

**FACULTY OF ENGINEERING TECHNOLOGY**

**COMPUTER SCIENCE DEPARTMENT**

**COMP4388, MACHINE LEARNING**

**ASSIGNMENT 1**

**Prepared by:**

Zaid Nedal Mousa 1221833

**Instructor:** Dr. Radi Jarrar

Contents

[Task 1 4](#_Toc185870244)

[1. EDA 4](#_Toc185870245)

[2. Distribution for class label (churn) 5](#_Toc185870246)

[3. Amount of churn in each sub-group in age group 6](#_Toc185870247)

[4. Amount of Churn in each sub-group of Charge Amount 7](#_Toc185870248)

[5. Details of charge amount 8](#_Toc185870249)

[6. Correlation between features 9](#_Toc185870250)

[7. Split Data (Training and Test) 10](#_Toc185870251)

[Regression Task 11](#_Toc185870252)

[1. Linear regression using all independent attributes 11](#_Toc185870253)

[2. Linear regression using 2 most important features 12](#_Toc185870254)

[3. Linear regression using set of important features 13](#_Toc185870255)

[4. Compare between the performance for the 14](#_Toc185870256)

[Classification Task 15](#_Toc185870257)

[1. KNN 15](#_Toc185870258)

[2. Naive Bayes 15](#_Toc185870259)

[3. Decision Tree 15](#_Toc185870260)

[4. ROC/AUC and Compare the performance between classification 16](#_Toc185870261)

[ROC/AUC 17](#_Toc185870262)

[Confusion Matrix 18](#_Toc185870263)

[Naive Bayes 18](#_Toc185870264)

[Decision Tree 19](#_Toc185870265)

[KNN 20](#_Toc185870266)

[Logistic Regression 21](#_Toc185870267)

Table of Figures

[Figure 1: EDA code 4](#_Toc185870134)

[Figure2:EDA statistics 4](#_Toc185870135)

[Figure 3: churn distribution code 5](#_Toc185870136)

[Figure 4: Distribution on Terminal 5](#_Toc185870137)

[Figure 5: Distribution plot 5](#_Toc185870138)

[Figure 6:Histogram details change code 6](#_Toc185870139)

[Figure 7:Histogram for details churn with age group 6](#_Toc185870140)

[Figure 8:Histogram details charge amount and churn code 7](#_Toc185870141)

[Figure 9:Histogram for details churn with charge amount 7](#_Toc185870142)

[Figure 10: charge amount details code 8](#_Toc185870143)

[Figure 11: charge amount details 8](#_Toc185870144)

[Figure 12: correlation features code 9](#_Toc185870145)

[Figure 13: correlation matrix 9](#_Toc185870146)

[Figure 14:split data code 10](#_Toc185870147)

[Figure 15:LRM1 Regression code 11](#_Toc185870148)

[Figure 16: LRM1 performance 11](#_Toc185870149)

[Figure 17: LRM2 Code 12](#_Toc185870150)

[Figure 18: LRM2 performance 12](#_Toc185870151)

[Figure 19: LRM3 regression code 13](#_Toc185870152)

[Figure 20: LRM3 performance 13](#_Toc185870153)

[Figure 21: Comparing 14](#_Toc185870154)

[Figure 22: KNN classification 15](#_Toc185870155)

[Figure 23: Naive Bayes classification 15](#_Toc185870156)

[Figure 24: Decision Tree classification 15](#_Toc185870157)

[Figure 25: Logistic code 16](#_Toc185870158)

[Figure 26: comparing performance code 16](#_Toc185870159)

[Figure 27: performance between classification models 16](#_Toc185870160)

[Figure 28:ROC/AUC code 17](#_Toc185870161)

[Figure 29: ROC/AUC figure 17](#_Toc185870162)

[Figure 30: Confusion matrix Naive Bayes code 18](#_Toc185870163)

[Figure 31: Confusion matrix Naive Bayes 18](#_Toc185870164)

[Figure 32: decision tree matrix code 19](#_Toc185870165)

[Figure 33:Decision tree matrix 19](#_Toc185870166)

[Figure 34:KNN confusion matrix code 20](#_Toc185870167)

[Figure 35: KNN confusion matrix 20](#_Toc185870168)

[Figure 36: Logistic regression confusion matrix code 21](#_Toc185870169)

[Figure 37:Logistic regression Confusion matrix 21](#_Toc185870170)

