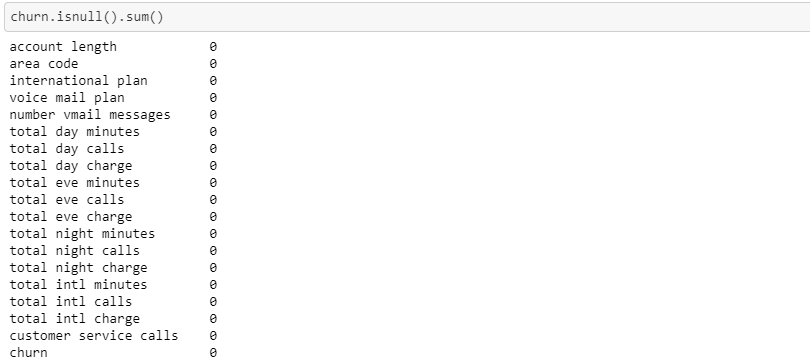
* Dropping irrelevant data

There may be data included that is not needed to improve our results. As it does not influence our predicted outcome, we drop the column using ‘drop ()’ function.



* Repacing / Droping of missing Values

It is important to handle missing data. The values can be identified by the ‘isnull()’. After identifying the null values, it depends on each case whether to fill the missing value with the mean, median or the mode, data drop the entry completely.

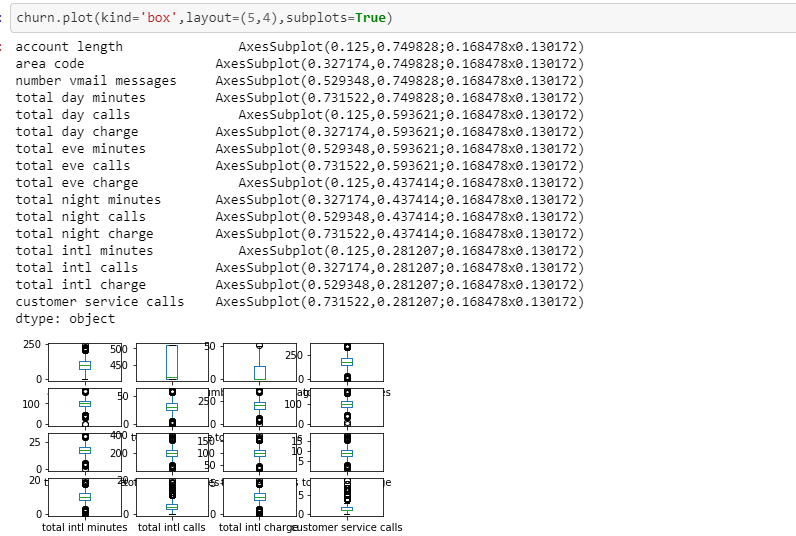


There are no null values present in the dataset.

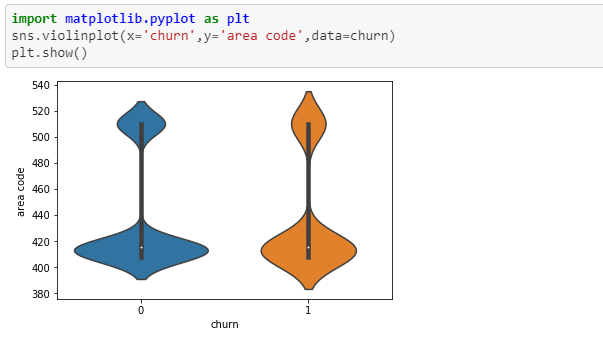
1. **EDA Remarks:**

EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

* Visualising the data using different plot

****

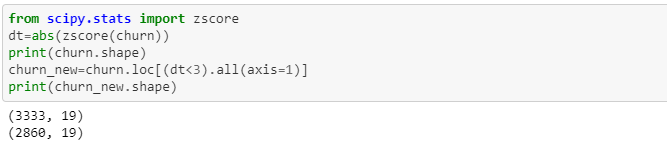
It shows outliers present in the distribution.



