Advanced Data Analytics Mini Project

Churn Modelling Prediction in Banking Sector

Student Information

- A. Mahendranadh Chowdary - 19BCE7058
- Ch. Rakesh – 19BCE7014
- Ch. Surya Varshit – 19BCE7494
- K. Gokul – 19BCD7006

Abstract

In the banking sectors one of the main problems is that customers switch their banking services from one bank to another and close the account very frequently. This can lead to the loss and some of the branches have to be closed. To avoid this many banking sectors use this Churn Modelling Prediction and based on it they can predict the customers who may leave the banking services in the future and the banks can use various strategies to prevent the customers from leaving their services by giving them relaxations and getting them involved in new schemes and policies.

Main Objective

The main objective of this Churn Modelling Prediction is to analyze how many customers have left a particular service in the given period of time and how it impacts the business of the company. In this Mini Project, the Churn Modelling Prediction is done for Banking Sector where we evaluate the number of customers who have exited from the banking services

Related Works

- https://www.researchgate.net/publication/342424673_Prediction_of_Customer_Churn_in_Banking_Industry

- https://www.researchgate.net/publication/357539438_Customer_churn_analysis_in_banking_sector_Evidence_from_explainable_machine_learning_models

- https://arxiv.org/ftp/arxiv/papers/1912/1912.11346.pdf

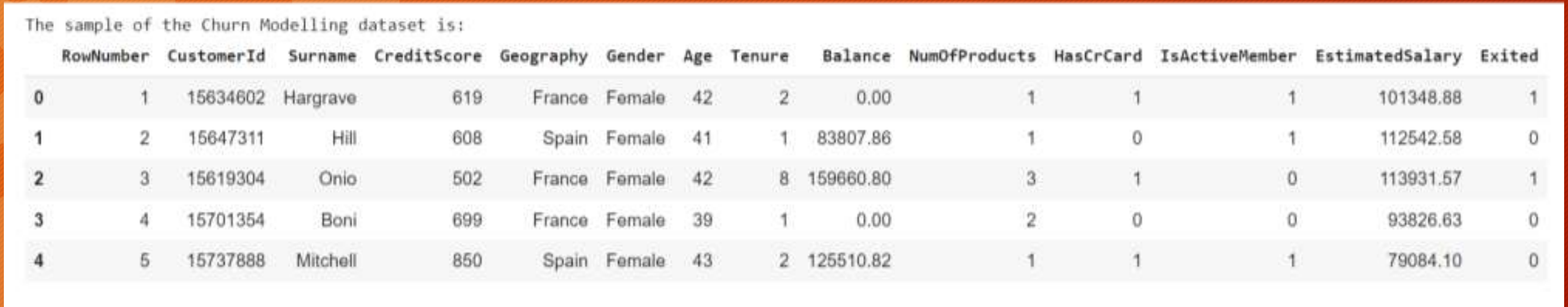
- https://link.springer.com/article/10.1007/s00521-022-07067-x#Sec12

- https://drive.google.com/file/d/1xB8MDR5d6FOfvVvfA4fYPXgn1Qn2LZNu/view?usp=sharing

Contribution

Over all working Principle Structure

Understanding the Dataset



The sample of the Churn Modelling dataset is:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

This is the initial dataset which is being used for the Churn Modelling Prediction. The details regarding the various attributes regarding the dataset have been explained in the upcoming slides.

Attributes of the Dataset

Feature Attributes

- Row Number: This attribute defines the indexing of records
- Customer Id: This attribute is the unique Id given to each customer
- Surname: This attribute contains the surnames of customers
- Credit Score: This attribute depicts a customer’s credit worthiness
- Geography: This attribute contains geographical location of the customer
- Gender: This attribute contains the gender of the customer
- Age: This attribute contains the age of the customer
- Tenure: This attribute determines the number of years the customer has been using the banking services

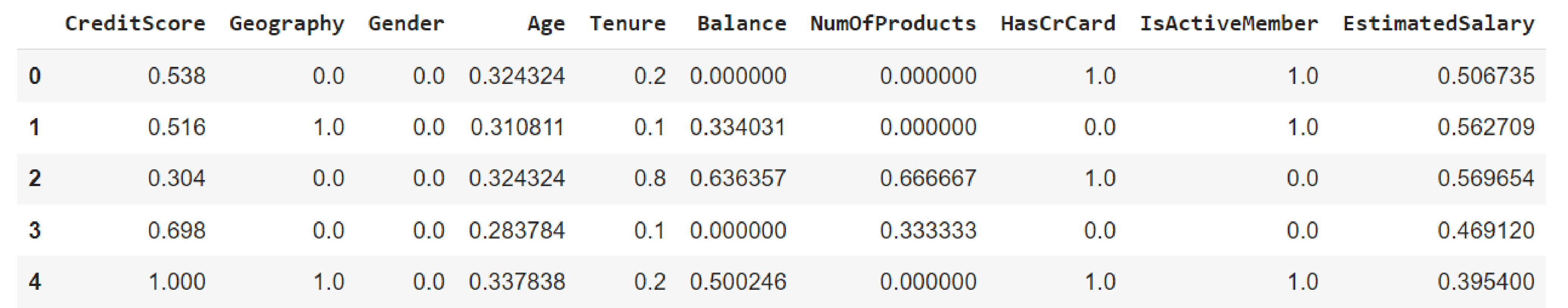
- Balance: This attribute contains the customer’s balance in the bank
- Number Of Products: This attribute determines the number of products the customer bought through the bank.
- Has Cr Card: This attribute depicts if customer has credit card or not
- Is Active Member: This attribute shows if the customer has been active or not
- Estimated Salary: This attribute contains the customer’s estimated salary

Target Attribute

- Exited: This attribute determines if the customer left the banking services or not.

