**TELECOM CUSTOMER CHURN ANALYSIS AND PREDICTION**

Sai Teja Saluri - 11547895

Devi Kari -11543834

Bhavani Gudhollu -11554910

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** | **PAGE NO.** |
| **ABSTRACT** | 2 |
| **CHAPTER 1 INTRODUCTION** | 3 |
| 1.1 Introduction | 3 |
| 1.2 Motivation | 3 |
| 1.3. Problem Statement | 4 |
| **CHAPTER 2 METHODOLOGY** | 4 |
| 2.1. Existing Systems | 4 |
| 2.2 Proposed Systems | 4 |
| 2.2.1. Data Collection | 5 |
| 2.2.2. Data Source | 5 |
| 2.2.3. Data Pre-processing | 5 |
| 2.2.4. Balancing of Data | 5 |
| 2.2.5. Prediction of Churn | 6 |
| **CHAPTER 3 WORKING OF SYSTEM** | 6 |
| 3.1. System Architecture | 6 |
| 3.2. Machine Learning | 7 |
| 3.3. Decision Cycle | 9 |
| 3.3.1. Algorithms | 9 |
| 3.3.2. Decision Tree | 9 |
| 3.3.3. Random Forest | 10 |
| 3.3.4. KNN | 11 |
| **CHAPTER 4 EXPERIMENTAL ANALYSIS** | 12 |
| 4.1. System Configuration | 12 |
| 4.1.1. Software Requirements | 12 |
| 4.2.2. Hardware Requirements | 13 |
| 4.2. Code | 13 |
| 4.3. Dataset Details | 19 |
| 4.4. Performance Analysis | 21 |
| **4.6. INPUT AND OUTPUT** | 23 |
| 4.7. Result | 23 |
| **CHAPTER 5 CONCLUSION AND FUTURE WORK** | 24 |

**1.Abstract:**

Many large businesses face a big problem when they lose customers, which is one of their biggest problems. The biggest problems that they have to deal with Because customer turnover has a direct effect on company revenue, especially in the telecommunications industry, companies are putting a lot of effort into creating technologies that can accurately predict how likely it is that a customer will leave. To keep customers from leaving, it is important to create a churn prediction model that helps telecom companies find the customers who are most likely to leave. This is one of the most important result of our research. To reduce customer turnover, it is important to figure out what factors contribute to it. The most important thing that came out of our research was making a churn prediction model that helps telecom companies figure out which customers are most likely to leave. The reason for we did this study is to get to model. During this research, a model that used machine learning techniques applied to big data platforms was used to come up with a unique strategy for feature engineering and selection. This strategy was then put into action. With the help of the dataset, which had all of the customer data, the system was built, tested, and judged. We used both decision trees and random forests to build our algorithms so that we could predict churn. The churn forecasting model used these techniques to put the data into groups.

Quantified algorithms used: Decision tree and Random forest.

1.1. **Introduction:**

People who stop using a product or service after a certain amount of time are called "churners." Churn is the word for a customer of a telecommunications company who has decided to stop using that company's services. The churn model was able to correctly pick out the employee who is most likely going to leave the company soon. The churn model is an example of a common way that data mining techniques are used. These techniques are used in many different fields. Different mobile phone companies in different parts of the world are almost done with their own churn models. Also, the results of customer turnover can be used effectively for a number of other goals, which can help keep clients you already have. In reality, the Churn Management technique is just the first step in building a model. In general, the project needs a churn model that is implemented in the most effective way, rather than taking a single approach that has the best lift. Because of this, we've made an automated application to use as a backup for the near future. It's not unusual for a customer of one company to also use the services of one or more telecommunication companies in this day and age. Some of us might use different carriers because we live far away from them, and others might use multiple carriers because they offer different plans. When doing the study with the help of machine learning, it's often helpful to use what you've learned from past customer experiences. Some people have changed the company that gives the services once in a while. Due to multiple responsibilities that has come with the job, the calling price could be either go up or to down. Multiple scenarios can be shown, depending on the data that is available.

**1.2. Motivation:**

Utilizing the predictive analytics algorithm for data mining will allow this research to accomplish its primary goal, which is to enhance the methods now utilized in the telecommunications industry for estimating the rates at which customers leave their service.

Consideration was given to various aspects of the following supplemental objectives:

i.) Dividing a part of audiences for promoting and advertising to certain groups of humans.

ii) The mining of important patterns in the accumulated data has a significant impact on the growth and income of the telecom sector. This is the fact that the sector receives a huge amount of data.

**1.3. Problem Statement:**

The Problems That Are Causing Customers to Leave Their Telecom Providers and How to Fix Them Businesses in the telecommunications industry are worried about reducing the amount of customer turnover because it is expensive to attract new customers. In most cases, customers will freely terminate their service with their respective telecommunications providers.

**METHODOLOGY:**

**2.1. Existing system:**

In order to find its churning customer base, the telecommunications industry utilized a wide variety of different tactics. The majority of these approaches are based on the use of data mining and machine learning. The majority of the connected papers looked at different churn prediction approaches, while others concentrated on data mining techniques to extract information from data sets.

In order to assist a large Chinese telecom firm with 5.23 million subscribers in lowering their rate of customer turnover, Heetal developed an algorithm that is based on neural networks. We computed the accuracy rate in order to conduct an overall assessment of the reliability of the forecasts.

**2.2. Proposed System:**

Following the collection of data and the determination of the primary properties, the system will be ready to begin its operation. After that, the necessary data will have the structure that is appropriate for it applied to it. After that, the data is split up into sets that will be used for training and for testing. The process of applying the algorithms to the training set results in the model being "trained." Inputting test data into the system in order to validate the functionality. The following elements are necessary for the successful operation of this system.

1. Data Collection
2. Data Source
3. Data Pre-processing
4. Balancing of Data

**2.2.1. Data Collection:**

The vast majority of responses came from patrons who were asked to participate in a survey for research purposes. The questionnaire for the survey was built with the help of an add-on for Google Drive. Many individuals participated in the survey's training component, whereas some individuals participated in the survey's testing component.

**2.2.2. Data Source:**

Because of the data that is available to us, we are able to learn about the lives of our customers at different ages. Without the information that is provided by this dataset, the churn study would not be possible. Some of the details that are included in this dataset include Partner, dependents, Tenure, multiple lines, and others. This data set is useful for churn analysis, and it has been thoughtfully partitioned into training and testing sets for your convenience. The following is a list of some of the columns and values that can be discovered in the data collection that was just presented.

**2.2.3. Data Pre-processing:**

There was a potential that the training dataset and the testing dataset that were used in this inquiry had information that was either missing, duplicated, or inconsistent. Dealing with missing data and removing duplicate values are the two primary focuses of the data pre-processing stage. RapidMiner is currently being utilized in order to get the data ready for mining and analysis. Before the data types can be employed in cluster analysis, Pearson's chi-square, or the creation of predictive models, they must first be translated to numbers.

