**Telecom Customer Churn Prediction**

* 1. **Introduction:**

The telecommunication sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators raised the levels of competition. Companies are working hard to survive in this competitive market depending on multiple strategies. Three main strategies have been proposed to generate more revenues such as acquire new customers, upsell the existing customers, and increase the retension period of customer. However, comparing these strategies taking the value of revenue on investment of each into account has shown that the third strategy is the most profitable strategy, proves that retaining an existing customer cost much lower than acquiring a new one, in addition to being considered much easier than the upselling strategy. To apply the third strategy, companies have to decrease the potential of customer's churn, known as the customer movement from one provider to another. Customer's churn is a considerable concern in service sector with high competitive services. On the other hand, predicting the customer who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase. Many research confirmed that machine learning technology is highly efficient to predict this situation.

* 1. **Objectives of Research:**
* The mainobjective of the project is to predict the customer churn
* Highlighting the main variables/factors influencing customer churn.
* To build prediction models various Machine Learning algorithms are used, evaluate the accuracy and performance of prediction models
* finding out the best model for business case & providing executive summary
* In telecommunications industry churn rates are more useful, because most customers have multiple options from which to choose within a geographic location.

**1.3 Problem Statement:**

Customer churn is a big problem for service providers because losing customers results in losing revenue and could indicate service deficiencies. There are many reasons why customers decide to leave services. With data analytics and machine learning, we can identify the important factors of churning, create a retention plan, and predict which customers are likely to churn.

**2. Review of Literature:**

Telecom business enterprises from their operational activities, generated, collected and stored huge amounts of historical data and are now capitalizing these data assets now. Such historical data have buried within them patterns relating to the effectiveness of their various business processes. Machine Learning has been successfully applied in many industries as a practical tool for knowledge discovery. Clustering and classification are extremely important tasks in a wide variety of application domains. There are a large number of techniques available to perform this task and often selection of the most appropriate technique poses a big challenge to the practitioner. Both the techniques have found profound use in the field of Customer Relationship Management (CRM) and Churn Prediction.

**3. Data Collection:**

There are many types of data used to build the churn model. These types are classified as follow:

1. ***Customer data*** it contains all data related to customer’s service and contract information. In addition to all offers, packages, and services subscribed to by the customer.
2. ***Network logs data*** contains the internal sessions related to internet, calls for each transaction in Telecom operator, like the time needed to open a session for the internet and call ending status. It could indicate if the session dropped due to an error in the internal network.

This data has a large size and there is a lot of detailed information about it. We spent a lot of time to understand it and to know its sources and storing format. In addition to these records, the data must be linked to the detailed data stored in relational databases that contain detailed information about the customer.

**4.Methodology:**

For our model, we first looked at several different machine learning algorithms to see which ones to move forward with. Our first step was to split our data into training and testsets using train-test-split, which would allow us to cross-validate our results later. We alsostratified the train-test-split**,** to ensure that the same proportion of our target variable was found in both our training and test sets.

# Stratify our train-test-split so that we have a balanced split  
x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=0, stratify=y)

**Brief Description of Algorithms Used:**

**Decision Tree Classifier:**

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes.

**Random Forest:**

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it.

**K-Nearest Neighbors:**

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use a data and classify new data points based on a similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the most nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.

**Accuracy Measures:**

Decision Tree, Random Forest, KNN algorithms are used in this research work. Experiments are performed using internal cross-validation 10-folds. Accuracy, confusion matrix and ROC (Receiver Operating Curve) measures are used for the classification of this work.

**4.1 Exploratory Data Analysis:**

Solutions In this project working with a Telecom customer churn dataset, we will try to predict which customer is likely to churn. This dataset contains the following features:

***Partner***-Whether the customer has a partner or not (Yes, No)

***Dependents***-Whether the customer has dependents or not (Yes, No)

***Tenure***-Number of months the customer has stayed with the company

***PhoneService***-Whether the customer has a phone service or not (Yes, No)

***MultipleLines***-Whether the customer has multiple lines or not (Yes, No)

***OnlineSecurity***-Whether the customer has online security or not (Yes, No)

***OnlineBackup***-Whether the customer has online backup or not (Yes, No)

***DeviceProtection***-Whether the customer has device protection or not (Yes, No,)

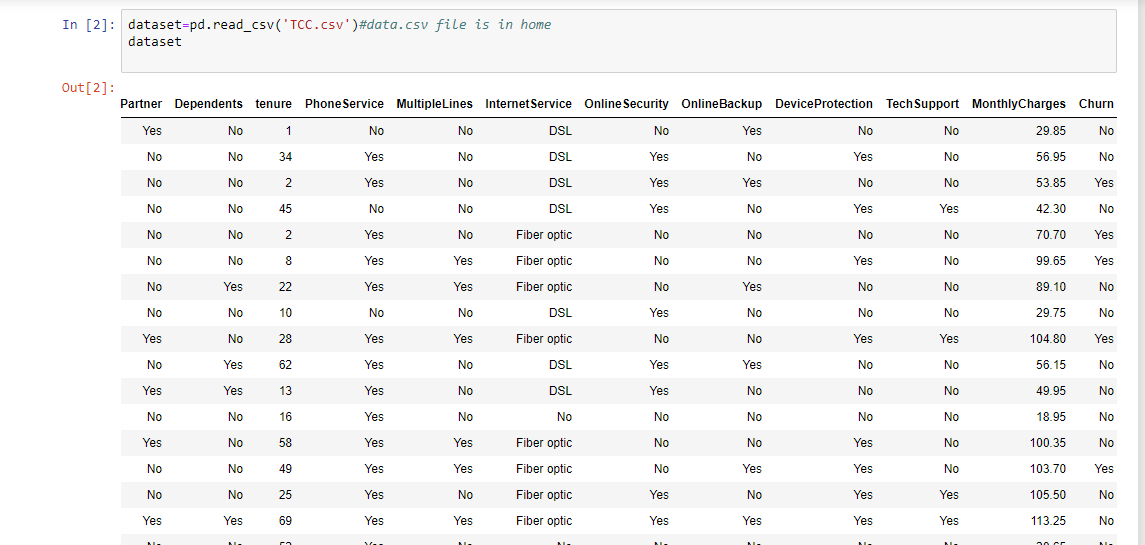
***TechSupport***-Whether the customer has tech support or not (Yes, No)