The above graph represents that there are more outliers in churn when it is 1

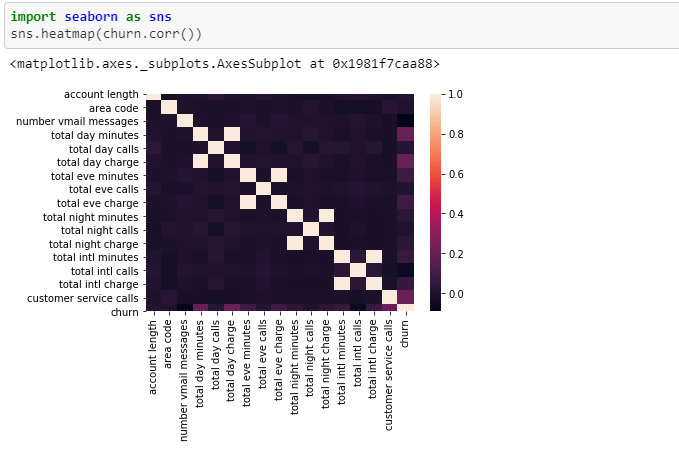
* Removing outliers

Outliers are the unusual values in the dataset. They cause measurement errors and sampling problems while building a predictive model. They can be removed using zscore



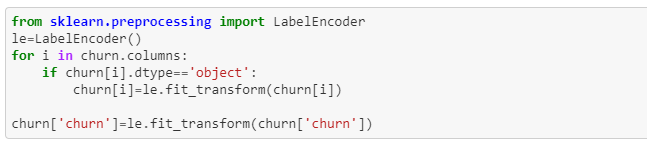
* Correlation

Correlation is a measure of how strongly one variable depends on another var-iable. It is an important tool in building machine learning models. Predictors which are uncorrelated with the objective variable are probably good candidates to trim from the model. Correlation coefficients are a useful tool for exploring relationships within your data



* Converting Categorical data into numerical

As we cannot calculate anything with string values, we have to convert these values into numeric data. This can be done using Label Encoder.



* Splitting the data set

Our model needs to be trained and tested. Therefore it is best to have two different dataset.X is the data with the independent variables, Y is the data with the dependent variable. The test size variable determines in which ratio the data will be split.



1. **Pre-processing Pipeline:**

A data pipeline architecture is a system that captures, organizes, and routes data so that it can be used to gain insights. Raw data contains too many data points that may not be relevant. Data pipeline architecture organizes data events to make reporting, analysis, and using data easier. It is primarily applied to help data improve targeted functionality. This allows for easier tuning and better access to the configuration of the entire model. They are cyclical and iterative as every step is repeated to continuously improve the accuracy of the model and achieve a successful algorithm.

1. **Model Training:**

Since we are dealing with a classification problem, i.e. we can only predict either ‘Churn’ or “not Churn”, therefore we must choose an Classifier algorithms.



The trained model we can predict if a customer churned or not for our test dataset. The results are saved in ‘ypred’. Later the best fit model is decided by calculating parameters like accuracy , confusion matrix , classification report , cross val score and AUC score.

* **Accuracy**:

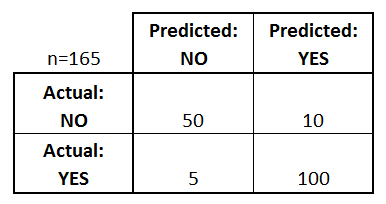
 It is the ratio of number of correct predictions to the total number of input samples.



It works well only if there are equal number of samples belonging to each class.

* **Confusion Matrix:**

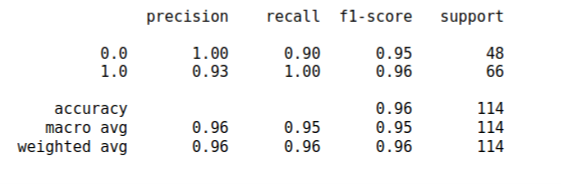
Confusion Matrix gives us a matrix as output which describes the complete performance of the model.



* **Classification Report:**

Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. The report shows the main classification metrics precision, recall and f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives

1. TN / True Negative: when a case was negative and predicted negative
2. TP / True Positive: when a case was positive and predicted positive
3. FN / False Negative: when a case was positive but predicted negative
4. FP / False Positive: when a case was negative but predicted positive



* **Precision**:

It is the number of correct positive results divided by the number of positive results predicted by the classifier.



* **Recall**:

It is the number of correct positive results divided by the number of *all*relevant samples (all samples that should have been identified as positive).