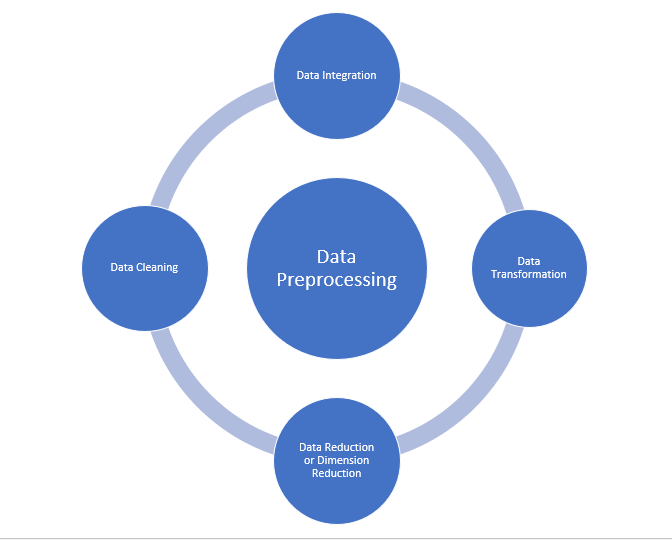


Figure: Data pre-processing.

**2.2.4. Balancing of Data:**

There are two approaches that can be taken to better balance out unbalanced datasets. They are both not sampling enough and trying to sample too much.

a) Under sampling: When using under sampling, the size of the huge class is reduced so that the dataset can be more evenly distributed. This strategy will be looked into further whenever there is sufficient information available.

b) Oversampling: This technique involves increasing the number of sparse samples in a dataset in order to make the distribution of the samples more even. This approach is investigated whenever there is an insufficient amount of information.

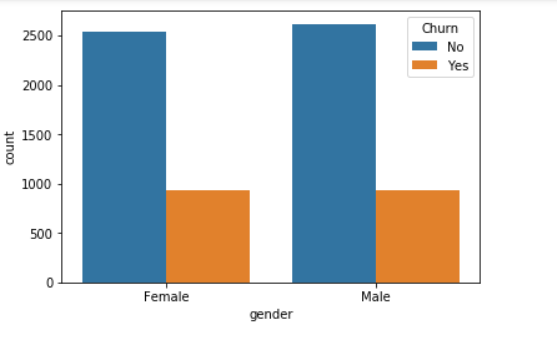


Figure: Balancing of the data.

### 2.2.5. Prediction of Churn

Many different machine learning algorithms, such as Decision Tree and Random Forest, are utilized throughout the categorization process. A comparison of different algorithms is carried out, and the one that yields the most accurate results is selected to be used for the prediction of customer churn.

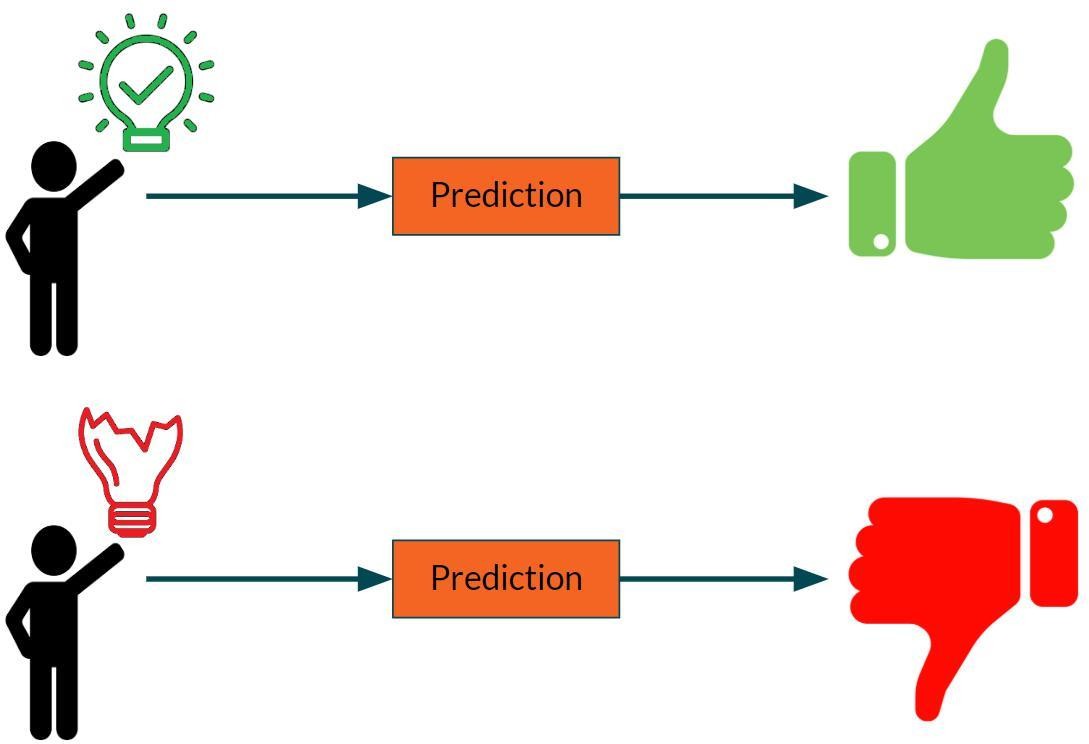


Figure: Prediction of Customer Churn

**3. WORKING ON A SYSTEM**

**3.1. SYSTEM ARCHITECTURE**

**The architecture of the system describes its overall structure and operation.**

**Work Flow Diagram:**

The process of acquiring information about customers is known as dataset collection. Attribute selection refers to the process of identifying beneficial characteristics for the prediction of churn analysis. Following the discovery of available data resources, they are selected, cleaned, and translated into the proper format. As previously stated, several classification algorithms will be applied on pre-processed data to estimate the accuracy of churn predication. The accuracy measure compares the precision of several classifiers.

### 

### Fig. System Architecture

### 3.2. MACHINE LEARNING

### In the field of machine learning, the term "classification" refers to a predictive modelling task known as "classification," in which a class label is predicted for a given sample of input data.

1. **Supervised Learning:**

A kind of machine learning known as supervised learning uses well-labeled training data to train machines, which then predict output based on that data. Data that has been labeled means that some input data has already been annotated with the appropriate output.

During supervised learning, the training data that is given to the computers serves as a supervisor, instructing them on how to correctly predict the output. It applies the same theory that a student would learn while being guided by a teacher.

The act of giving appropriate input and output data to a machine learning model is known as supervised learning. A supervised learning algorithm's objective is to identify a mapping function that links the input variable (x) to the desired outcome (y).