***MonthlyCharges***-The amount charged to the customer monthly

***Churn***-Whether the customer churned or not (Yes or No)

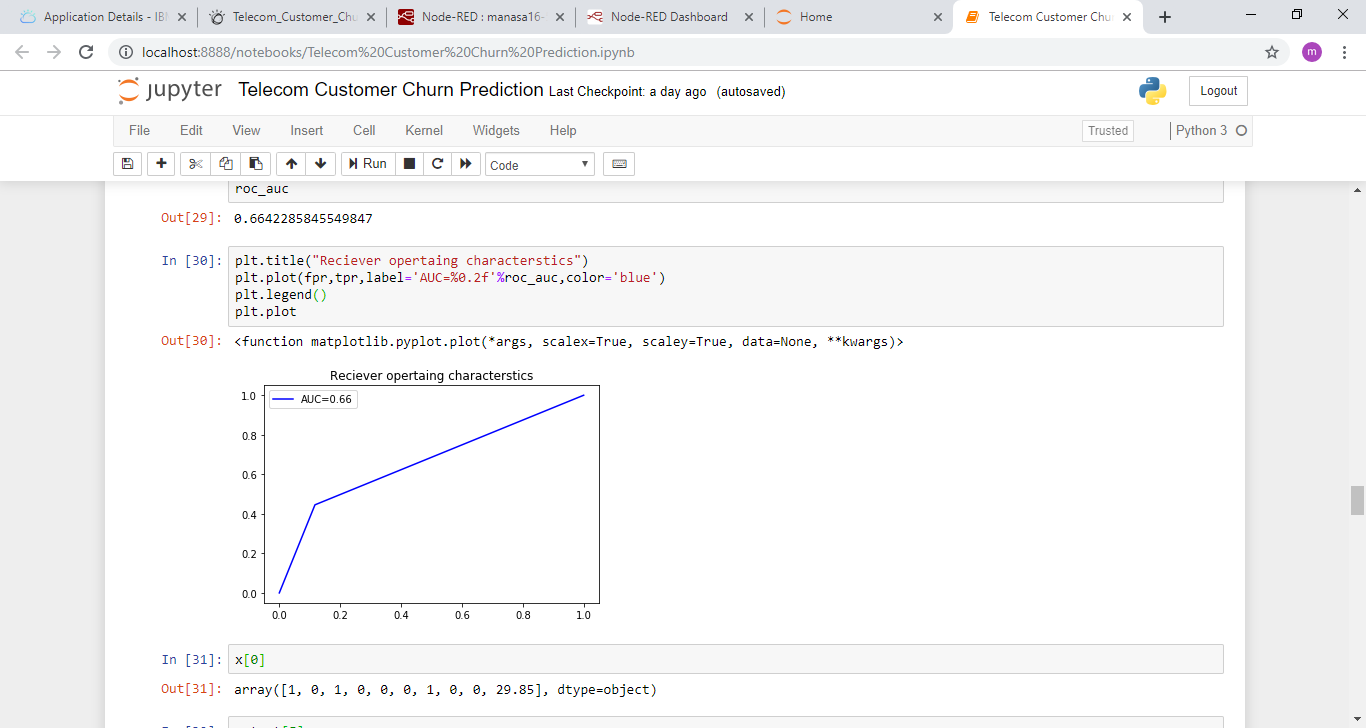