Dataset After Pre-Processing

- The dataset has been checked for null values, removal unnecessary columns like Row Number, Surname, and Customer Id has been done.
- Label encoded the categorical attributes geography and gender, and normalized for better prediction. Now, the dataset of feature attributes is :



	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	0.538	0.0	0.0	0.324324	0.2	0.000000	0.000000	1.0	1.0	0.506735
1	0.516	1.0	0.0	0.310811	0.1	0.334031	0.000000	0.0	1.0	0.562709
2	0.304	0.0	0.0	0.324324	0.8	0.636357	0.666667	1.0	0.0	0.569654
3	0.698	0.0	0.0	0.283784	0.1	0.000000	0.333333	0.0	0.0	0.469120
4	1.000	1.0	0.0	0.337838	0.2	0.500246	0.000000	1.0	1.0	0.395400

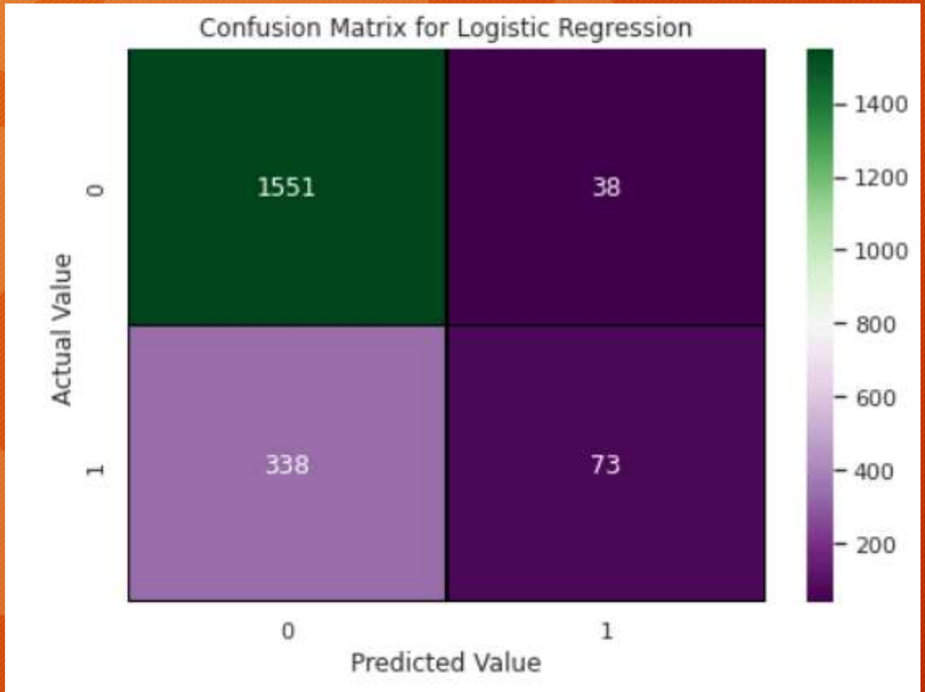
Classification Models for Prediction

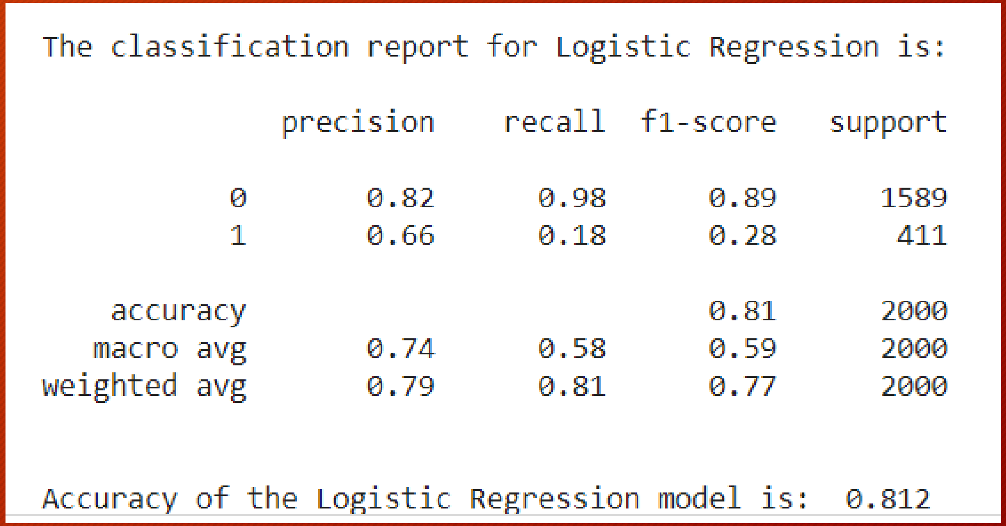
- For this Churn Modelling Prediction, the list of models being used for classification are as follows:

❖ Logistic Regression Model
❖ Decision Tree Model
❖ Random Forests Model
❖ K-Nearest Neighbours Model
❖ Support Vector Machine Model
❖ Naïve Bayesian Model
❖ Artificial Neural Networks Model

Logistic Regression Model

This is the confusion matrix plot and classification report for the Logistic Regression Model, which gives an accuracy of 81.2%.