1. **Unsupervised Learning**

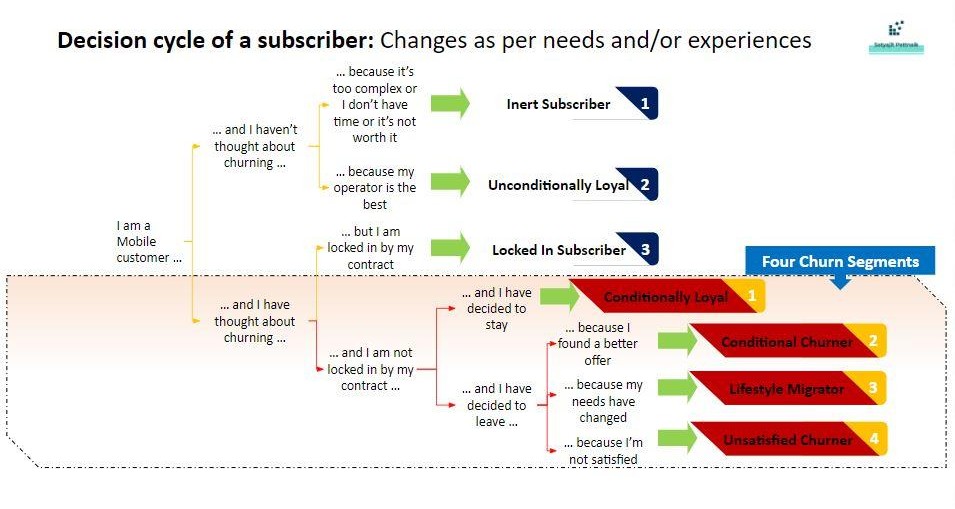
In contrast to supervised learning, unsupervised learning merely provides the opportunity for input data, which precludes its usage for immediate applications such as regression or classification. Discovering the structure of a dataset, classifying its components according to the similarities they share, and presenting the findings in a condensed form are the objectives of unsupervised learning. It is to one's great advantage to gain insights from data through the process of unsupervised learning.

* Unsupervised learning is closer to actual artificial intelligence because it replicates the way people learn to think, which is by doing. This is because unsupervised learning is based on reinforcement learning.
* The capacity to use data that has not been labelled or categorized increases the value of unsupervised learning.
* Unsupervised learning is required in the real world since we do not always have data inputs that match the outputs.

**3. Reinforcement Learning**

One of the numerous subfields that fall under the umbrella of machine learning is called "reinforcement learning." To get the most out of a situation, one must behave in a manner that is consistent with getting the most out of it. It is used by computers to figure out the best way to proceed in a given situation. Reinforcement learning is an alternative to supervised learning, which requires the model to be given incorrect solutions in order to instruct it on how to function. In supervised learning, the correct answer is presented alongside the training data. In the process of reinforcement learning, there is no one right answer; therefore, the agent is responsible for determining what actions to take on its own. It won't be able to learn anything without a dataset to train on, so everything will be up to it to figure out on its own

# **Decision cycle:**



**3.3. Algorithms:**

**3.3.3. DECISION TREE ALGORITHM**

The Decision Tree algorithm is a form of supervised learning that may be utilized for a variety of different types of tasks, including classification and regression. However, the principal application for it is in the solving of problems associated with categorisation. A classifier that takes the form of a tree, with internal nodes that store attributes of the dataset, branches that store decision rules, and leaf nodes that store the final results. The two points at which a Decision Tree comes to an end are known as the Leaf Node and the Decision Node.

Decision nodes, which serve as the centres of activity within a decision-making network, frequently give rise to a large number of offshoots. The results of these decisions are displayed at leaf nodes, which do not have any child nodes of their own. The dataset that is being provided is utilized in order to carry out any necessary evaluations or investigations. It is a graphical representation of all the possible solutions or decisions to a certain issue or situation. The term "decision tree" comes from the fact that its fundamental structure, which consists of a "root node" from which other branches emerge, is modelled after the structure of trees. Utilizing the CART method, a tree should be created. Data classification is accomplished through the application of the CART algorithm. A Decision Tree is a straightforward method that involves asking a series of yes/no questions in order to select an action to do.

The Decision Tree Algorithm is one example of a machine learning approach that is supervised in nature. It can be applied to solve problems involving regression as well as classification, among other things.

The objective of this tactic is to construct a model that is able to make reliable projections regarding the value that will be attained by a dependent variable in the future. The decision tree is a type of diagram that depicts the scenario as a tree. The leaf node of the tree is a class label, and the inner nodes of the tree display the attributes, which allows the problem to be solved.

When constructing a model for machine learning, one of the most important considerations is picking the suitable method to apply to the particular dataset and problem. The Decision Tree can be helpful for a variety of reasons, including the following: • Decision trees are easy to understand since they are based on the way that people actually make decisions. Because of its basic tree-like structure, the decision tree's operation is very easy to understand.

The basic objective of a Decision Tree is to determine the qualities that are shared by the root node of each level. This procedure is referred to by its technical name, attribute selection. There are two basic approaches to picking traits, which are as follows:

1. Information Gain
2. GINI Index
3. IDichotomiser 3 (ID3)
4. C4.5
5. Classification and Regression Tree.

**Working:**

1. Step 1: S recommends starting the tree from the root node, which represents the entire dataset.
2. Step 2: Determine which data point is most important by employing the Attribute Selection Measure (ASM).
3. Step 3: The S should be broken down into smaller sets, each of which could contain values for the most important characteristics.
4. Step 4: The best attribute should be used to create the node in the Decision Tree.
5. Step 5: Recursively generate new decision trees from the step-3 dataset subsets. Keep going until you reach a point where you can no longer classify the nodes further, at which point you will have reached a leaf node.

### 3.3.4. RANDOM FOREST ALGORITHM

The Random Forest Algorithm is a way of learning through supervision called supervised learning. Integration of bagging with machine learning classifiers, with the goal of improving Decision Tree performance. Every tree relies on its own random vector, which is gathered from a different pool of data. The distance between individual trees remains consistent throughout the entire forest. Instead of subdividing the predictors into groups according to the variables, Random Forests chooses the subset of predictors from each node that is the most informative. The worst-case time for learning using Random Forests is O(M(dnlogn), provided that there are M growing trees, n instances, and d dimensions to the data.

It is adaptable enough to be used for regression analysis as well as classification work. This algorithm is the most flexible and user-friendly one that is currently available. Trees are the building blocks of a forest. Many people are under the impression that the number of trees in a forest directly correlates to its level of resistance. Random Forests construct numerous Decision Trees using random data samples, gather forecasts from each tree, and then poll the population to decide which tree's predictions are more correct. This process is called "collecting forecasts from each tree." Additionally, it is an encouraging sign that the addition is significant in some way.

Random Forests have a wide variety of applications, some of which include the selection of features, the recommendation of features, and the classification of images. It is possible to utilize it to detect dishonest behavior, rank loan applicants according to their loyalty, and even diagnose illness. In point of fact, it is used as the basis for yet another well-known data-mining strategy, which is known as the Boruta approach.

Random Forest is one of the many algorithms that are used in supervised machine learning, but it is one of the most well-known. Both classification and regression are possible applications for it when employed as a tool for machine learning. The idea of ensemble learning is used by it in order to improve a model's performance and come up with a solution to a difficult problem.

As its name suggests, "Random Forest is a classifier that employs many decision trees on various subsets of the given dataset and takes the average to enhance the accuracy of predictions for that dataset." "Random Forest" is an acronym for "many decision trees on various subsets of the given dataset and taking the average." Instead of relying on just one decision tree to reach a result, the random forest compiles all of the predictions made by the individual trees and utilizes the one with the highest number of votes to draw the ultimate conclusion.