# Task 1

## EDA

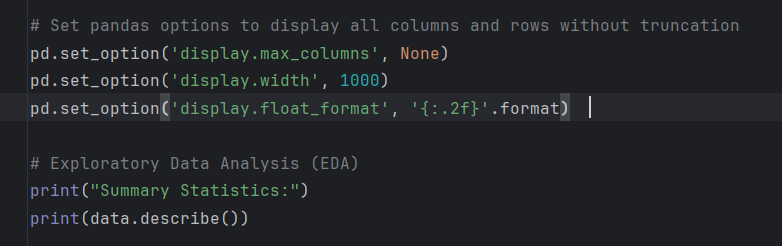


Figure : EDA code

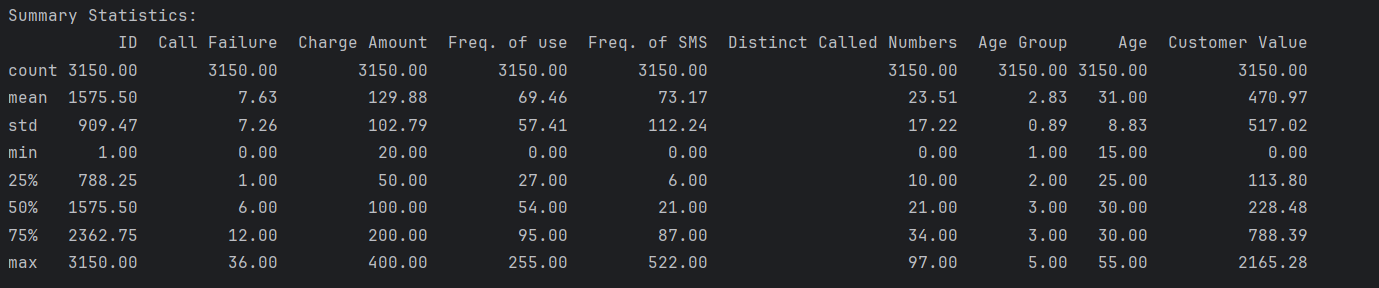


Figure:EDA statistics

This summarize the data and describe it and hive some information about the(count of them, mean, Standard Deviation, min and max, and the quartiles)

