Project Report Template

Introduction :

The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators raised the level of competition . Companies are working hard to survive in this competitive market depending on multiple strategies.

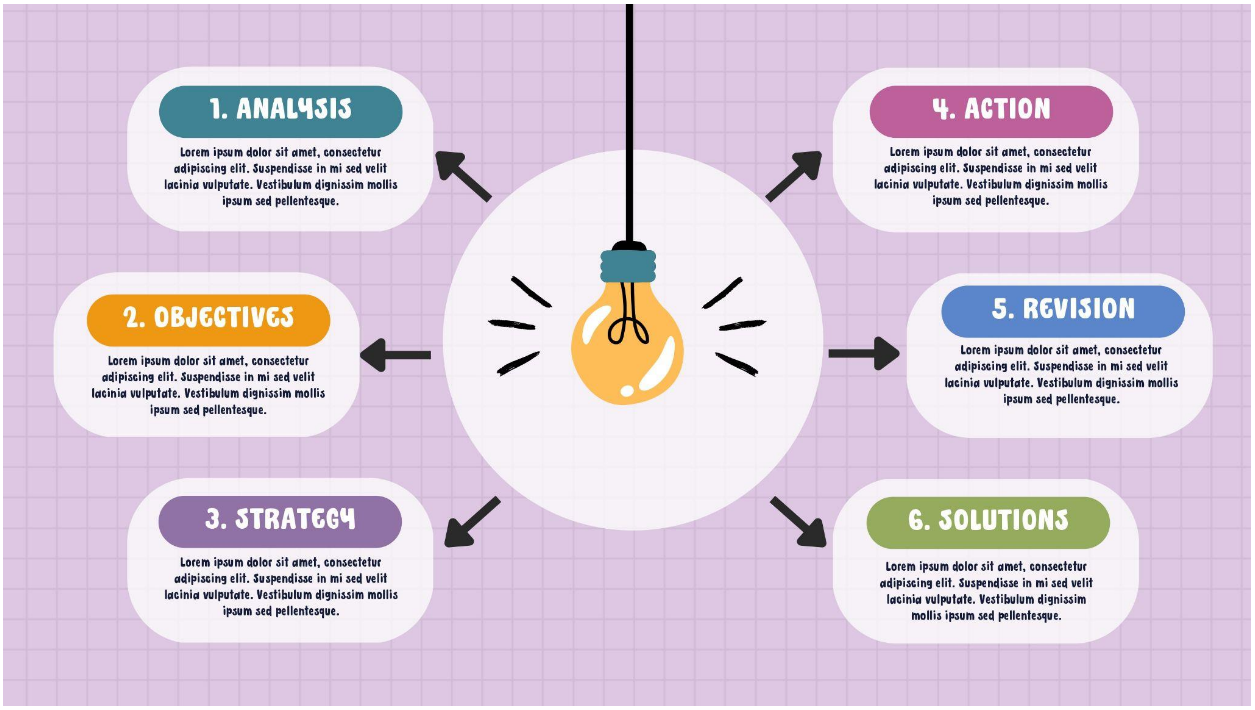
Customers' churn is a considerable concern in service sectors with high competitive services. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase.

Many research confirmed that machine learning technology is highly efficient to predict this situation. This technique is applied through learning from previous data

The data used in this research contains all customers' information throughout nine months before baseline. The volume of this dataset is about 70 Terabyte on HDFS "Hadoop Distributed File System", and has different data formats which are structured, semi-structured, and unstructured. The data also comes very fast and needs a suitable big data platform to handle it. The dataset is aggregated to extract features for each customer.

We built the social network of all the customers and calculated features like degree centrality measures, similarity values, and customer's network connectivity for each customer. SNA features made good enhancement in AUC results and that is due to the contribution of these features in giving more different information about the customers.

Problem definition & design thinking



Ideation & brainstorming map

Person 1

- Obesity and hypertension: Many medical conditions can cause KD.
- Family history: If anyone in your family has kidney disease, dialysis, or kidney transplantation, you may be more likely to develop kidney disease than someone without this family history.
- Medicines: Some medicines can cause or exacerbate kidney disease, such as over-the-counter pain medicines.
- Age and race: older people and certain racial groups may have a higher chance of developing renal disease.

The diagnosis of kidney disease in early stage saves the patient from serious complications. To predict the kidney diseases, the factors that cause it must be studied carefully.