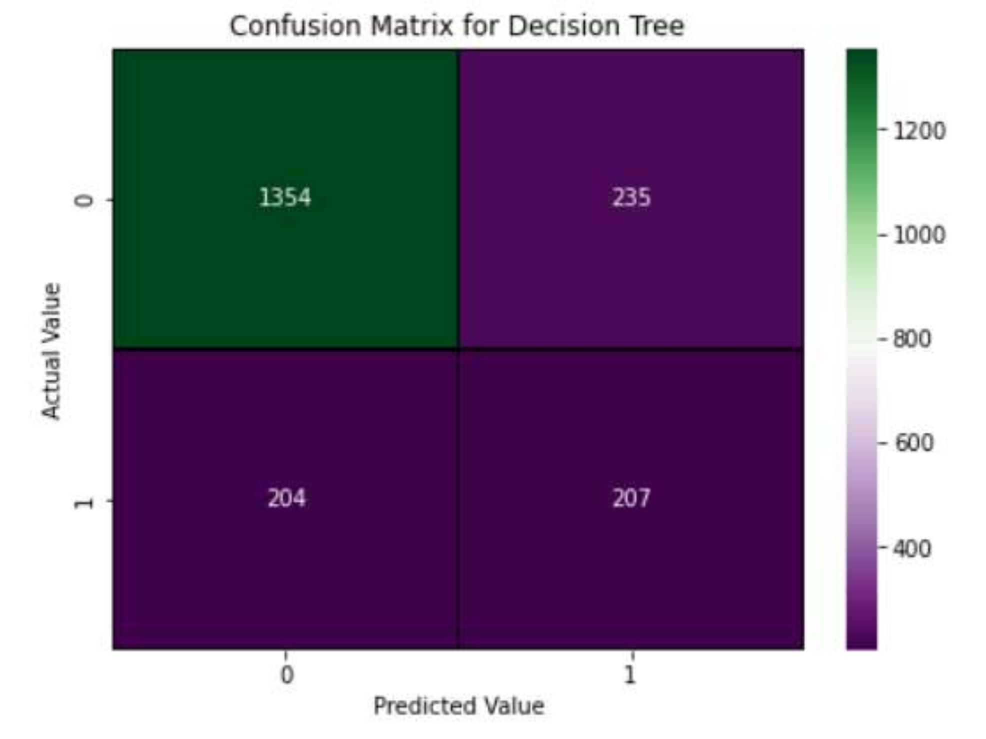


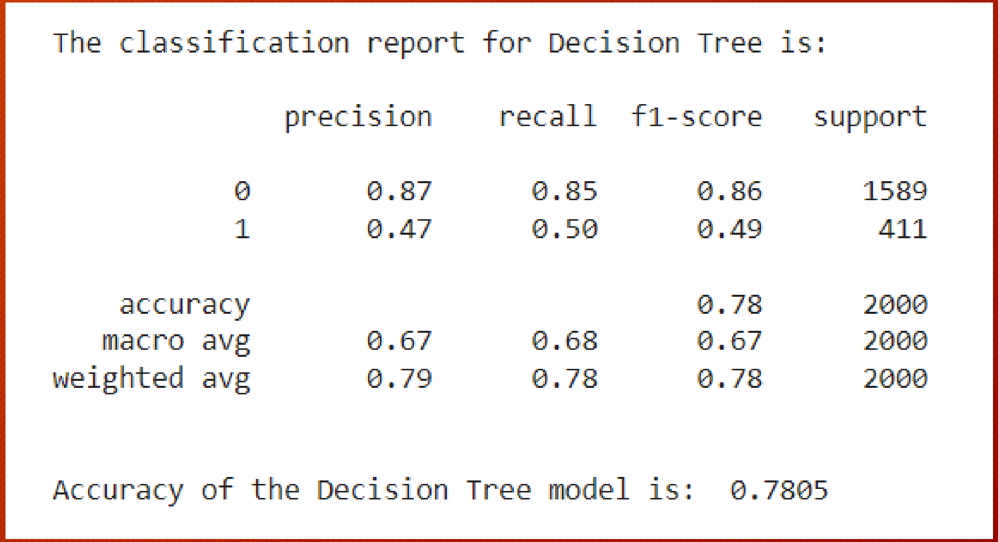


```
The classification report for Logistic Regression is:

              precision    recall  f1-score   support

           0       0.82      0.98      0.89      1589
           1       0.66      0.18      0.28       411

    accuracy                           0.81      2000
   macro avg       0.74      0.58      0.59      2000
weighted avg       0.79      0.81      0.77      2000


Accuracy of the Logistic Regression model is:  0.812
```

Decision Tree Model

This is the confusion matrix plot and classification report for the Decision Tree Model, which gives an accuracy of 78.5%.

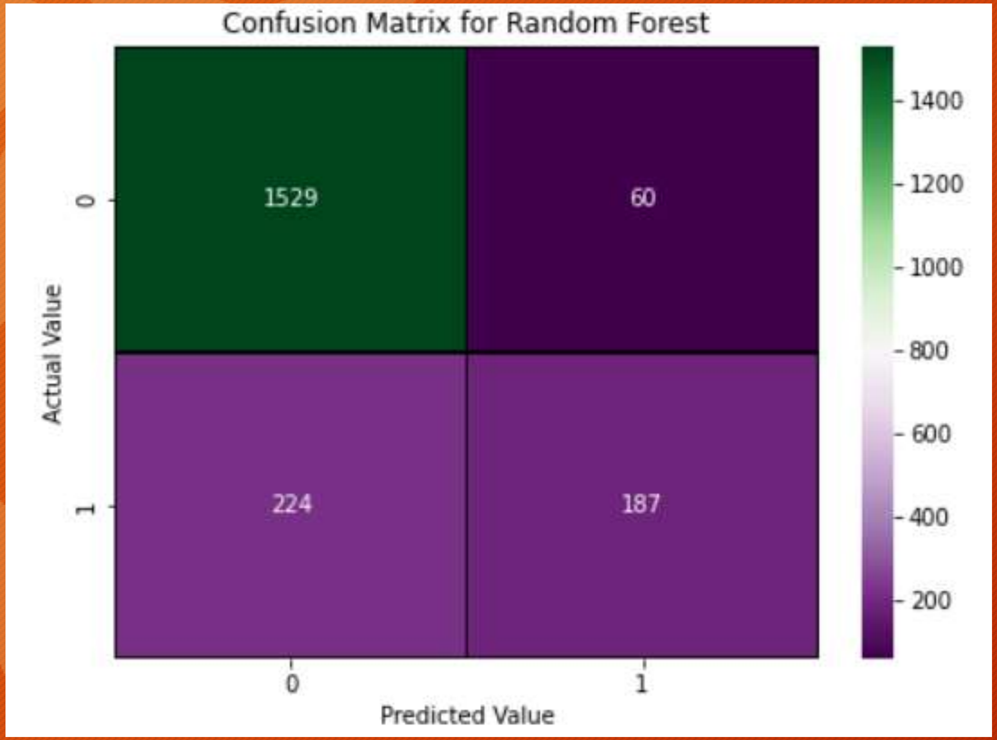


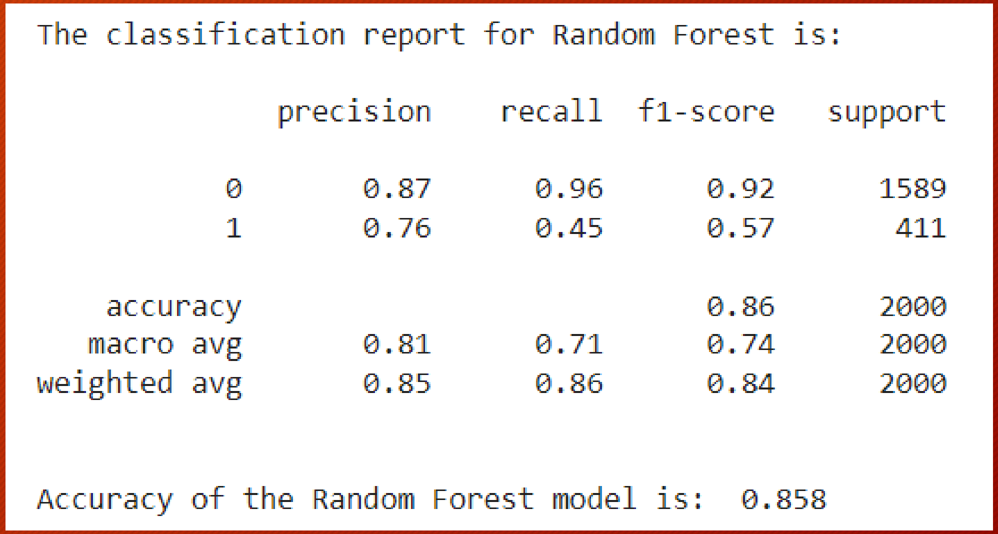


```
The classification report for Decision Tree is:

              precision    recall  f1-score   support

           0       0.87      0.85      0.86      1589
           1       0.47      0.50      0.49       411

    accuracy                           0.78      2000
   macro avg       0.67      0.68      0.67      2000
weighted avg       0.79      0.78      0.78      2000


Accuracy of the Decision Tree model is:  0.7805
```

Random Forest Model

This is the confusion matrix plot and classification report for the Random Forest Model, which gives an accuracy of 85.8%.