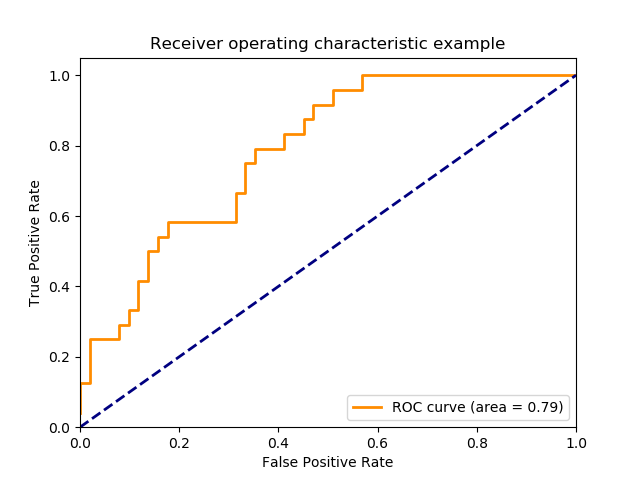
* **F1** **Score**:

F1 Score is the Harmonic Mean between precision and recall. The greater the F1 Score, the better is the performance of our model.



* **AUC** **Score**:

Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. It is the graphical representation of False Positive Rate vs True Positive Rate at different points in [0, 1].



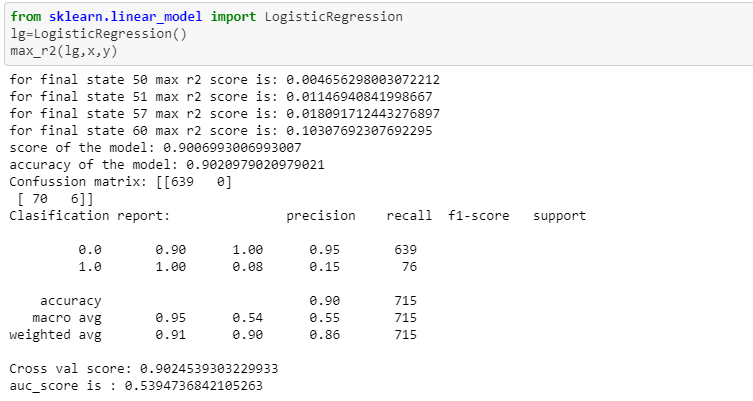
The greater the value, the better is the performance of our model.

* **Logistic Regression:**

Logistic Regression is one of the most used supervised machine learning classification algorithms mainly used when the dependent variable (here churn 1 or churn 0) is categorical. The independent variables in contrary can be categorical or numerical.

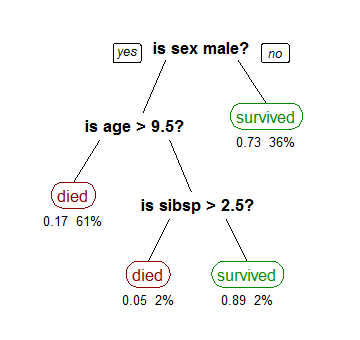


Import Logistic Regressor function using slearn.linear\_model.



* **Decision Tree Classifier:**

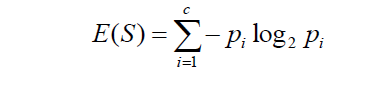
A decision tree is one of most frequently and widely used supervised machine learning algorithms that can perform both regression and classification tasks. For each attribute in the dataset, the [decision tree](https://en.wikipedia.org/wiki/Decision_tree_learning) algorithm forms a node, where the most important attribute is placed at the root node. For evaluation we start at the root node and work our way down the tree by following the corresponding node that meets our condition or "decision". This process continues until a leaf node is reached, which contains the prediction or the outcome of the decision tree.



There are two primary characteristics in the decision tree calculation. One is Information Gain and another is the Gini index.

1) Information Gain is the proportion of Change in entropy. Higher the entrop y more the instructive substance, where the entropy is a proportion of vulnerability of arbitrary variable.