Classifying data with missing values is a challenge.
The used dataset has missing values, which reduce the efficiency, so it must be removed before analyzing data.
The missing values can be determined in two points of view, cases (records) or attributes. In cases(record) point of view, the missing values degree may be simple, medium, or complex. It is simple degree if the case (record) has a missing value in one attribute at most. It is medium if the case (record) has missing values in 2% to 50% of the total number of attributes. While It is complex if the case (record) has missing values in at least 50% up to 80% of attributes.

Deep Belief Networks are interactive systems that built on stacking RBM that trained with CD The algorithm that determines the optimum locale for each layer and the next stacked RBM layer takes those optimally trained values and searches for the optimum locale again that is the cause of the greedy algorithm for learning works of DBN training layer by layer as shown in

Person 2

Chronic kidney disease (CKD) is one of the most life-threatening disorders. To improve survivability, early discovery and good management are encouraged. In this paper, CKD was diagnosed using multiple optimized neural networks against traditional neural networks on the UCI machine learning dataset, to identify the most efficient model for the task. The study works on the binary classification of CKD from 24 attributes. For classification, optimized CNN (OCNN), ANN (OANN), and LSTM (OLSTM) models were used as well as traditional CNN, ANN, and LSTM models.

The highest validation accuracy among the tradition models were achieved from CNN with 92.71%, whereas OCNN, OANN, and OLSTM have higher accuracies of 98.75%, 96.25%, and 98.5%, respectively. Additionally, OCNN has the highest AUC score of 0.99 and the lowest compilation time for classification with 0.00447 s, making it the most efficient model for the diagnosis of CKD.

One of the non-communicable diseases with the quickest growth rate is chronic kidney disease (CKD), a significant cause of death and disease. It has affected more than 10% of the world's population, and millions of people die each year [1]. According to the Global Burden of Disease Study, almost 697.5 million cases of all-stage CKD were registered in 2017, resulting in a global prevalence of 9.1%, up 29.3% from 1990

Chronic kidney disease treatment
is both expensive and ineffective.
In contrast, only about 5% of
ividuals with early CKD are
ware of their condition.
erular damage has reached
0% and is usually irreversible
e CKD is identified. In this
rd, accurate chronic renal
se prognosis can be highly
beneficial.

A convolutional neural network with a gated recurrent unit (CNN-GRU), deep belief network (DBN), and kernel extreme learning machine (KELM) are proposed. They achieved the highest accuracy of 96.91% using the EDL-CDSS approach. Akter, et al. [8], in 2021, deployed seven state-of-the-art deep learning algorithms, ANN, LSTM, GRU, bidirectional LSTM, bidirectional GRU, MLP, and simple RNN, for CKD prediction and classification along with the numerous clinical features of CKD that have been proposed.

Person 3

They employed the deep neural network (DNN) model to predict if CKD would be present in a patient. The DNN model generated a 98% accuracy rate. Of the 11 variables, creatinine and bicarbonate impact CKD prediction most. In 2020, Ma, et al. [10] suggested chronic kidney illness utilizing a heterogeneous modified artificial neural network based on deep learning.

used classification techniques including a artificial neural network (ANN) and a support vector machine (SVM). Using the mean of the corresponding attributes, they replaced all missing values in the datasets. Additionally, they employed a 10-fold cross-validation procedure to divide the training and test datasets according to the ratio (90:10). In their proposed method, ANN performs better. Using the optimized features, the accuracy is 99.75%, while, from SVM, the accuracy is 97.75%.

They used a hybrid deep learning convolution neural network–support vector machine (CNN-SVM) model to make predictions. The proposed model is put to the test in experiments, and its performance is compared to that of a traditional CNN.

- Performing data preprocessing to confirm accuracy through the detection of extreme situations, removing noisy data and missing values.
- Choosing the best classifier, by contrasting regularly used classification methods with CKD studies from the literature review and ablation study.
- An optimized model based on CNN architecture is proposed.
- The precision, recall, specificity, and F1 score are calculated to support the model accuracy. The effectiveness of the models is evaluated using the loss function as well.
- The AUC value is computed in order to assess the proposed model