**4.1.1Figures and Tables:**

Dataset for Telecomm customer churn

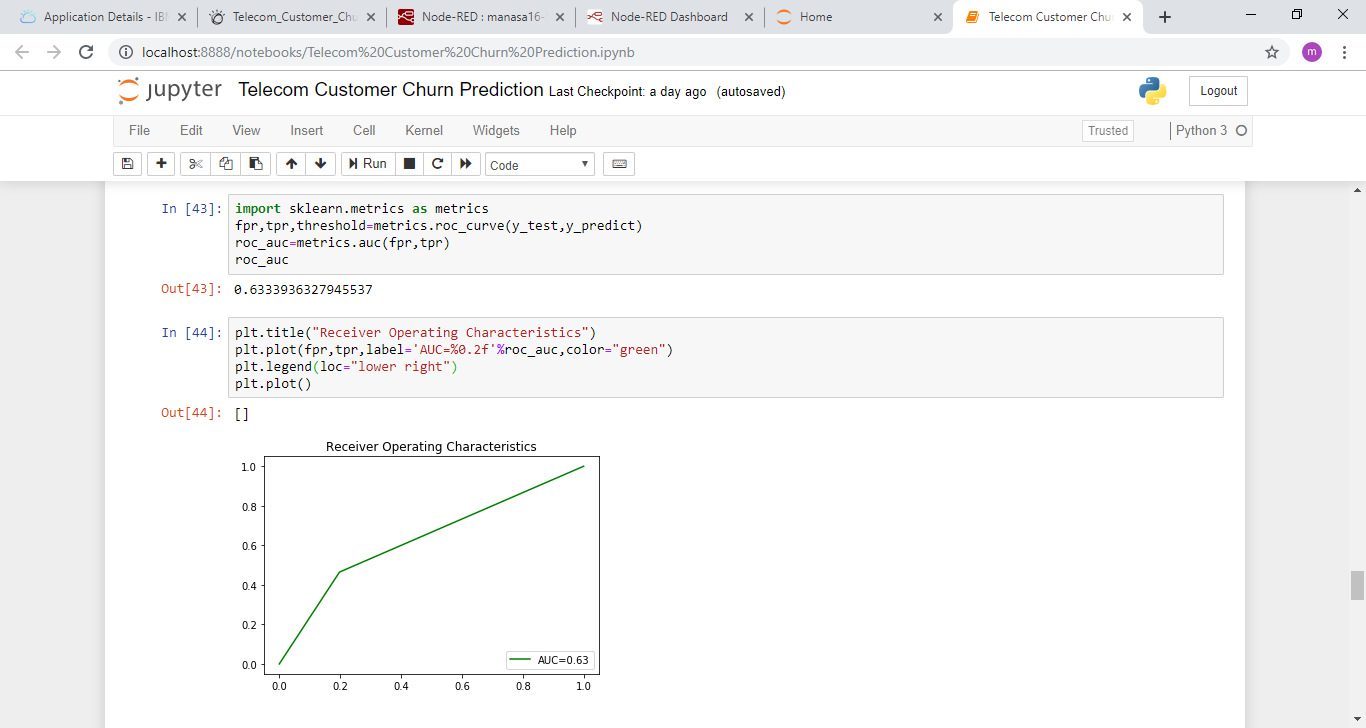


AUC-ROC curves

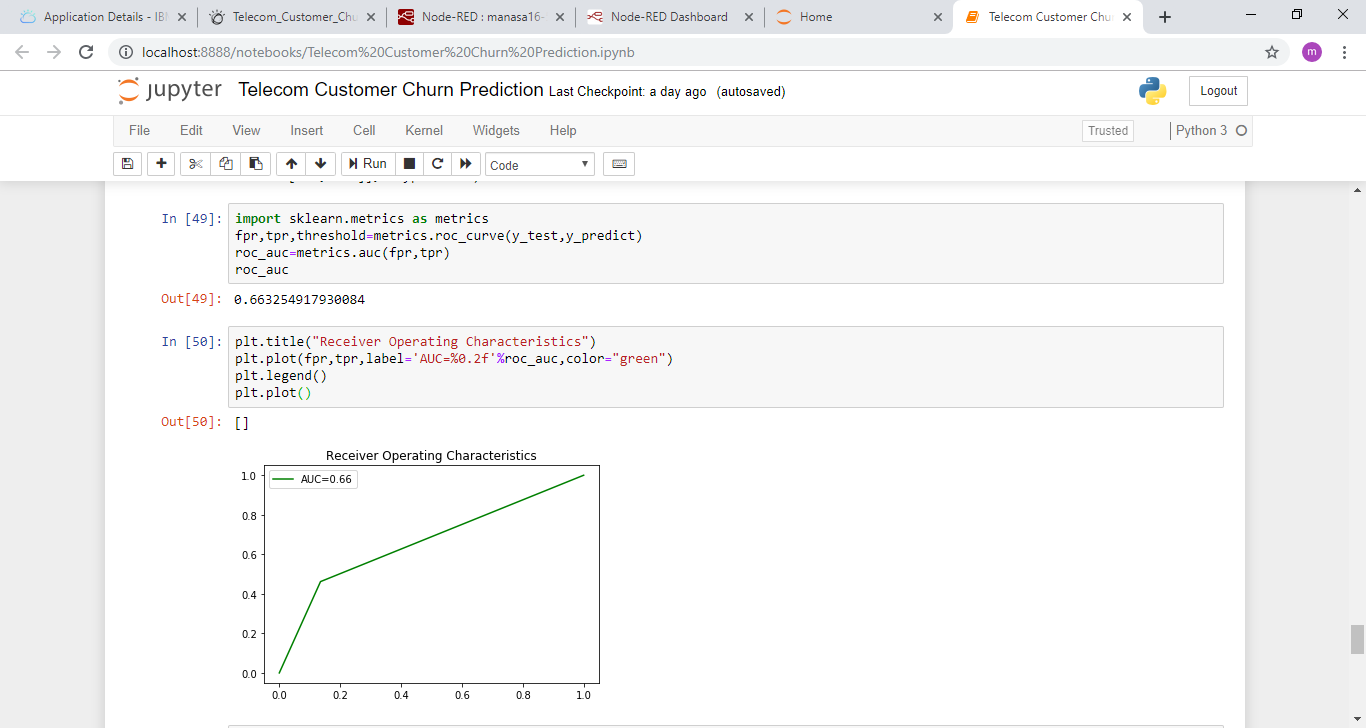
Random forest:



Decision tree:

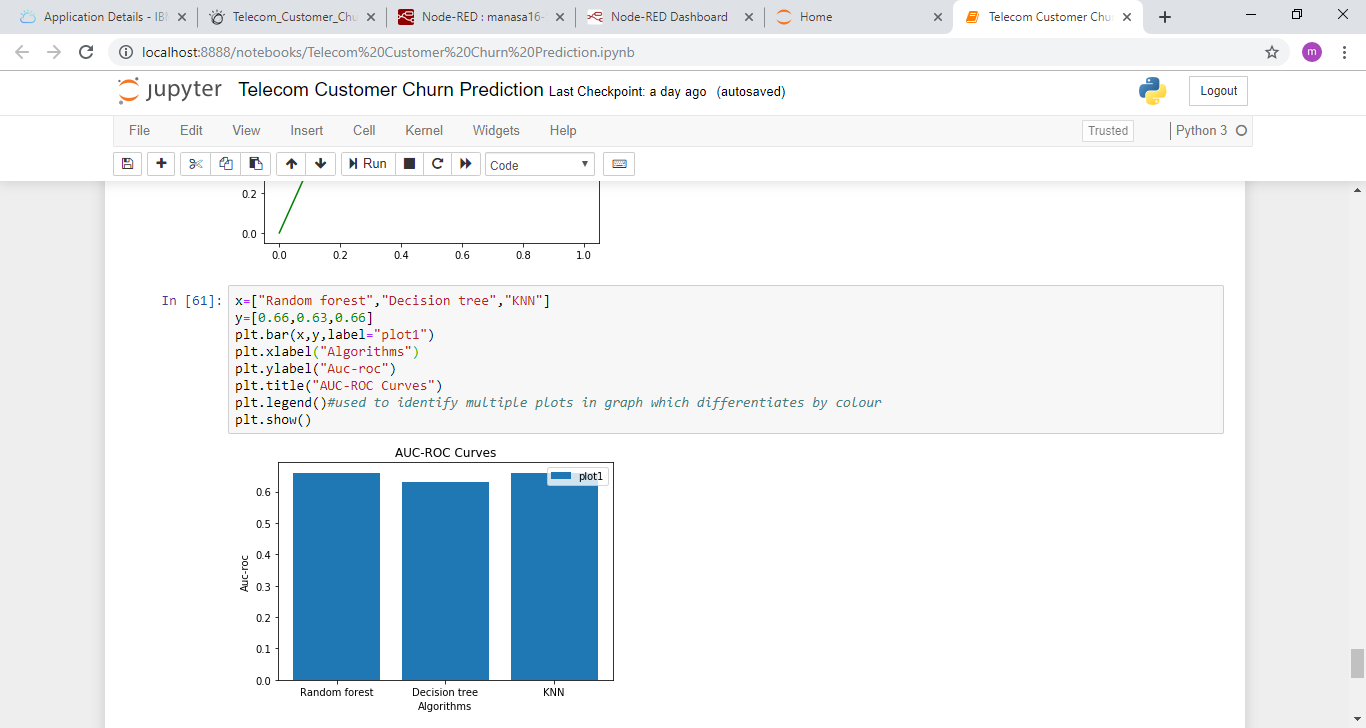


K Nearest Neighbour :