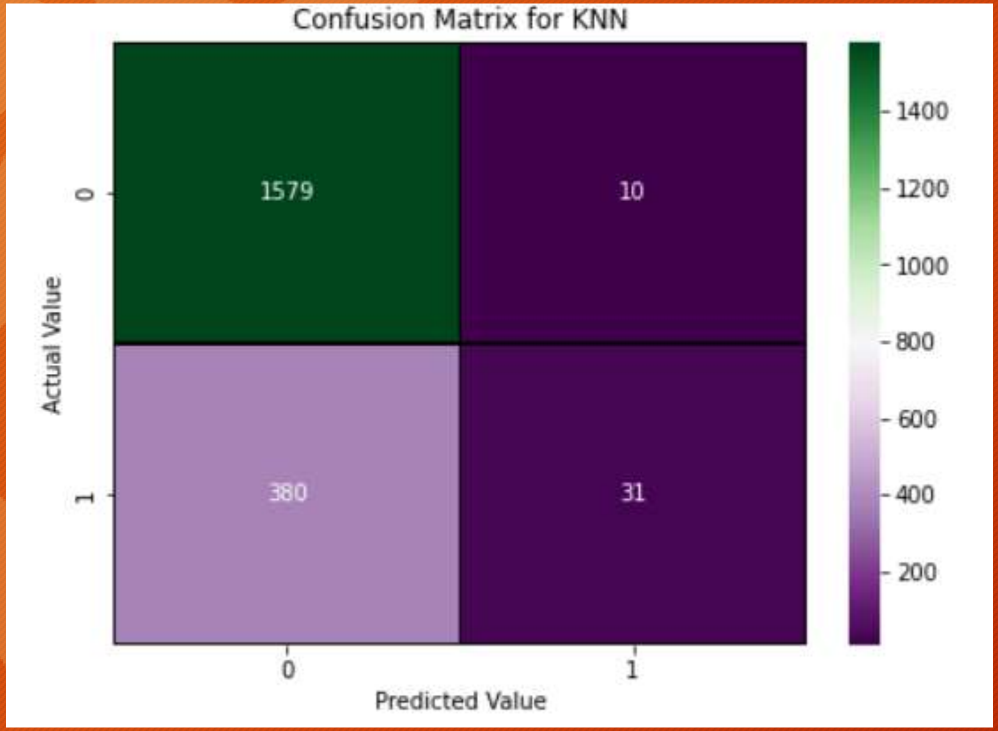


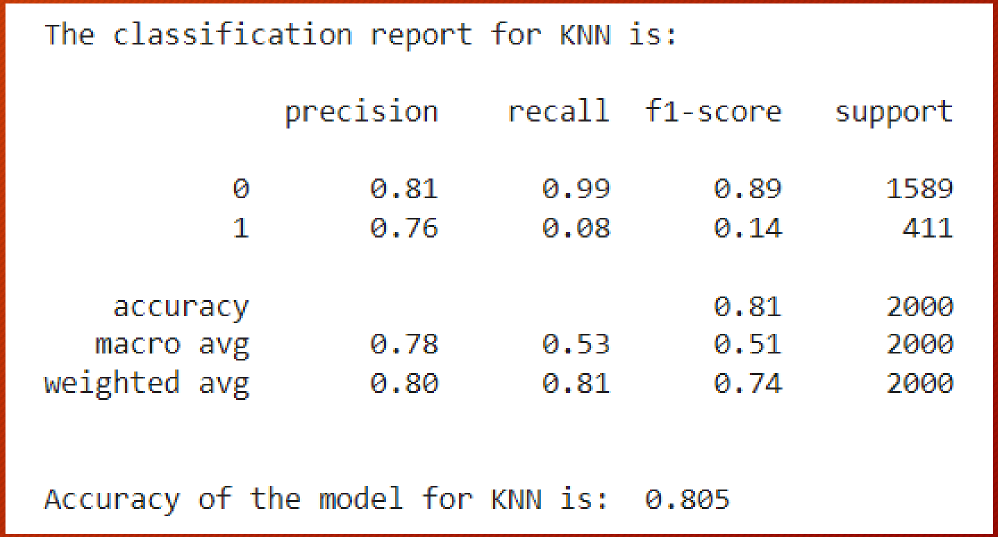


```
The classification report for Random Forest is:

              precision    recall  f1-score   support

           0       0.87      0.96      0.92      1589
           1       0.76      0.45      0.57       411

    accuracy                           0.86      2000
   macro avg       0.81      0.71      0.74      2000
weighted avg       0.85      0.86      0.84      2000


Accuracy of the Random Forest model is:  0.858
```

K-Nearest Neighbours Model

This is the confusion matrix plot and classification report for the K-Nearest Neighbours Model, which gives an accuracy of 80.5%.