## Distribution for class label (churn)

A computer screen shot of text

Description automatically generated

Figure : churn distribution code

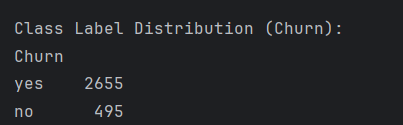


Figure : Distribution on Terminal

A blue rectangular bar graph

Description automatically generated

Figure : Distribution plot

These shows the Distribution for the class label **Churn**

## Amount of churn in each sub-group in age group

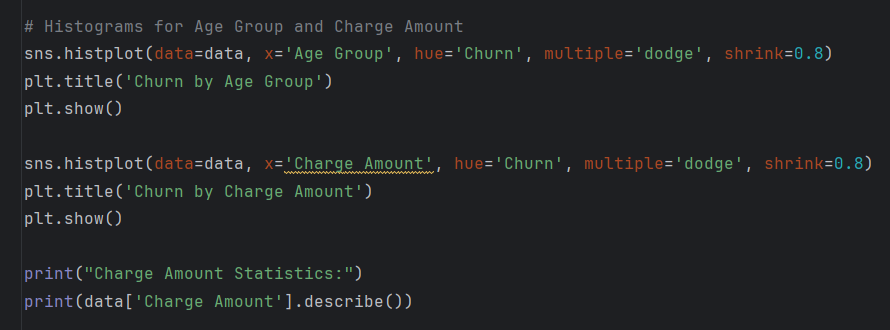


Figure :Histogram details change code

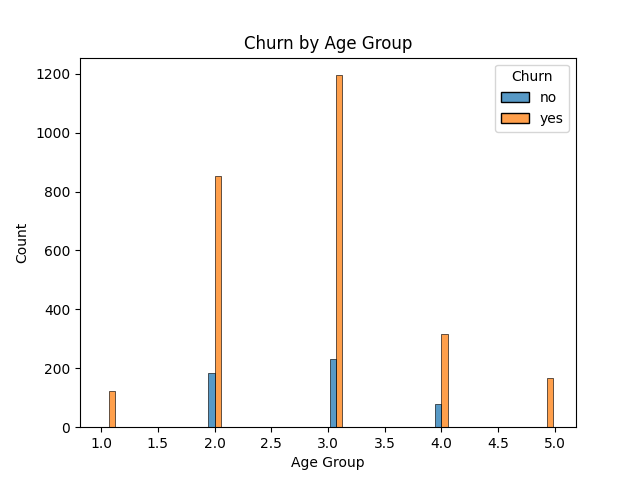


Figure :Histogram for details churn with age group

## Amount of Churn in each sub-group of Charge Amount

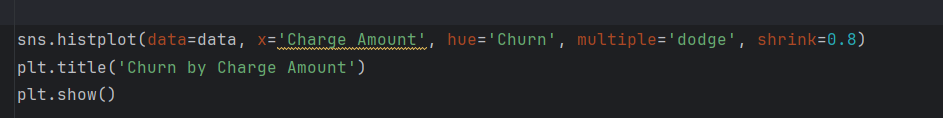


Figure :Histogram details charge amount and churn code

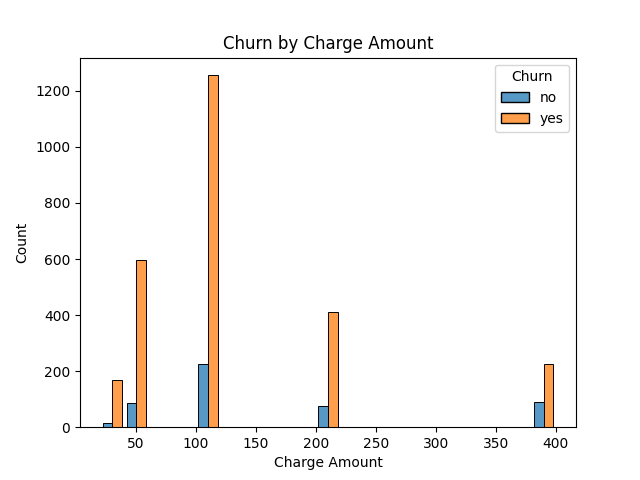


Figure :Histogram for details churn with charge amount

## Details of charge amount

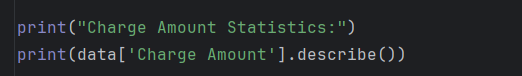


Figure : charge amount details code

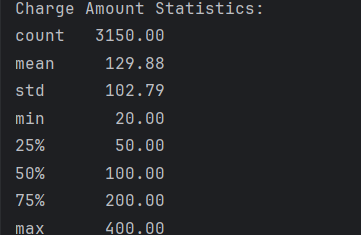


Figure : charge amount details

## Correlation between features

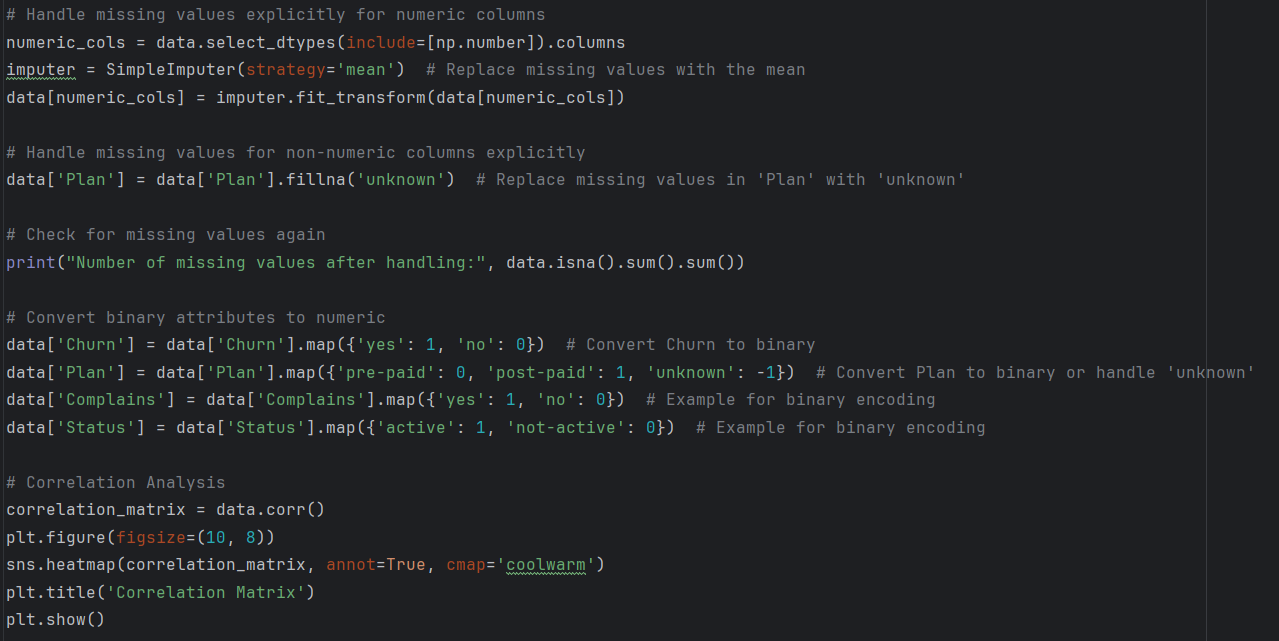


Figure : correlation features code

A screenshot of a graph

Description automatically generated

Figure : correlation matrix

From the matrix, it is clear that Customer Value, Freq. of use, and Complains have significant relationships with other features and are particularly relevant for predicting churn. These features should be prioritized when developing a churn prediction model.