The Mathematical formula for Entropy is as follows -

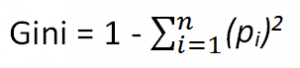


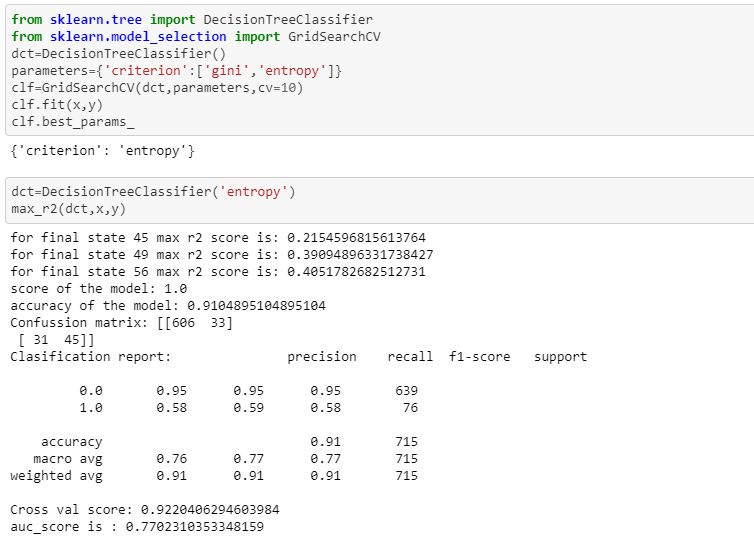
Where ‘Pi’ is simply the frequentist probability of an element/class ‘i’ in our data.

Next we need a metric to measure the reduction of this disorder in our target variable/class given additional information (features/independent variables) about it. This is where Information Gain comes in. Mathematically it can be written as:



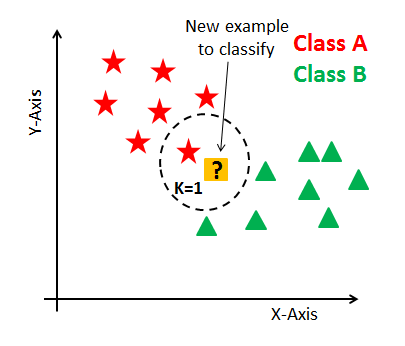
2) Gini Index is a component that measures how frequently an arbitrarily picked component would be mistakenly distinguished. It implies a characteristic with a lower Gini index ought to be liked. The Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one. The Mathematical formula for Gini Index is as follows -

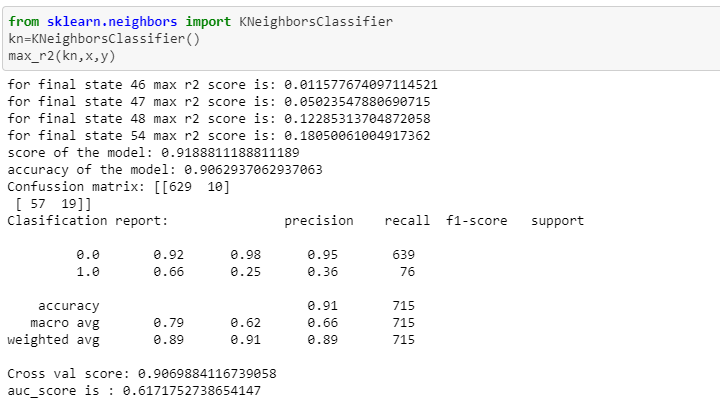




* **KNeighbors Classifier:**

K Nearest Neighbor (KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms.  KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach. K is the number of nearest neighbors. The number of neighbors is the core deciding factor.



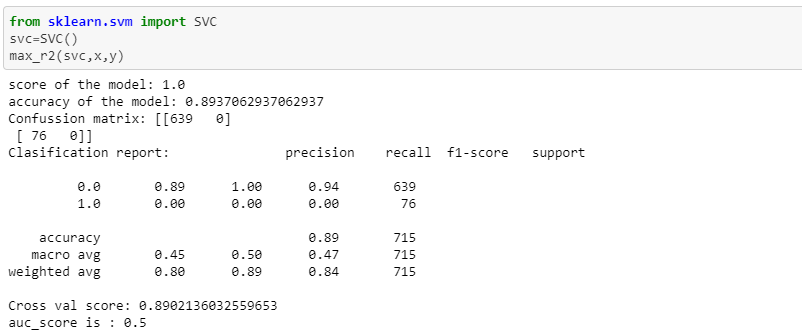


* **SVC:**

‘Support Vector Machine’ (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM 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 coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