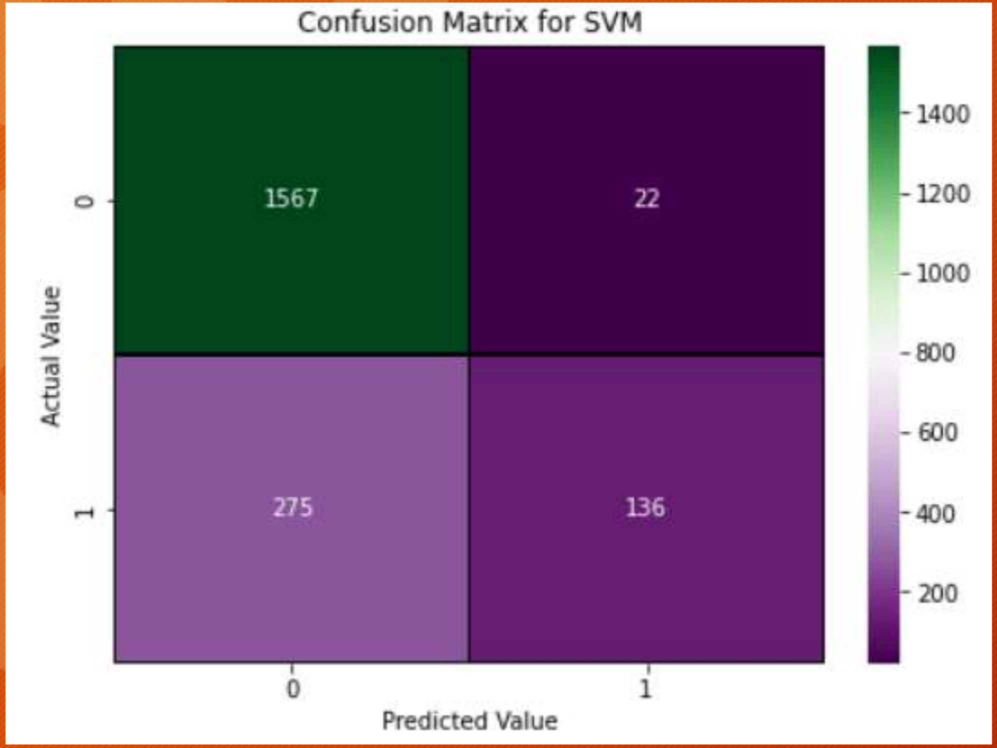


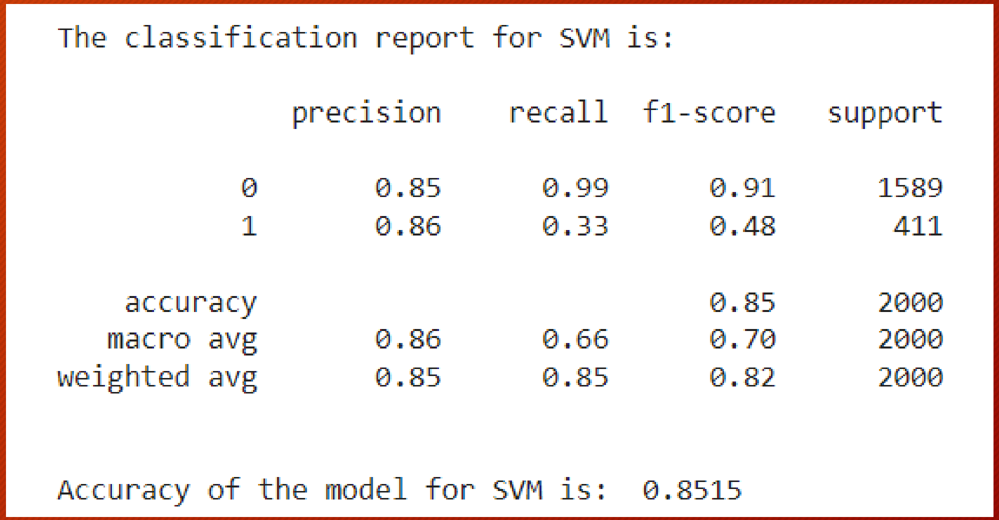


```
The classification report for KNN is:

              precision    recall  f1-score   support

           0       0.81      0.99      0.89      1589
           1       0.76      0.08      0.14       411

    accuracy                           0.81      2000
   macro avg       0.78      0.53      0.51      2000
weighted avg       0.80      0.81      0.74      2000


Accuracy of the model for KNN is:  0.805
```

Support Vector Machine Model

This is the confusion matrix plot and classification report for the Support Vector Machine Model, which gives an accuracy of 85.15%.