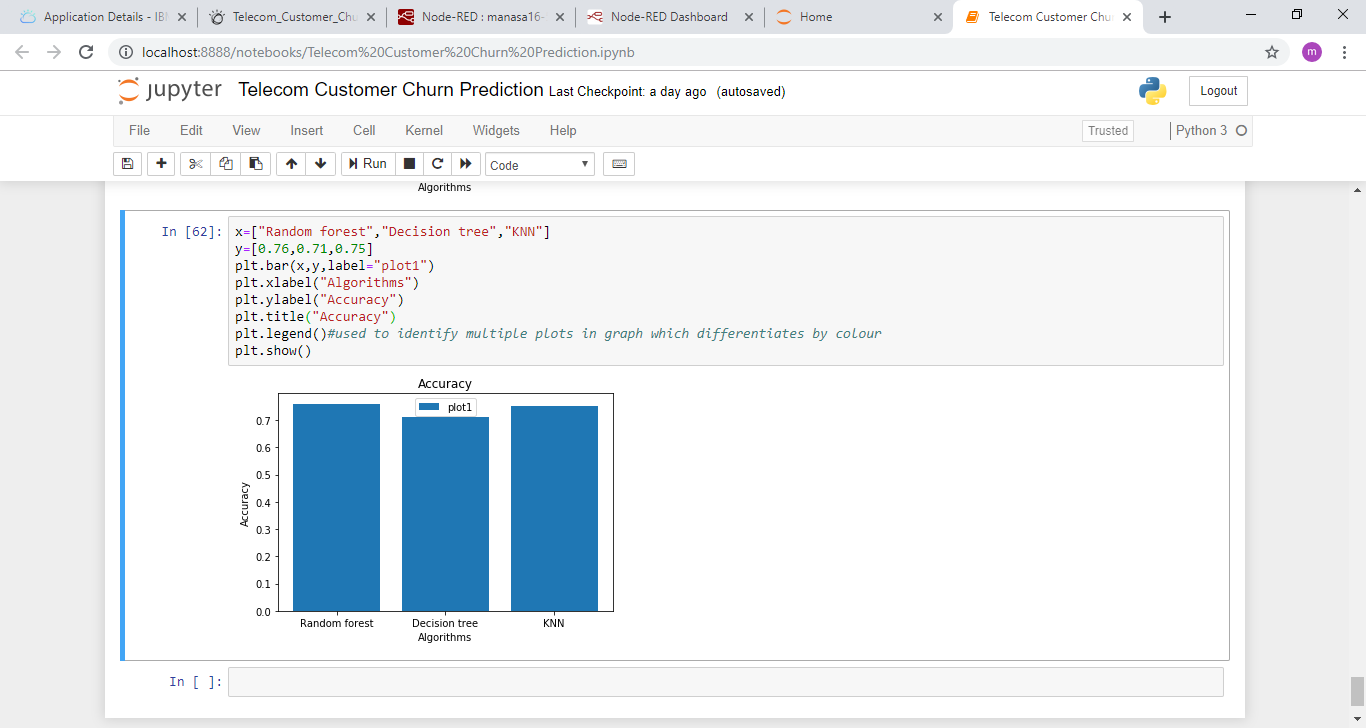


**Bar graphs:**

AUC-ROC:



Accuracy:

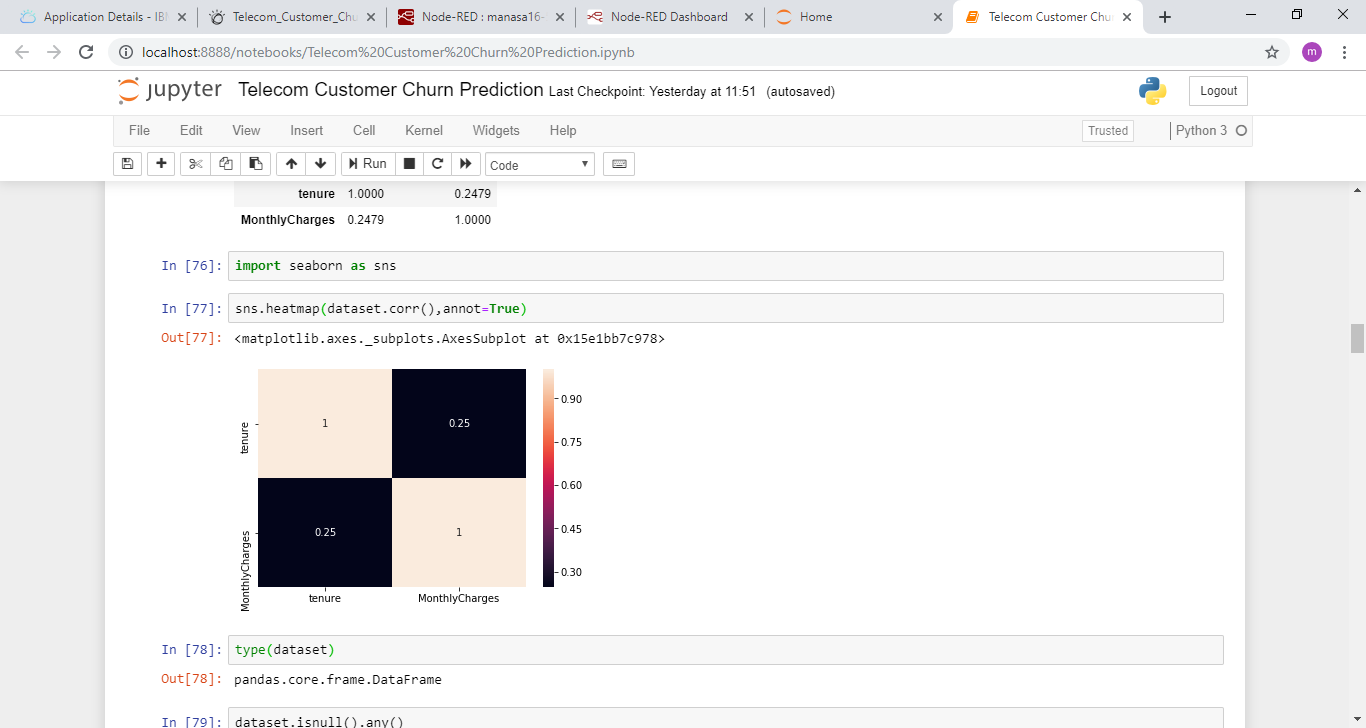


**4.2 Statistical techniques and data visualization:**

By importing matplotlib.pyplot library we have drawn graphs to demonstrate the AUC-ROC curves and by using bar graphs we have visualized the percentage levels of different techniques.

And we have used the co-relation function to demonstrate the impact of every factor on each other.

**Correlation graph:**



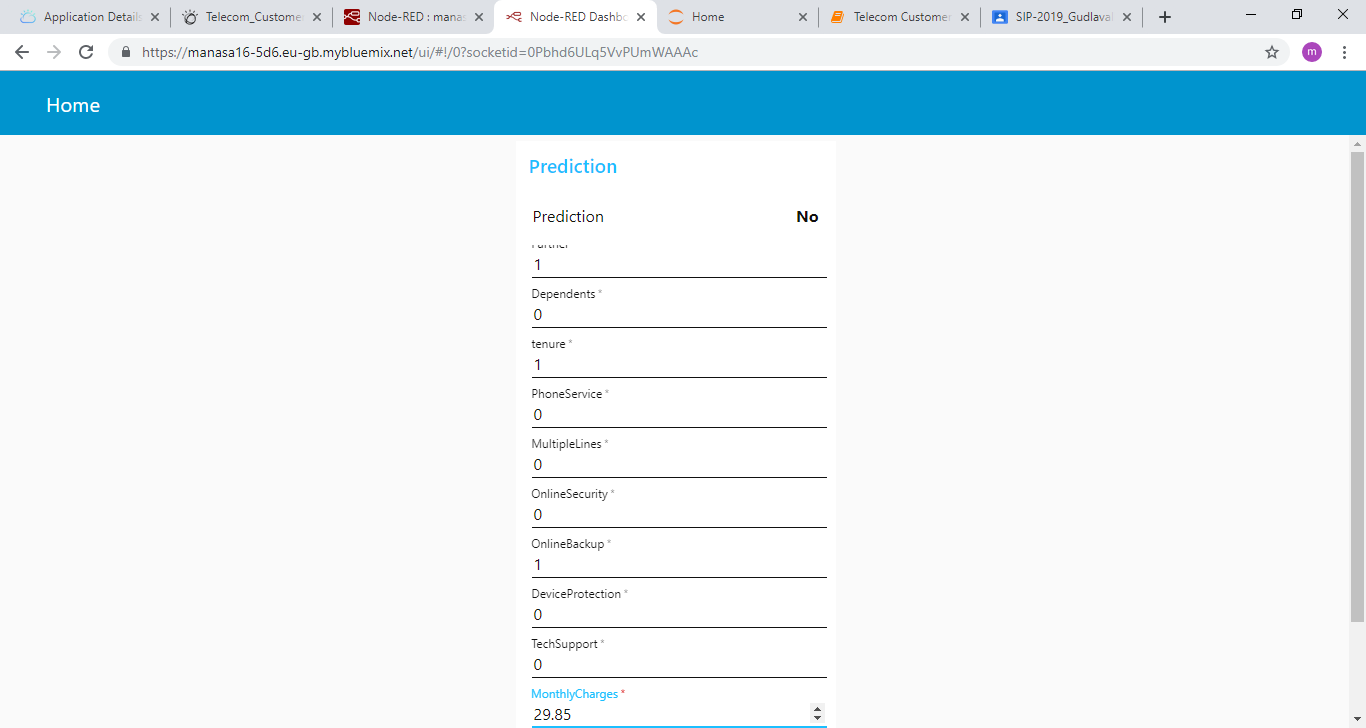
**4.3 Data Modeling Using Supervised ML techniques:**