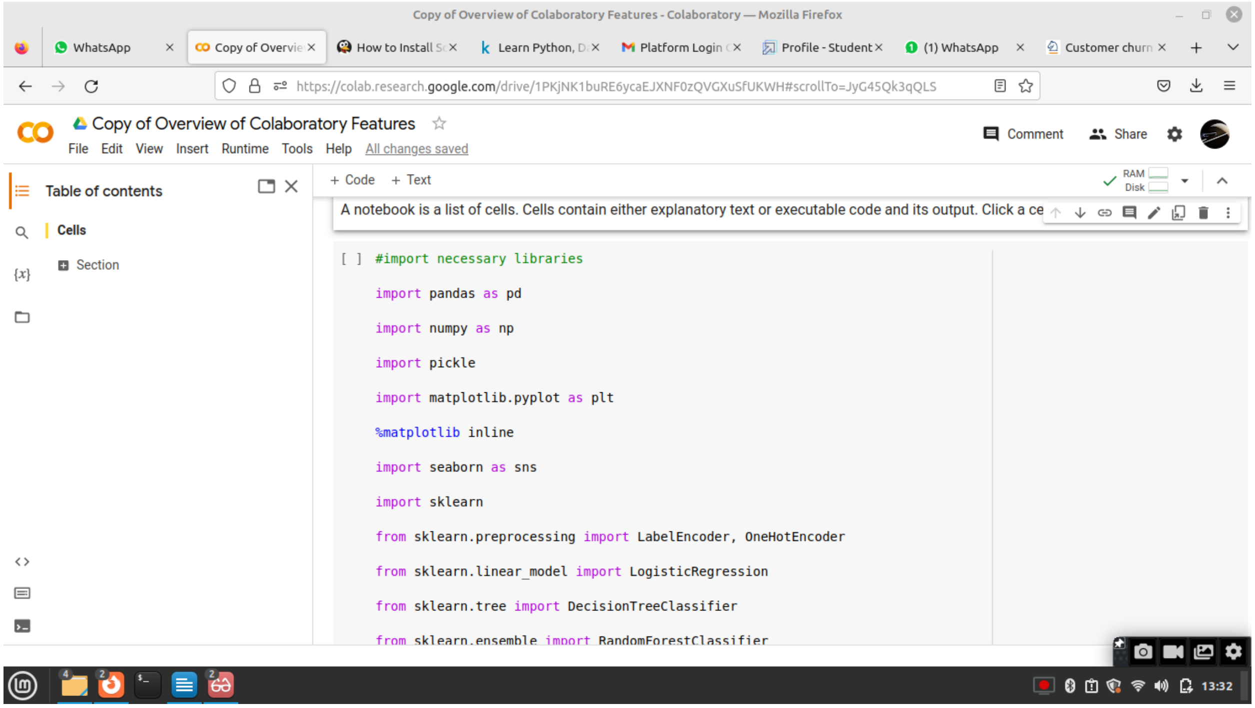
Person 4

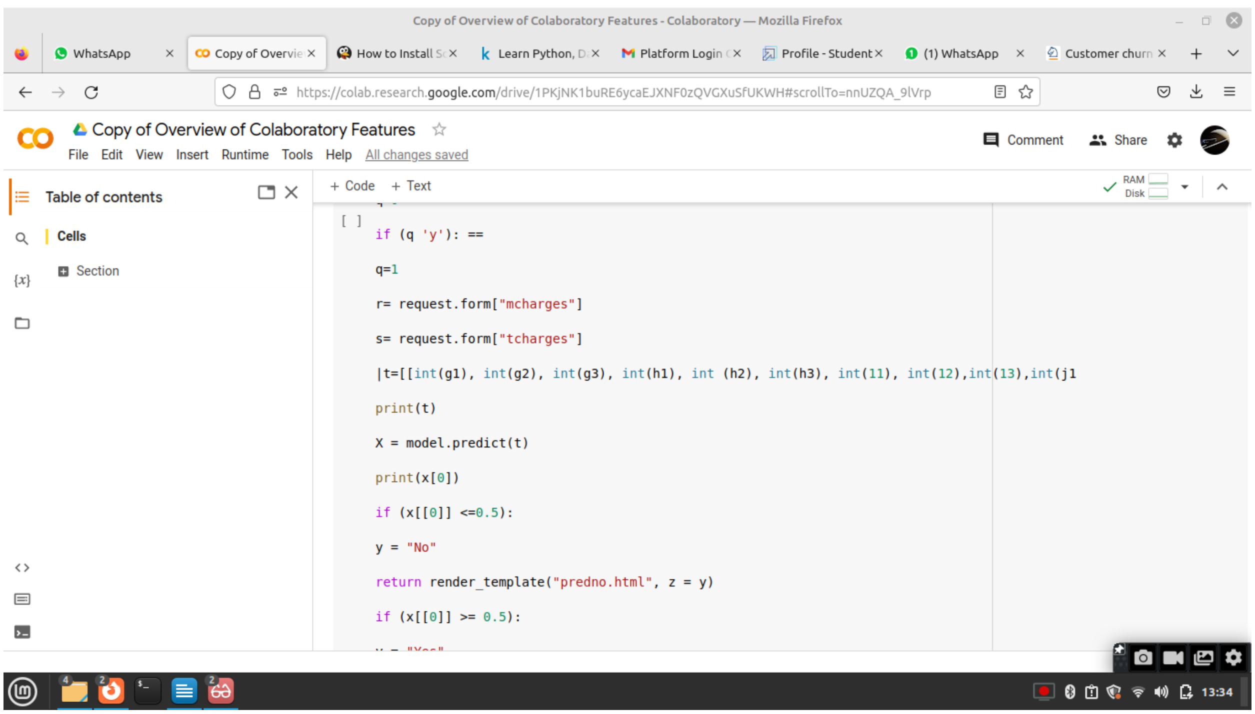
Our suggested approach is built on kidney disease datasets. We split our dataset into train and test (80% data on train and 20% data on test) and showed that the model was free of overfitting issues. All the classifiers introduced were designed and obtained the best accuracy from the dataset. Figure 1 presents the complete aspects of our approach.