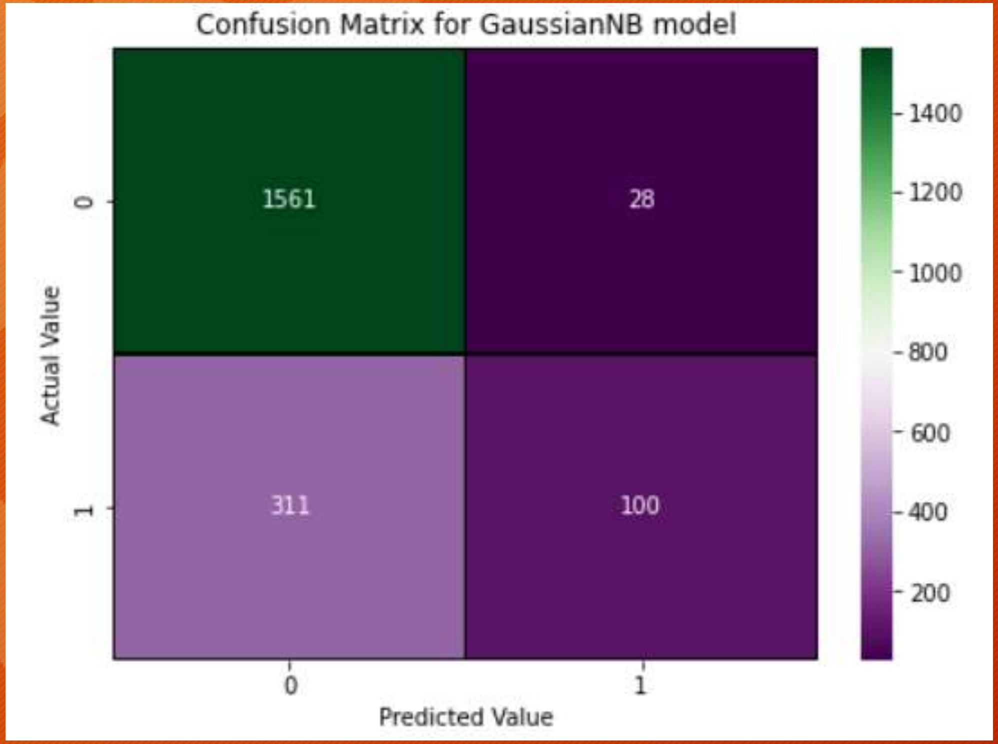


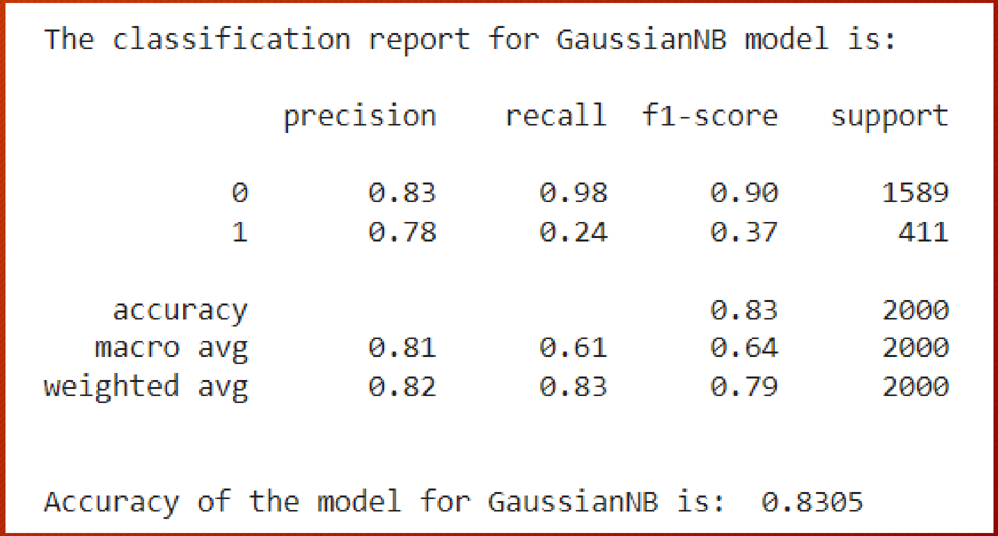


```
The classification report for SVM is:

              precision    recall  f1-score   support

           0       0.85      0.99      0.91      1589
           1       0.86      0.33      0.48       411

    accuracy                           0.85      2000
   macro avg       0.86      0.66      0.70      2000
weighted avg       0.85      0.85      0.82      2000


Accuracy of the model for SVM is:  0.8515
```

Naïve Bayesian Model

This is the confusion matrix plot and classification report for the Gaussian Naïve Bayesian Model, which gives an accuracy of 83.05%.

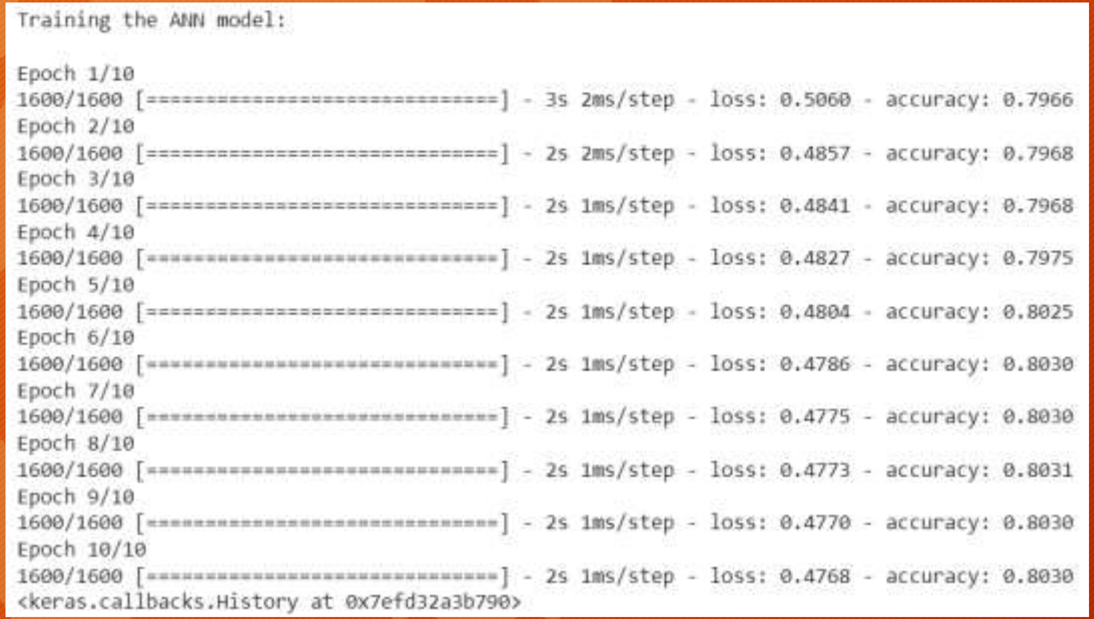




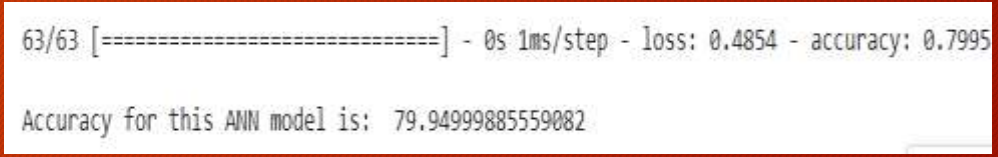
```
The classification report for GaussianNB model is:

              precision    recall  f1-score   support

           0       0.83      0.98      0.90      1589
           1       0.78      0.24      0.37       411

    accuracy                           0.83      2000
   macro avg       0.81      0.61      0.64      2000
weighted avg       0.82      0.83      0.79      2000


Accuracy of the model for GaussianNB is:  0.8305
```

Artificial Neural Networks Model

This is the training and accuracy for the Artificial Neural Networks Model, which gives an accuracy of 79.94%.