Both the accuracy of the model and the likelihood that it will be overfitted to the data will improve as the number of trees in the forest grows.

**Algorithm Steps:**

It operates in four steps, which are as follows:

1. Choose from a pre-existing collection of data.
2. Construct a decision tree for each of the samples, and then make educated guesses about the results.
3. Cast your vote for each of the potential outcomes.
4. If you are looking for a forecast that is supported by the majority of people, you should choose the option that garners the most votes.

**3.3.5. KNN**

Introduction

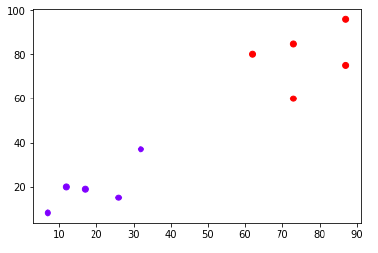
The supervised machine learning (ML) method K-nearest neighbors (KNN) may be applied to both classification and regression prediction problems. It is mostly used in industry, though, to solve categorization and prediction problems. The two factors stated below give a clear definition of KNN.

KNN is a lazy learning algorithm since it utilizes the entire data to train and classification instead of having a distinct training phase. method for non-parametric learning KNN is a non-parametric learning method as well since it doesn't make any assumptions data.

Making use of the KNN Algorithm

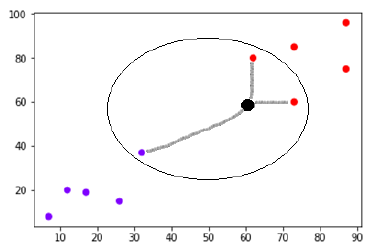
The K-nearest neighbors (KNN) approach predicts new data items and values based on "feature similarity," which means that the new data point will now be assigned a value based on how closely it resembles the points in the training set.

The KNN Algorithm in Action

The K-nearest neighbors (KNN) technique predicts the values of new data points based on 'feature similarity,' which means that a value will be assigned to the new data point based on how closely it resembles the points in the training set. The following stages will assist us in understanding how it works. −

The K concept

Now we must categorize a fresh data point with a black dot (at location 60,60) as blue or red. We'll assume K = 3, which means it'll find the three closest data points. It is seen in the diagram below.



**4. EXPERIMENTAL ANALYSIS:**

# **System configuration and Software Requirements:**

* Operating system: Windows- 64 GB RAM
* Programming tool: Jupyter notebook.
* Programming language: Python.
* Rest API: flask.
* Cloud: Google cloud platform.
* Client-side: HTML, flask
* Server-side: python

|  |
| --- |
| **4.1.2. Hardware requirements:** |

* Processor : Any processor
* RAM: Minimum 4GB
* Hard Disk: Minimum 100 GB
  1. **CODE:**

**Sample python code :**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn import metrics

from flask import Flask, request, render\_template

import pickle

app = Flask("\_\_name\_\_")

df\_1=pd.read\_csv("first\_telc.csv")

q = ""

@app.route("/")

def loadPage():

return render\_template('home.html', query="")

@app.route("/", methods=['POST'])

def predict():

'''

SeniorCitizen

MonthlyCharges

TotalCharges

gender

Partner

Dependents

PhoneService

MultipleLines

InternetService

OnlineSecurity

OnlineBackup

DeviceProtection

TechSupport

StreamingTV

StreamingMovies

Contract

PaperlessBilling

PaymentMethod

tenure

'''

inputQuery1 = request.form['query1']

inputQuery2 = request.form['query2']

inputQuery3 = request.form['query3']

inputQuery4 = request.form['query4']

inputQuery5 = request.form['query5']

inputQuery6 = request.form['query6']

inputQuery7 = request.form['query7']

inputQuery8 = request.form['query8']

inputQuery9 = request.form['query9']

inputQuery10 = request.form['query10']

inputQuery11 = request.form['query11']

inputQuery12 = request.form['query12']

inputQuery13 = request.form['query13']

inputQuery14 = request.form['query14']

inputQuery15 = request.form['query15']

inputQuery16 = request.form['query16']

inputQuery17 = request.form['query17']

inputQuery18 = request.form['query18']

inputQuery19 = request.form['query19']

model = pickle.load(open("model.sav", "rb"))

data = [[inputQuery1, inputQuery2, inputQuery3, inputQuery4, inputQuery5, inputQuery6, inputQuery7,

inputQuery8, inputQuery9, inputQuery10, inputQuery11, inputQuery12, inputQuery13, inputQuery14,

inputQuery15, inputQuery16, inputQuery17, inputQuery18, inputQuery19]]

new\_df = pd.DataFrame(data, columns = ['SeniorCitizen', 'MonthlyCharges', 'TotalCharges', 'gender',

'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService',

'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',

'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling',

'PaymentMethod', 'tenure'])

df\_2 = pd.concat([df\_1, new\_df], ignore\_index = True)

# Group the tenure in bins of 12 months

labels = ["{0} - {1}".format(i, i + 11) for i in range(1, 72, 12)]

df\_2['tenure\_group'] = pd.cut(df\_2.tenure.astype(int), range(1, 80, 12), right=False, labels=labels)

#drop column customerID and tenure

df\_2.drop(columns= ['tenure'], axis=1, inplace=True)

new\_df\_\_dummies = pd.get\_dummies(df\_2[['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'PhoneService',

'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup',

'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies',

'Contract', 'PaperlessBilling', 'PaymentMethod','tenure\_group']])

#final\_df=pd.concat([new\_df\_\_dummies, new\_dummy], axis=1)

single = model.predict(new\_df\_\_dummies.tail(1))

probablity = model.predict\_proba(new\_df\_\_dummies.tail(1))[:,1]

if single==1:

o1 = "This customer is likely to be churned!!"

o2 = "Confidence: {}".format(probablity\*100)

else:

o1 = "This customer is likely to continue!!"

o2 = "Confidence: {}".format(probablity\*100)

return render\_template('home.html', output1=o1, output2=o2,

query1 = request.form['query1'],

query2 = request.form['query2'],

query3 = request.form['query3'],

query4 = request.form['query4'],

query5 = request.form['query5'],

query6 = request.form['query6'],

query7 = request.form['query7'],

query8 = request.form['query8'],

query9 = request.form['query9'],

query10 = request.form['query10'],

query11 = request.form['query11'],

query12 = request.form['query12'],

query13 = request.form['query13'],

query14 = request.form['query14'],

query15 = request.form['query15'],

query16 = request.form['query16'],

query17 = request.form['query17'],

query18 = request.form['query18'],

query19 = request.form['query19'])

app.run()

**Homepage code :**

<html>

<head>

<link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css">

<script src="https://ajax.googleapis.com/ajax/libs/jquery/3.4.1/jquery.min.js"></script>

<script src="https://cdnjs.cloudflare.com/ajax/libs/popper.js/1.14.7/umd/popper.min.js"></script>

<script src="https://maxcdn.bootstrapcdn.com/bootstrap/4.3.1/js/bootstrap.min.js"></script>