On the other hand, features with weak correlations, such as Plan, Status, and Call Failure, may not contribute much to the prediction and might either require transformation to improve their relevance or could be excluded altogether to simplify the model and enhance its performance. This approach ensures that the focus remains on the most impactful variables.

## Split Data (Training and Test)

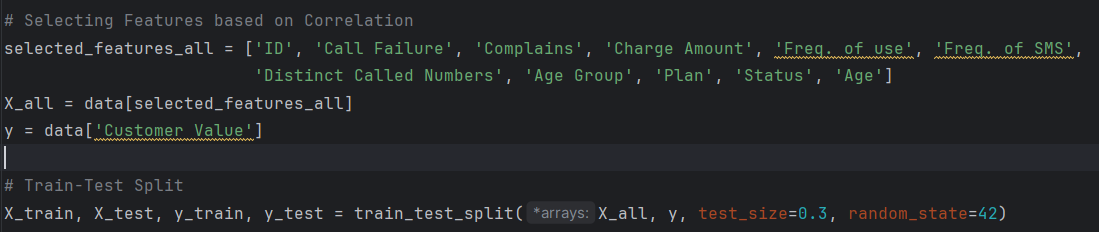


Figure :split data code

Here we just split the data to: Training data 70% and Test Data 30%

# Regression Task

## Linear regression using all independent attributes

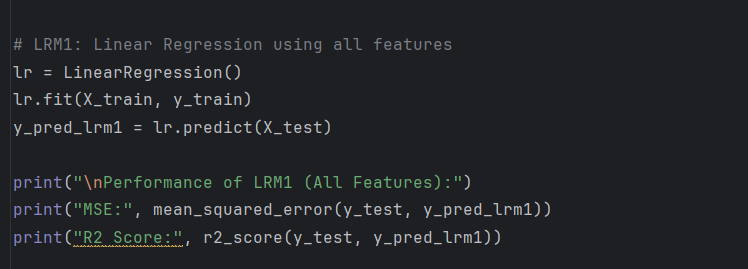


Figure :LRM1 Regression code

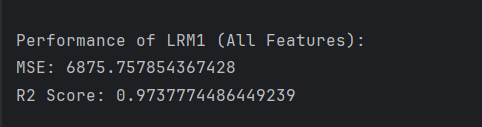


Figure : LRM1 performance

## Linear regression using 2 most important features

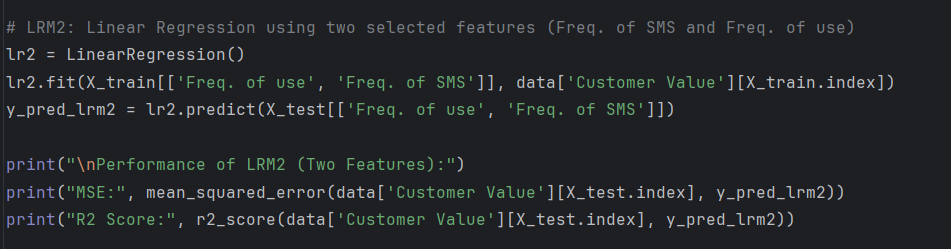


Figure : LRM2 Code

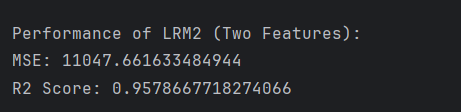


Figure : LRM2 performance

We selected **Freq. of Use** and **Freq. of SMS** for LRM2 because they reflect customer engagement with the service, which directly influences customer value. High usage typically indicates a more valuable customer.

**Freq. of Use** reflects customer engagement and activity levels. Customers who frequently use the service are more likely to generate higher value. This engagement often translates to higher spending or longer retention.

**Freq. of SMS**: this feature complements it by capturing another dimension of service usage. A high number of SMS messages indicates frequent interaction with the telecommunication service, which may correlate with higher customer value.