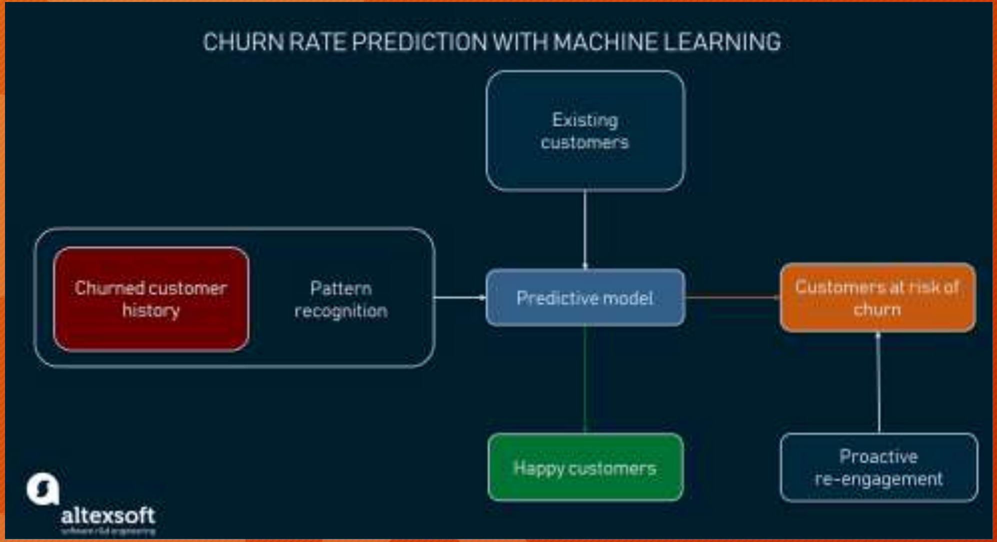
```
Training the ANN model:

Epoch 1/10
1600/1600 [==============================] - 3s 2ms/step - loss: 0.5060 - accuracy: 0.7966
Epoch 2/10
1600/1600 [==============================] - 2s 2ms/step - loss: 0.4857 - accuracy: 0.7968
Epoch 3/10
1600/1600 [==============================] - 2s 1ms/step - loss: 0.4841 - accuracy: 0.7968
Epoch 4/10
1600/1600 [==============================] - 2s 1ms/step - loss: 0.4827 - accuracy: 0.7975
Epoch 5/10
1600/1600 [==============================] - 2s 1ms/step - loss: 0.4804 - accuracy: 0.8025
Epoch 6/10
1600/1600 [==============================] - 2s 1ms/step - loss: 0.4786 - accuracy: 0.8030
Epoch 7/10
1600/1600 [==============================] - 2s 1ms/step - loss: 0.4775 - accuracy: 0.8030
Epoch 8/10
1600/1600 [==============================] - 2s 1ms/step - loss: 0.4773 - accuracy: 0.8031
Epoch 9/10
1600/1600 [==============================] - 2s 1ms/step - loss: 0.4770 - accuracy: 0.8030
Epoch 10/10
1600/1600 [==============================] - 2s 1ms/step - loss: 0.4768 - accuracy: 0.8030
<keras.callbacks.History at 0x7efd32a3b790>
```

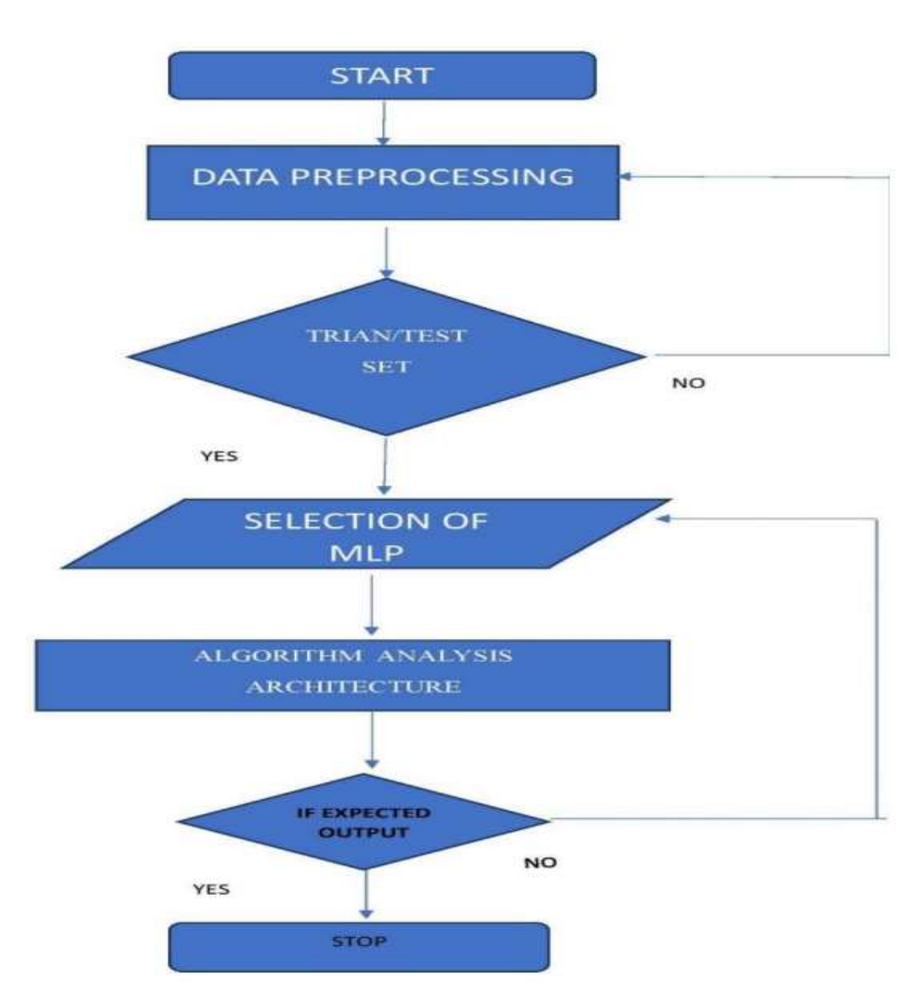


```
63/63 [==============================] - 0s 1ms/step - loss: 0.4854 - accuracy: 0.7995

Accuracy for this ANN model is:  79.94999885559082
```