<body style="background-color:#1c87c9;">

<h1> Telecom Customer churn analysis </h1>

</body>

</head>

<title>Churn Prediction</title>

<div class="container">

<div class="row">

<form action="http://localhost:5000/" method="POST">

<div class="col-sm-9">

<div class="form-group purple-border">

<label for="comment">SeniorCitizen:</label>

<textarea class="form-control" rows="2" id="query1" name="query1" rows="2" cols="5" autofocus>{{query1}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">MonthlyCharges:</label>

<textarea class="form-control" rows="2" id="query2" name="query2" rows="2" cols="5" autofocus>{{query2}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">TotalCharges:</label>

<textarea class="form-control" rows="2" id="query3" name="query3" rows="2" cols="5" autofocus>{{query3}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">gender:</label>

<textarea class="form-control" rows="2" id="query4" name="query4" rows="2" cols="5" autofocus>{{query4}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">Partner:</label>

<textarea class="form-control" rows="2" id="query5" name="query5" rows="2" cols="5" autofocus>{{query5}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">Dependents:</label>

<textarea class="form-control" rows="2" id="query6" name="query6" rows="2" cols="5" autofocus>{{query6}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">PhoneService:</label>

<textarea class="form-control" rows="2" id="query7" name="query7" rows="2" cols="5" autofocus>{{query7}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">MultipleLines:</label>

<textarea class="form-control" rows="2" id="query8" name="query8" rows="2" cols="5" autofocus>{{query8}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">InternetService:</label>

<textarea class="form-control" rows="2" id="query9" name="query9" rows="2" cols="5" autofocus>{{query9}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">OnlineSecurity:</label>

<textarea class="form-control" rows="2" id="query10" name="query10" rows="2" cols="5" autofocus>{{query10}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">OnlineBackup:</label>

<textarea class="form-control" rows="2" id="query11" name="query11" rows="2" cols="5" autofocus>{{query11}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">DeviceProtection:</label>

<textarea class="form-control" rows="2" id="query12" name="query12" rows="2" cols="5" autofocus>{{query12}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">TechSupport:</label>

<textarea class="form-control" rows="2" id="query13" name="query13" rows="2" cols="5" autofocus>{{query13}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">StreamingTV:</label>

<textarea class="form-control" rows="2" id="query14" name="query14" rows="2" cols="5" autofocus>{{query14}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">StreamingMovies:</label>

<textarea class="form-control" rows="2" id="query15" name="query15" rows="2" cols="5" autofocus>{{query15}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">Contract:</label>

<textarea class="form-control" rows="2" id="query16" name="query16" rows="2" cols="5" autofocus>{{query16}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">PaperlessBilling:</label>

<textarea class="form-control" rows="2" id="query17" name="query17" rows="2" cols="5" autofocus>{{query17}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">PaymentMethod:</label>

<textarea class="form-control" rows="2" id="query18" name="query18" rows="2" cols="5" autofocus>{{query18}}</textarea>

</div>

<div class="form-group purple-border">

<label for="comment">tenure:</label>

<textarea class="form-control" rows="2" id="query19" name="query19" rows="2" cols="5" autofocus>{{query19}}</textarea>

</div>

</div>

<div class="col-sm-3">

<button type="submit" class="btn btn-primary" name="submit">SUBMIT</button>

</div>

</form>

</div>

<div class="row">

<div class="col-sm-9">

<textarea class="form-control" rows="2" id="comment" name="query6" rows="2" cols="5" autofocus>{{output1}}</textarea>

<textarea class="form-control" rows="2" id="comment" name="query7" rows="2" cols="5" autofocus>{{output2}}</textarea>

</div>

</div>

</div>

</body>

</html>

**4.3. DETAILS OF DATA SET:**

* Churn analysis :

[**https://www.kaggle.com/datasets/blastchar/telco-customer-churn?resource=download**](https://www.kaggle.com/datasets/blastchar/telco-customer-churn?resource=download)

* Total 23 attributes, out of that “churn “is used for prediction of the output.