The UCI machine learning repository's chronic kidney disease dataset was used in this study. The dataset has 400 records, each composed of a set of 25 attributes [14]. The 'classification' variable indicates whether the patient has CKD or not. This variable is preserved as a dependent or target variable during the classification process. The rest variables are fed as input to the classifier model to predict the target class. The type for the variables Age, BP, Bgr, Bu, Sc, Sod, Pot, Hemo, Pcv, Wc, and Rc is numerical, whereas the variables Sg, Al, Su, Rbc, Pc, Pcc, Ba, Htn, Dm, Cad, Appet, Pe, Ane, and classification are nominal in type. Our target variable (classification) has two categories of nominal value (ckd and not ckd). To develop our proposed approach, we mapped this value into numerical values (0, 1).

In our view, Figure 2 represents correlated features with the predicted class attribute (classification). The attribute values define the strength of the correlated features at the right portion (range from −0.6 to 0.6), in accordance with the lightness of color. The Figure represents 'pcv' and 'rc' as having a strong correlation with 'htn', having the value of 0.74, 0.68; whereas 'sod','htn' has a lesser correlation with 'hemo'. having the value of −0.62, −0.5 approximately.

he implementation of optimized CNN, ANN, and LSTM classifiers is explained in this section. To obtain the training and testing datasets, the preprocessed data is divided in an 80:20 ratio. The training dataset is used to fit the classifier models, and the testing dataset is used to collect predictions. The optimized CNN classifier is implemented with a kernel regularization parameter of C = 1.0 and the activation function is ReLu. For the greatest known performance, the LSTM ...fier is learned at a rate of 1.0 and with ...imators. The predicted performance ... classifiers will be used to justify the ...posed classifier's performance.


Copy of Overview of Colaboratory Features - Colaboratory — Mozilla Firefox
WhatsApp
Copy of Overvie
How to Install S
Learn Python, D
Platform Login
Profile - Student
(1) WhatsApp
Customer churn
https://colab.research.google.com/drive/1PKjNK1buRE6ycaEJXNF0zQVGXuSfUKWH#scrollTo=JyG45Qk3qQLS
Copy of Overview of Colaboratory Features
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Table of contents
Cells
Section
+ Code + Text
RAM
Disk
A notebook is a list of cells. Cells contain either explanatory text or executable code and its output. Click a ce
[ ]
#import necessary libraries
import pandas as pd
import numpy as np
import pickle
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import sklearn
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
13:32


Copy of Overview of Colaboratory Features - Colaboratory — Mozilla Firefox
WhatsApp
Copy of Overvie
How to Install S
Learn Python, D
Platform Login
Profile - Student
(1) WhatsApp
Customer churn
https://colab.research.google.com/drive/1PKjNK1buRE6ycaEJXNF0zQVGXuSfUKWH#scrollTo=nnUZQA_9lVrp
Copy of Overview of Colaboratory Features
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Table of contents
Cells
Section
+ Code + Text
RAM
Disk
[ ]
if (q 'y'): ==
q=1
r= request.form["mcharges"]
s= request.form["tcharges"]
|t=[[int(g1), int(g2), int(g3), int(h1), int (h2), int(h3), int(11), int(12),int(13),int(j1
print(t)
X = model.predict(t)
print(x[0])
if (x[[0]] <=0.5):
y = "No"
return render_template("predno.html", z = y)
if (x[[0]] >= 0.5):
13:34