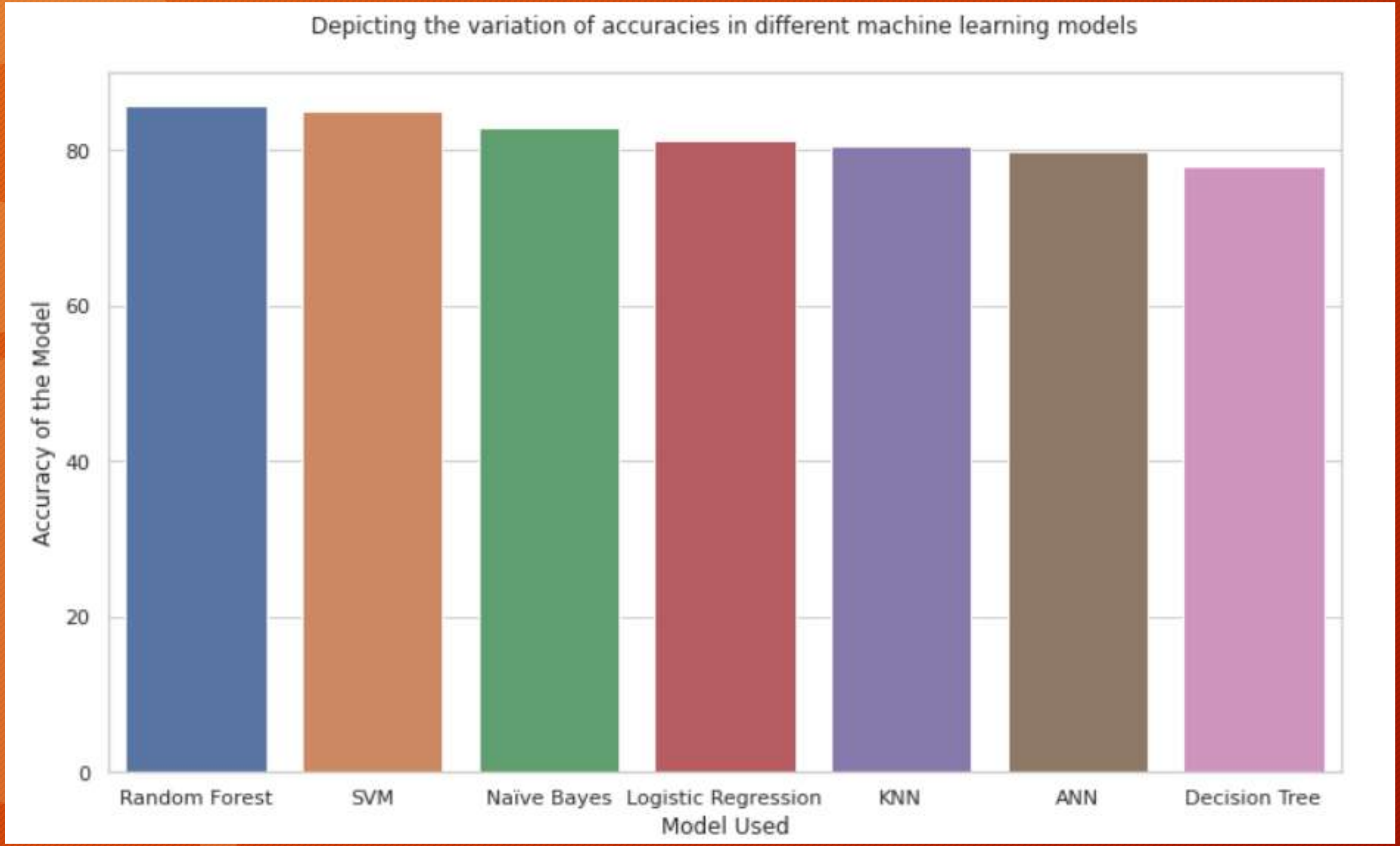
Contribution

- Data Flow Diagram



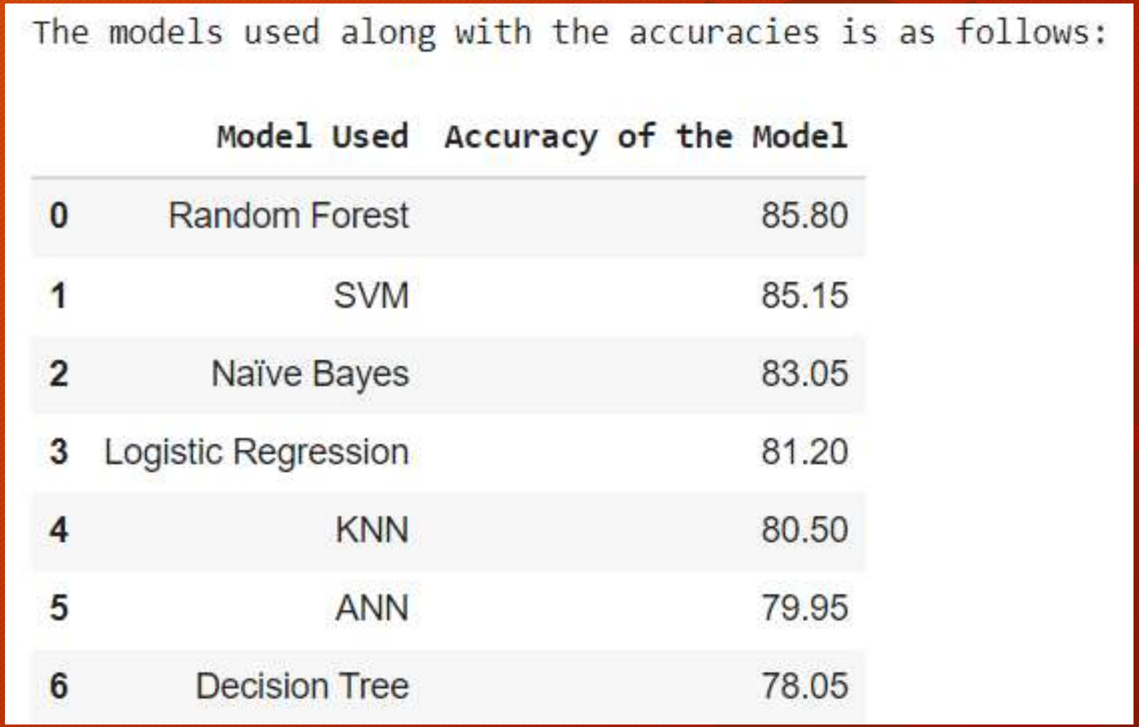


Visualizing the accuracies of Models



Final Research Findings

From the Data Frame on the right hand side, we can clearly see that the Random Forest Model gives the highest accuracy amongst the 7 different models. Then comes the SVM Model with an accuracy close to that of Random Forest and then the Naïve Bayesian Model. Also, the model giving the least accuracies are Decision Tree Model and Artificial Neural Networks Model.



```
The models used along with the accuracies is as follows:
```

	Model Used	Accuracy of the Model
0	Random Forest	85.80
1	SVM	85.15
2	Naïve Bayes	83.05
3	Logistic Regression	81.20
4	KNN	80.50
5	ANN	79.95
6	Decision Tree	78.05

Conclusion

So, from the above results we can come to a conclusion that for the given Churn Modelling dataset, Random Forest Model performs very well when compared to the other machine learning models. The reason for this is that it utilizes ensemble learning method for prediction and it selects random sample of training data and uses many single decision trees and considers the node values receiving the most votes among the many single decision trees. So, we get very accurate node values at each step. Also, it is resistant to overfitting and pruning is not necessary for it, and also each decision tree is independent so they can grow in different cores and computers for faster analysis. The founder of Random Forest algorithm Leo Breiman also suggests that they perform very well with large datasets having moderate number of columns.

References

- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
- https://scikit-learn.org/stable/modules/naive_bayes.html
- https://keras.io/guides/sequential_model/

End of Presentation

Thank You