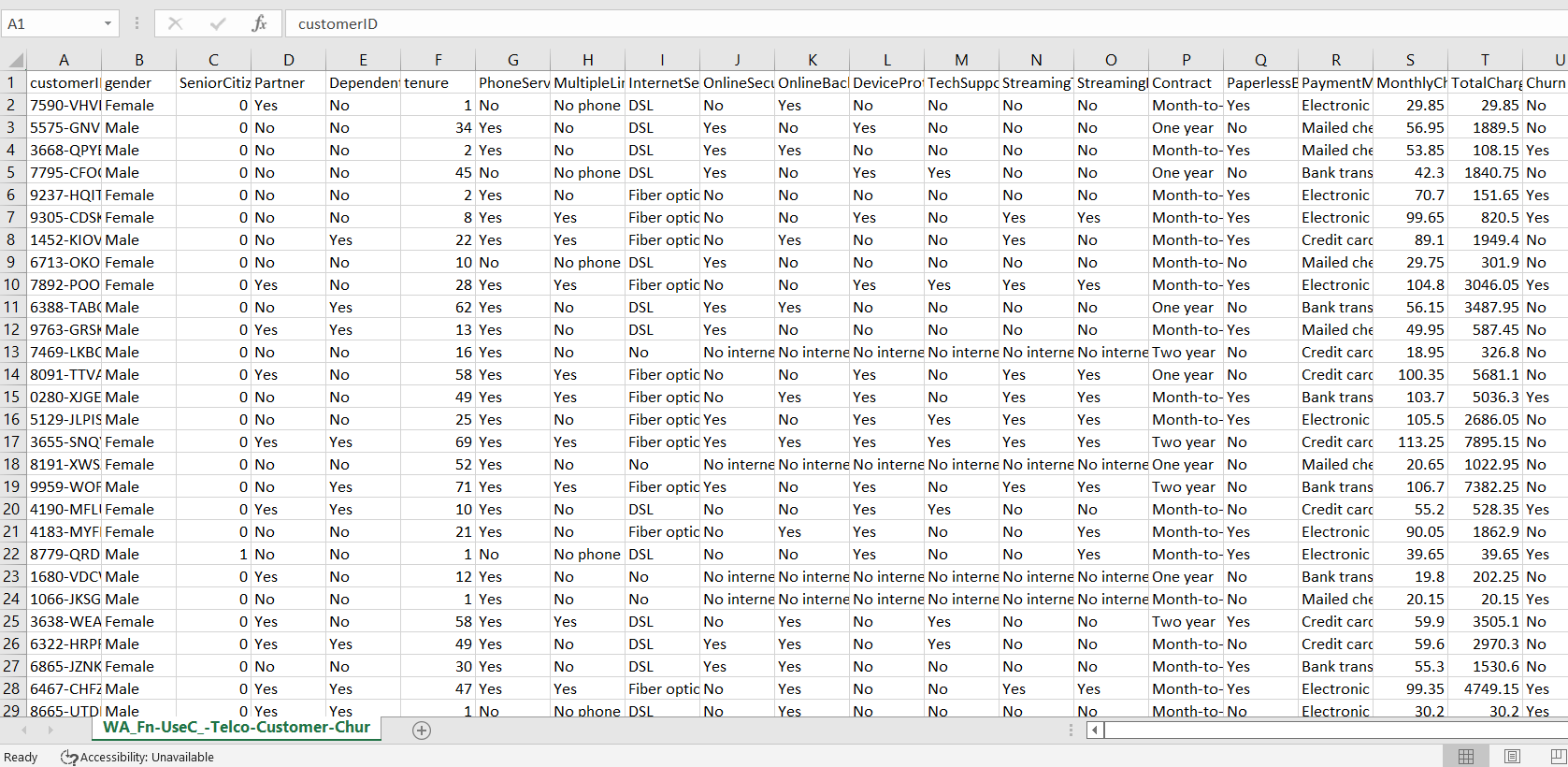


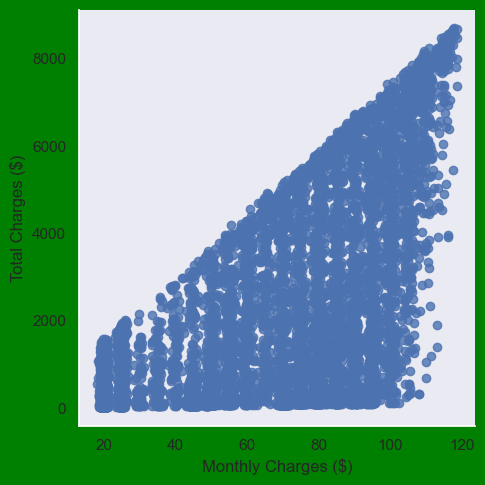
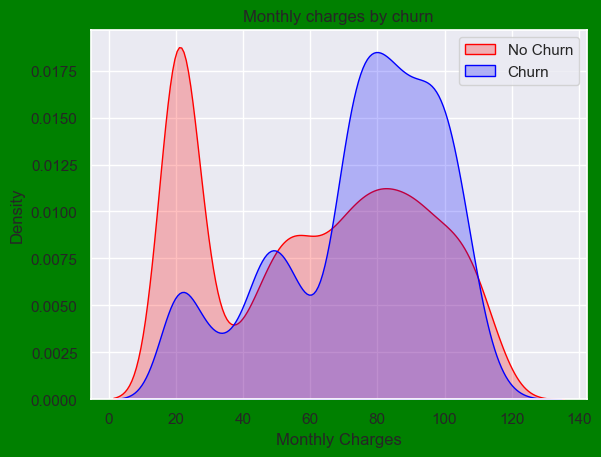
Fig. Dataset Attributes

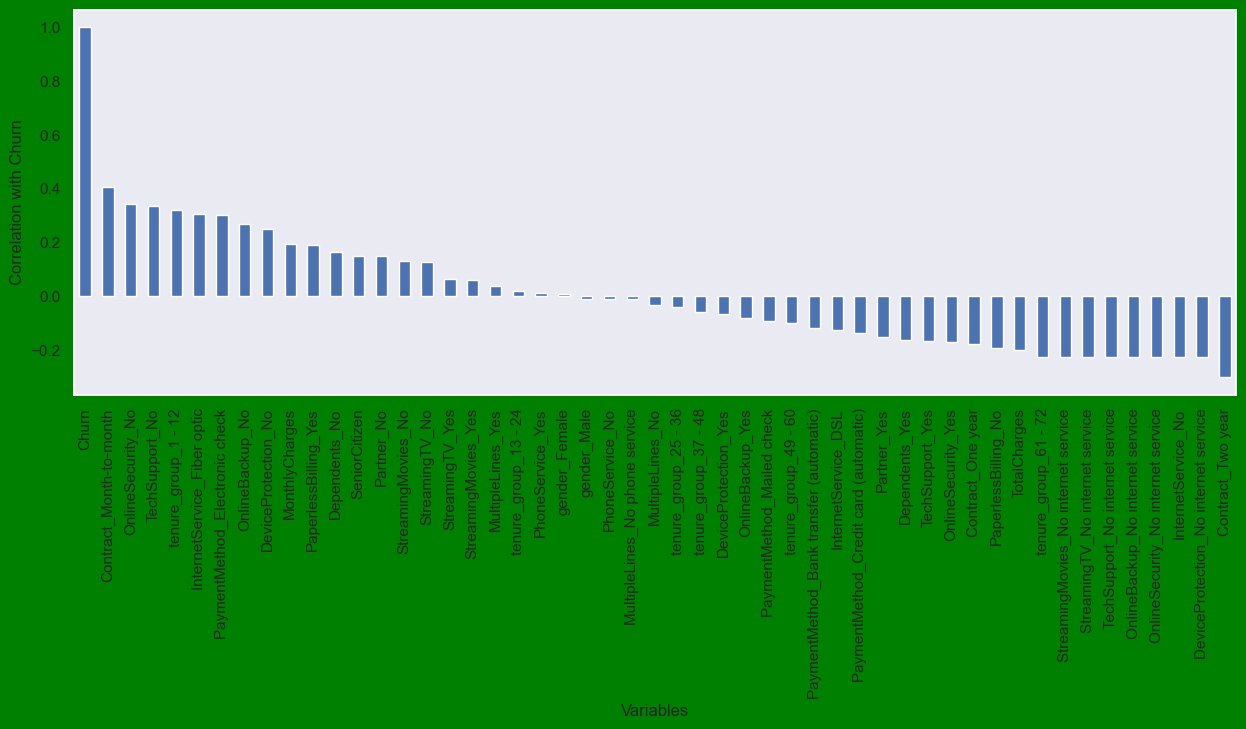
|  |  |  |  |
| --- | --- | --- | --- |
| S.no | Observations | Description | Values |
| i. | Cust\_id | Customer ID | string |
| ii. | Gender | Gender of the subject | Female or Male |
| iii. | Senior citizen | Senior citizens | 0 or 1 |
| iv. | Partner | Partners involved | Yes or No |
| v. | Dependents | Who are dependents | Yes or No |
| vi. | Tenure | Length of time of customer with the company | 1-100 values |
| vii. | Phone service | Type of service | Yes or No |
| viii. | Multiple lines | Having one or more services | Yes or No |
| ix. | Internet service | Internet provider | Fiber optic or DSL |
| x. | Online Security | Security provided | Yes or No |
| xi. | Online backup | Backup services | Yes or No |