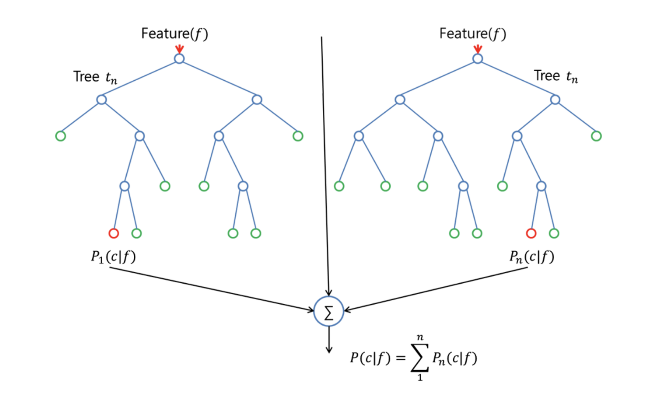
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_1.png)

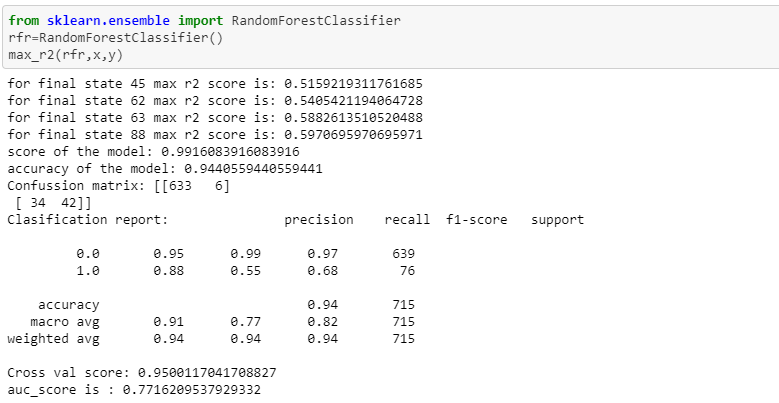
Support Vectors Classifier tries to find the best hyperplane to separate the different classes by maximizing the distance between sample points and the hyperplane.



* **Random Forest Classifier:**

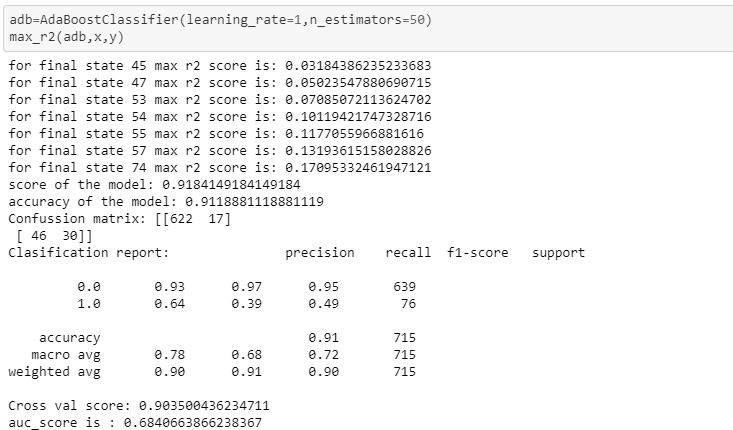
Random Forest is a supervised learning algorithm. It uses the ensemble learning technique (Ensemble learning is using multiple algorithms at a time or a single algorithm multiple times to make a model more powerful) to build several decision trees at random data points. Then their predictions are averaged. Taking the average value of predictions made by several decision trees is usually better than that of a single decision tree.





* **AdaBoost Classifier:**

AdaBoost is one of the first boosting algorithms to be adapted in solving practices. Adaboost helps you combine multiple ‘weak classifiers’ into a single ‘strong classifier’.



1. **Conclusion:**

The importance of this type of research in the telecom market is to help companies make more profit. It has become known that predicting churn is one of the most important sources of income to telecom companies. Hence, we aimed to build a system that predicts the churn of customers. These prediction models need to achieve high AUC values. Also the machine learning models are evaluated on different datasets are studied. Their accuracy and performances are evaluated and compared in order to get better result. From the above different Classifier Techniques we can see Random Forest model is performing really good in regards to all .Finally we will use this to predict our test data.