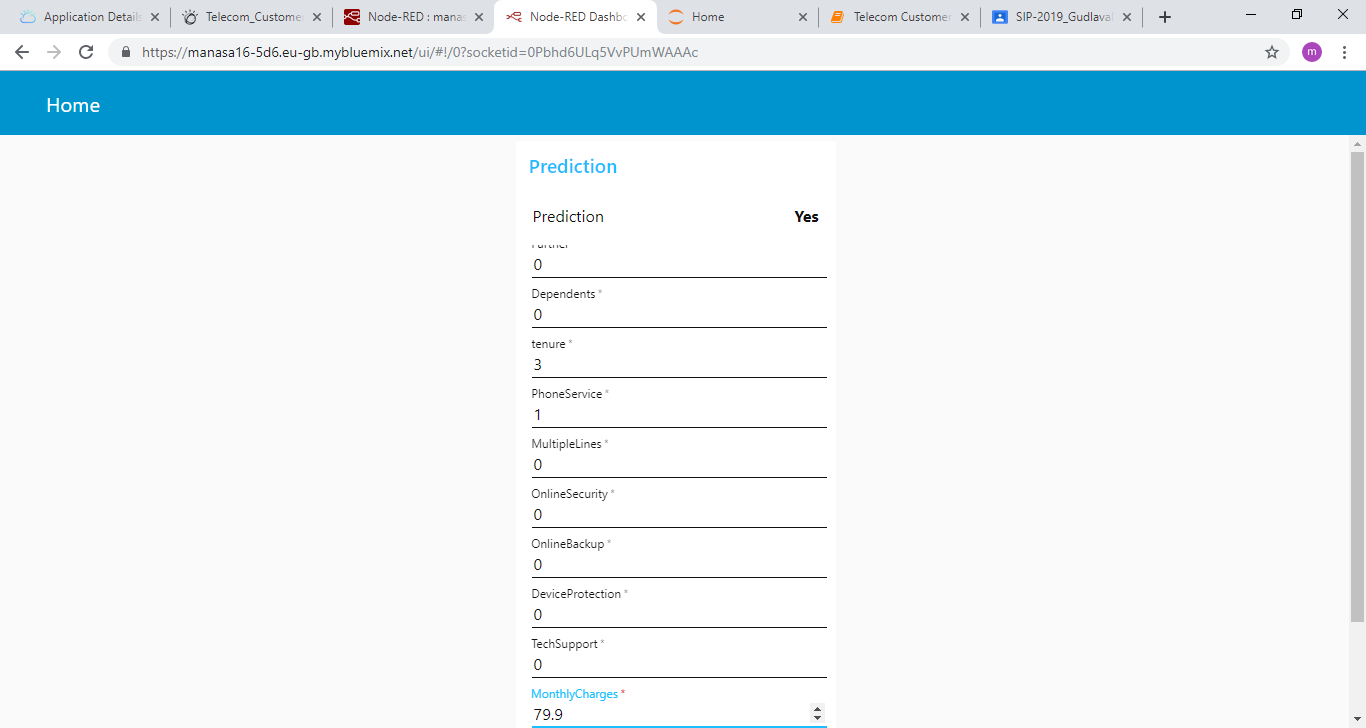
In general we have two types of learning algorithms, supervised and unsupervised learning algorithms. And in detail it consists of different techniques like,

* Support Vector Machines
* decision trees
* k-nearest neighbor algorithm

since our model comes under supervised learning algorithm we applied every technique of the algorithm and based on the accuracy values we obtained we chosen random forest technique. and our model has  dependent variable which is gives the output can be which customer is likely to churn.

**5. Finding and Suggestions:**

****

****

**6. Conclusion**

In this, initially, we introduced the churn prediction problem and the significance of using predictive modeling methods to overcome the problem of customer churn in telecom industry. We surveyed the existing churn prediction methods in detail and summarized them. Unlike other surveys, which primarily focused only on the prediction models and the accuracy of churn prediction, in this survey we presented the characteristics of the existing publicly available churn prediction datasets. Further, we focused on different customer related variables that are used for churn prediction and categorized them. Finally, we surveyed the list of the commonly used metrics proposed in the literature for evaluating the performance of various churn prediction methods