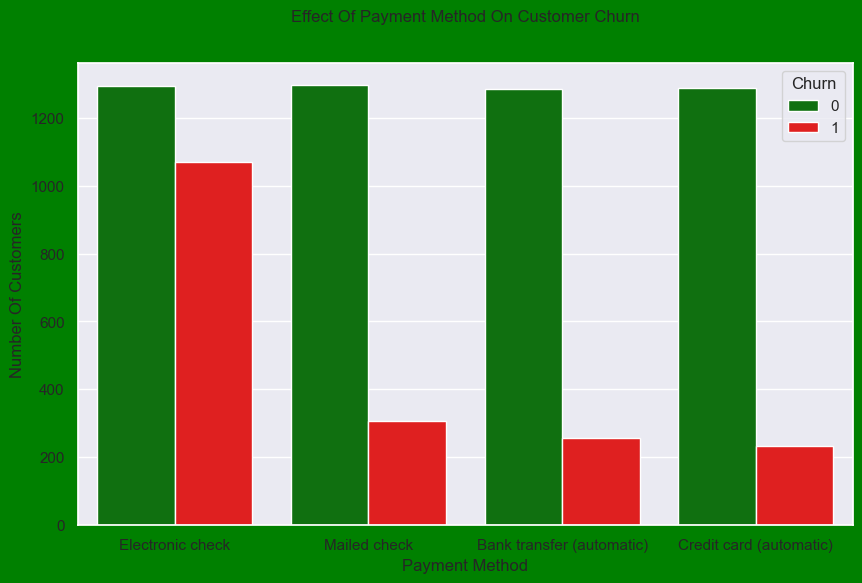
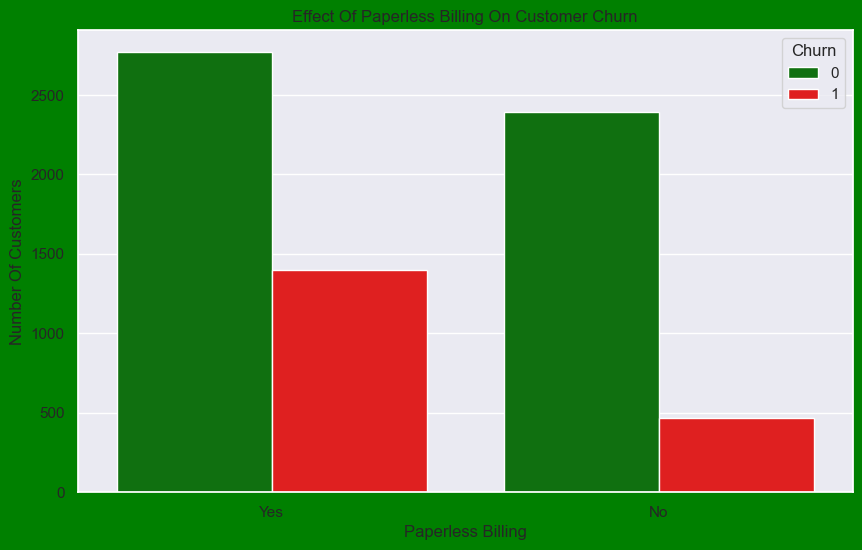
TABLE1: Attributes of the dataset

**EDA :**

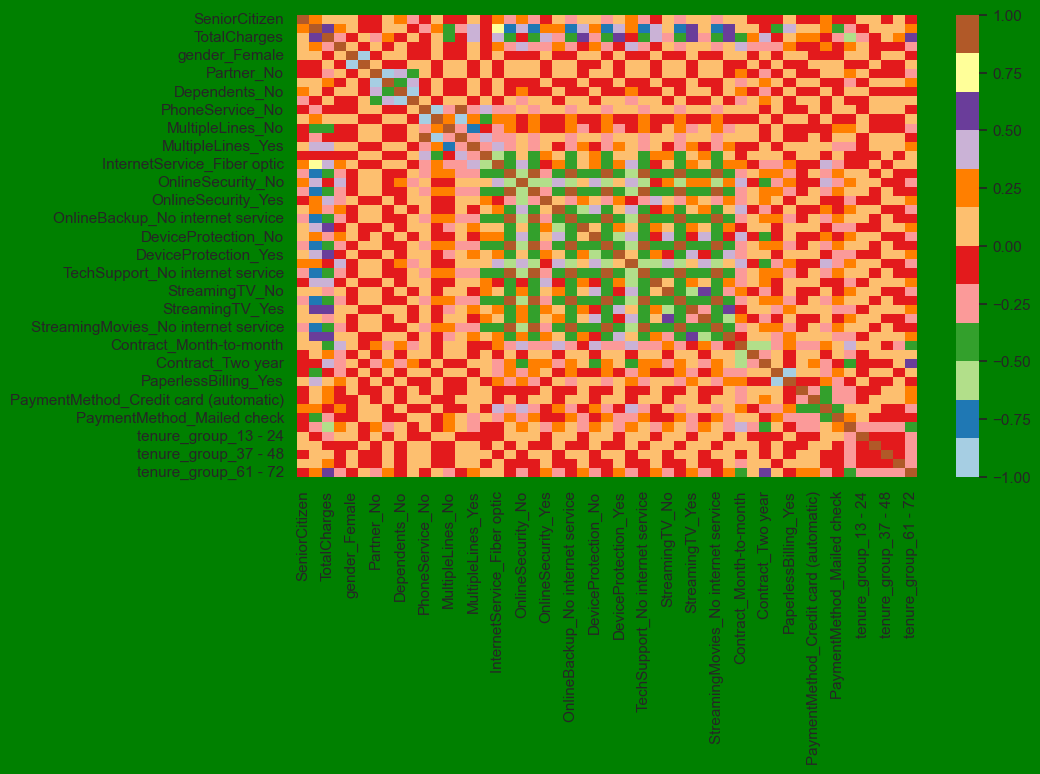
Below are the Analysis performed considering the attributes and their patterns

 ****

****

Heatmap :



**4.4. PERFORMANCE ANALYSIS:**

In this study, different machine learning approaches such as Decision Tree, Random Forest, and KNN are used to forecast customer churn analysis. The Churn dataset has 21 variables in all, but only one is utilized to predict heart churn. Several patient factors, such as gender, monthly costs, Internet service, monthly billing, and so on, are taken into consideration for this project. Individual algorithm accuracy must be assessed, and the algorithm with the highest accuracy is utilized to forecast churn. To evaluate the experiment, several evaluation metrics including as accuracy, confusion matrix, precision, recall, and f1-score are utilized.

**Accuracy**- Accuracy is defined as the percentage of correct predictions to total forecasts. It is shown as:

**Accuracy** = (TP + TN) /(TP+FP+FN+TN)

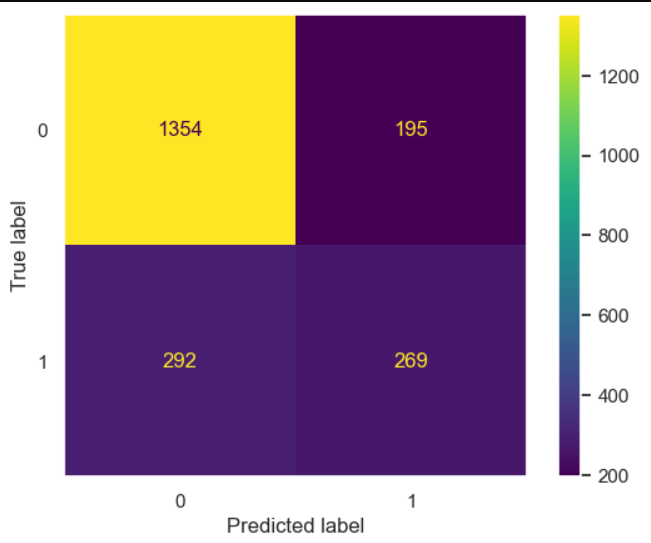
**Confusion Matrix**- It produces a matrix and calculates the overall performance of the system.