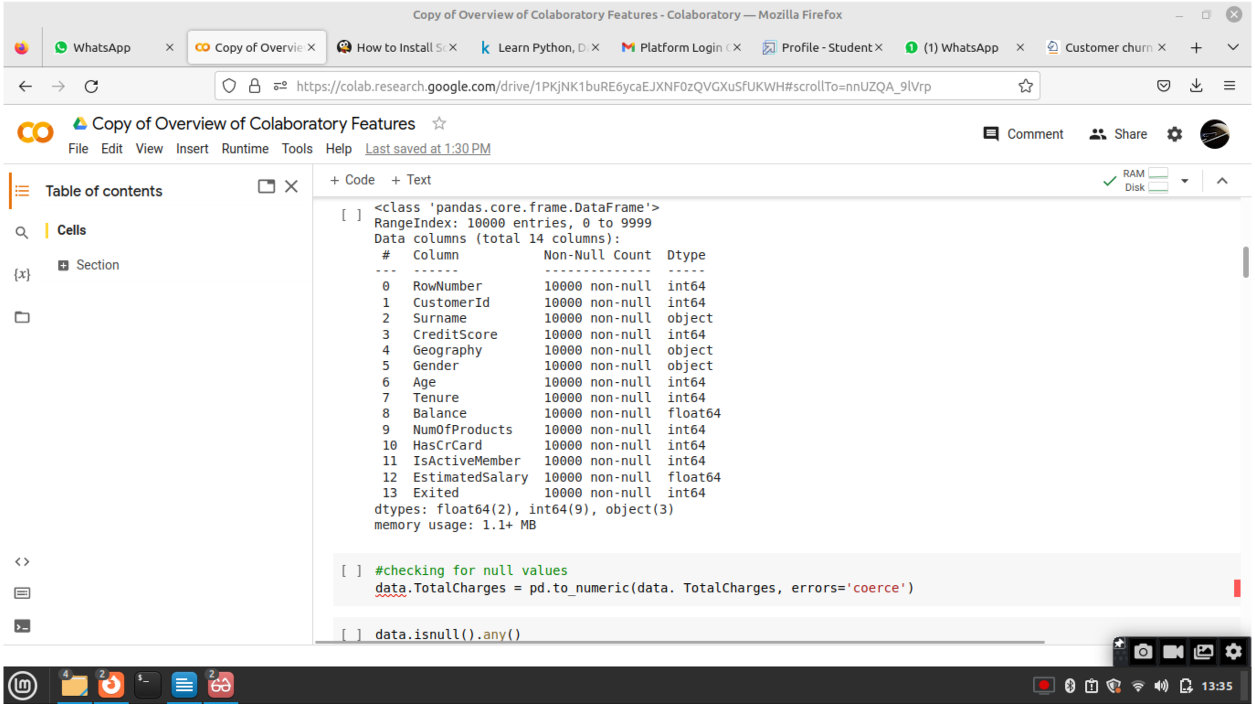
```
#printing the train accuracy and test accuracy respectively
RandomForest(x_train,x_test,y_train,y_test)
```