TP: True positive

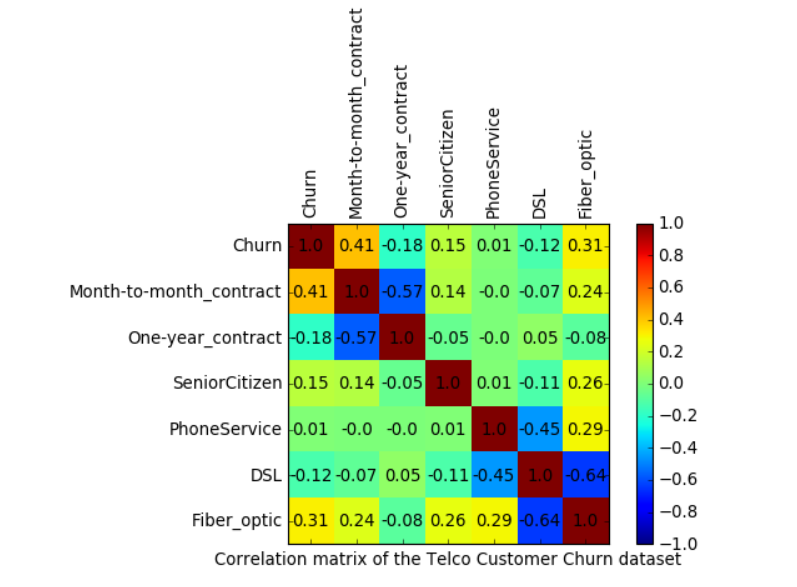
FP: False Positive

FN: False Negative

TN: True Negative



**correlation Matrix**: In machine learning, the correlation matrix is utilized for feature selection. It demonstrates the interdependence of several properties

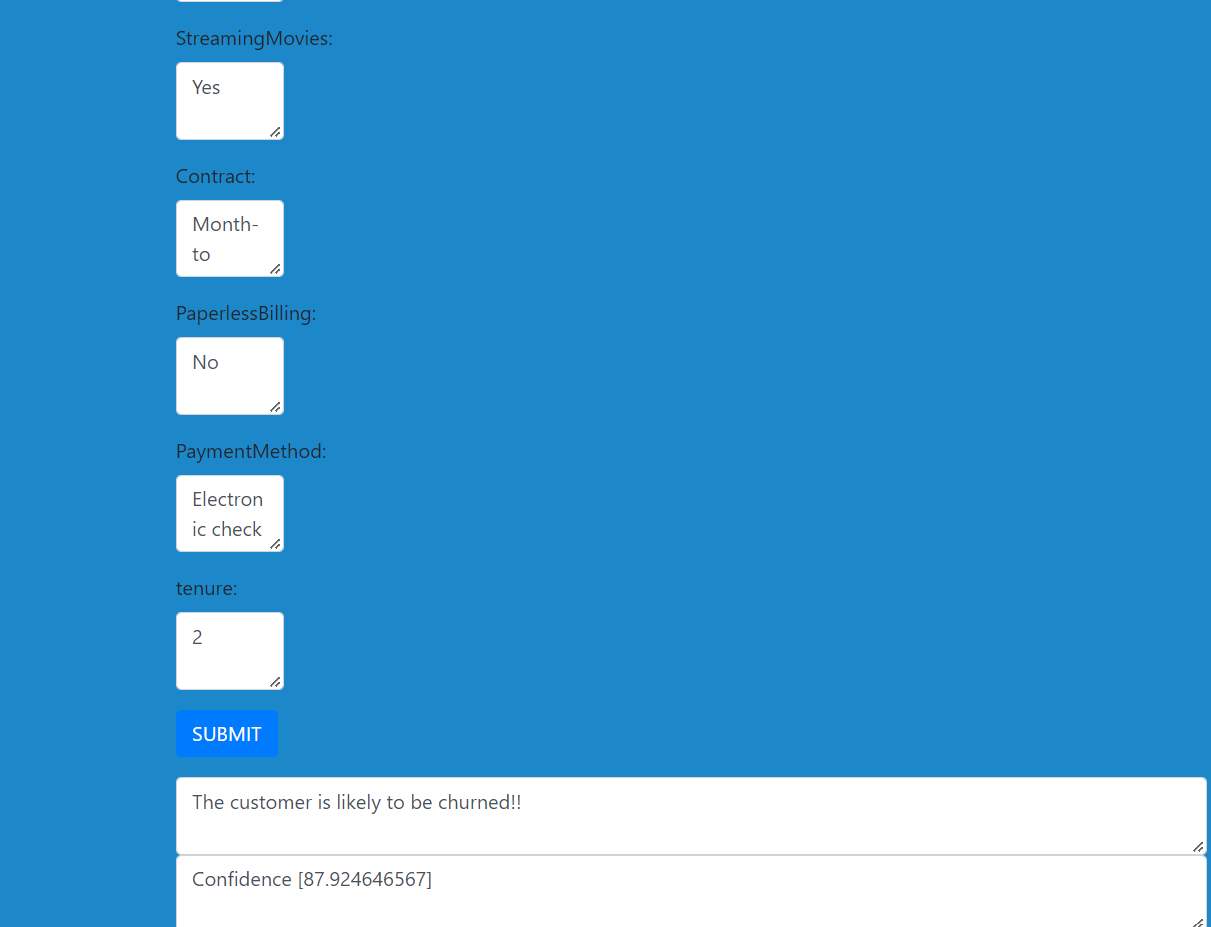
****

**Precision** is defined as the proportion of accurately predicted positive outcomes in relation to the total number of positive results predicted by the technique.

**Recall:** The recall statistic compares the number of positive results that were actually attained to the number predicted by the algorithm.

**F1 Score:** It is the best balance of being correct and remembering stuff. The findings can be used to conduct an examination of the test's reliability. In its range, this indicator can accept values ranging from 0 to 1

**4.6. Input/Output:**

****

**4.7. Result:**

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| **Decision tree** | **79.45%** |
| **Random forest** | **80.09%** |
| **KNN** | **74%** |
| **Decision tree(SMOTEENN)** | **90.81%** |
| **Random forest(SMOTEENN)** | **92.13%** |

**5. Conclusion:**

We have covered every aspect of a machine learning project by using the Telco customer Churn dataset. To begin, we performed analysis and cleanup of the data using data visualization. After that, the categorical data were transformed into numerical variables so that a model of machine learning could be built (feature engineering). After transforming the data, we then attempted to use Three different machine-learning algorithms with the settings that were predefined for each. In the end, we were able to perfect our model by playing around with the hyperparameters of both our decision tree and our random classifier.

**References:**

1. Qureshi SA, Rehman AS, Qamar AM, Kamal A, Rehman A. Telecommunication subscribers’ churn prediction model using machine learning. In: Eighth international conference on digital information management. 2013. p.131–6.
2. Umayaparvathi V, Iyakutti K. A survey on customer churn prediction in telecom industry: datasets, methods and metric. Int Res J Eng Technol.2016;3(4):1065–70.
3. Burez D, den Poel V. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.
4. Gerpott TJ, Rams W, Schindler A. Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommun Policy. 2001;25:249–69.