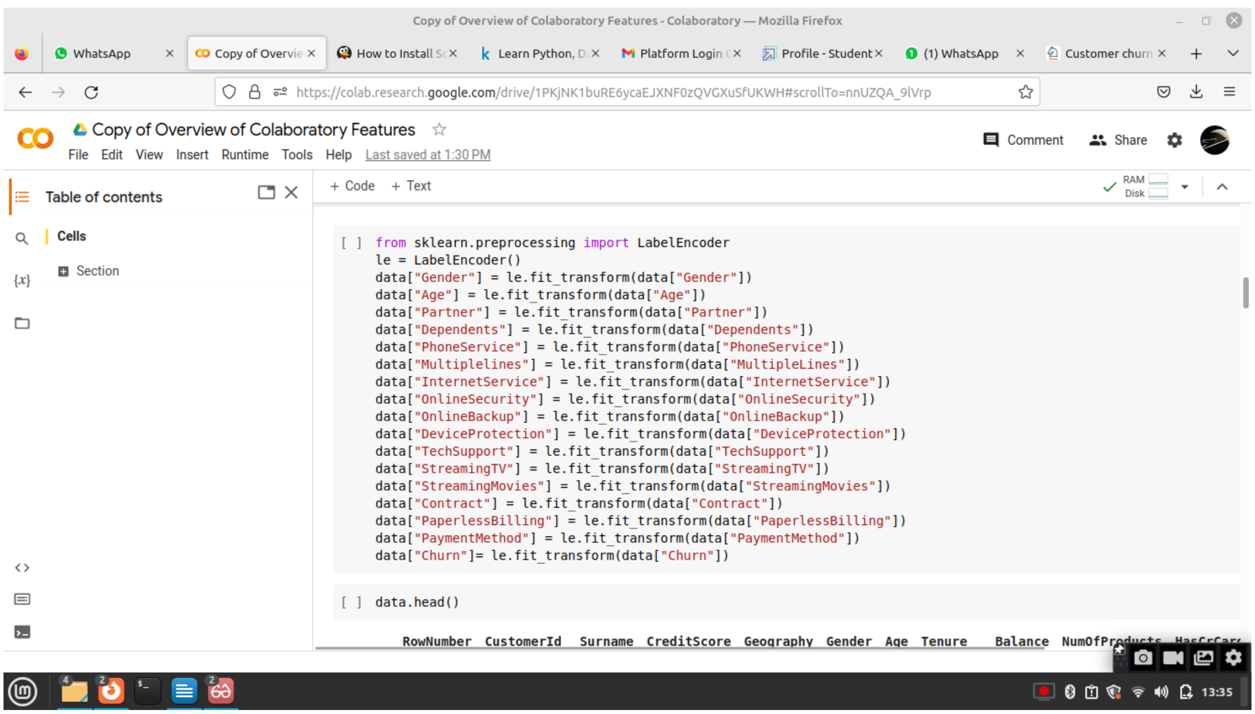
```
0.8570910848030925
0.7913043478260869
**KNN**
Confusion Matrix
 [[730 303]
  [129 908]]
Classification Report
                   precision    recall  f1-score   support

               0       0.85      0.71      0.77      1033
               1       0.75      0.88      0.81      1037

        accuracy                           0.79      2070
       macro avg       0.80      0.79      0.79      2070
    weighted avg       0.80      0.79      0.79      2070
```

```
# Importing the Keras libraries and packages
      import keras
     from keras.models import Sequential
     from keras.layers import Dense
```





Advantages of Telecommunication :

Quick and accessible communication

Lack of time period

- Saves time
- Saves gasoline (do not need to drive distance)
- More than two people can communicate with at least one another at an equivalent time
- Next “best thing” to being there
- Easy to exchange ideas and knowledge via phone and/or fax
- Worldwide access
- Easy access to the people you would like to contact.
- Less effort in using transportation just to satisfy a private personally.
- You can just occupy your home and use a telephone or a cellphone if you would like to speak to someone.
- Enable end-users to speak electronically and share hardware, software, and data resources.
- This make corporation to do the transaction at the point only and in a very fast way from many remote locations, exchange business documents electronically with customers and suppliers, or remotely monitor and control production processes.
- Interconnect the pc systems of a business so their computing power is often shared by end-users throughout an enterprise.
- Make the organization work with collaboration and communication among the staff inside and out of doors a corporation.
- Speed
- Develops new products and inventions

Disadvantages of Telecommunication :

- Cultural Barrier
- Misunderstanding
- Prank calls
- Sometimes expensive
- High electric bills
- Remote areas don’t have access
- Remote areas might not be ready to afford the necessary equipment
- Cannot see whom you’re speaking with
- Cannot see facial expressions, therefore results in misunderstandings
- Cultural barriers
- Poor connections or downed power lines during/after storms

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, this research aimed to build a system that predicts the churn of customers in SyriaTel telecom company. These prediction models need to achieve high AUC values. To test and train the model, the sample data is divided into 70% for training and 30% for testing. We chose to perform cross-validation with 10-folds for validation and hyperparameter optimization. We have applied feature engineering, effective feature transformation and selection approach to make the features ready for machine learning algorithms. In addition, we encountered another problem: the data was not balanced. Only about 5% of the entries represent customers’ churn. This problem was solved by undersampling or using trees algorithms not affected by this problem. Four tree based algorithms were chosen because of their diversity and applicability in this type of prediction. These algorithms are Decision Tree, Random Forest, GBM tree algorithm, and XGBOOST algorithm. The method of preparation and selection of features and entering the mobile social network features had the biggest impact on the success of this model, since the value of AUC in SyriaTel reached 93.301%.