## Linear regression using set of important features

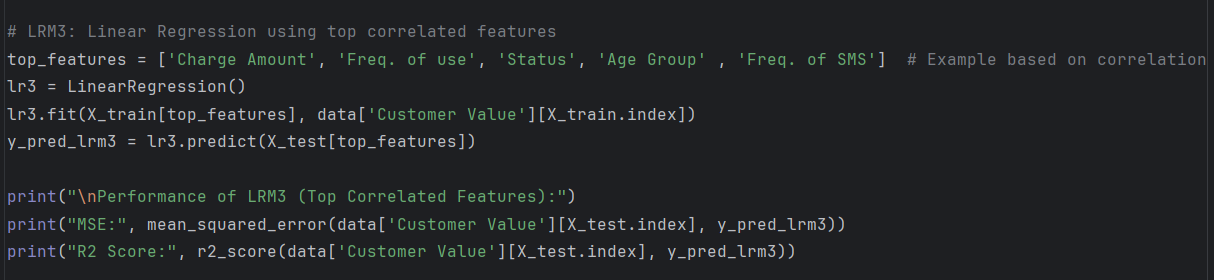


Figure : LRM3 regression code

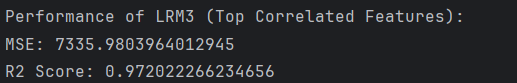


Figure : LRM3 performance

We chose features based on their strong correlations with customer value, including **Charge Amount** (financial contribution), **Freq. of Use** and **Freq. of SMS** (engagement), **Status** (active or inactive), and **Age Group** (demographic trends). These features were chosen as key drivers of customer behavior and value, combining both domain insights and statistical significance.

**Charge Amount** reflects the financial contribution of the customer. Customers with higher charges are generally more valuable and engaged.

**Freq. of Use**: The correlation matrix may confirm that frequent usage of calls correlates strongly with customer value.

**Status**: Active customers are more likely to generate value for the company, while inactive ones contribute little to no revenue. This is an important predictor of churn and overall customer value.

**Age Group**: Customer value often varies by age. Younger customers may use the service differently than older ones, impacting their overall contribution to the business.

**Freq. of SMS**: The correlation matrix likely confirms that frequent SMS usage correlates strongly with customer value.

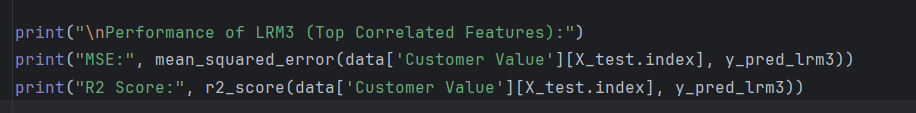
## Compare between the performance for the

A screen shot of a computer program

Description automatically generated

A screen shot of a computer code

Description automatically generated



A screenshot of a computer program

Description automatically generated

Figure : Comparing

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Features Used | MSE | R² Score |
| LRM1 | All Independent Attributes | 6875.75 | 0.973777 |
| LRM2 | Freq. of use, Freq. of SMS | 11047.66 | 0.957866 |
| LRM3 | Charge Amount, Freq. of use, Status, Age Group, Freq. of SMS | 7335.98 | 0.972022 |

# Classification Task

## KNN

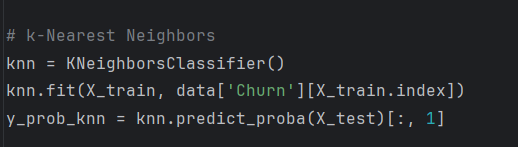


Figure : KNN classification

## Naive Bayes

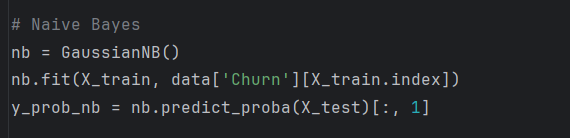


Figure : Naive Bayes classification

## Decision Tree

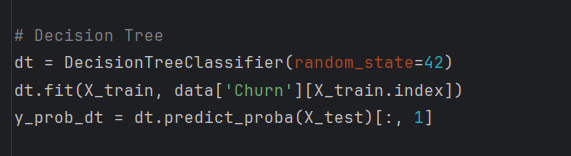


Figure : Decision Tree classification

## ROC/AUC and Compare the performance between classification

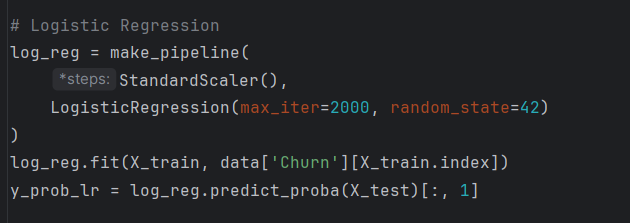


Figure : Logistic code

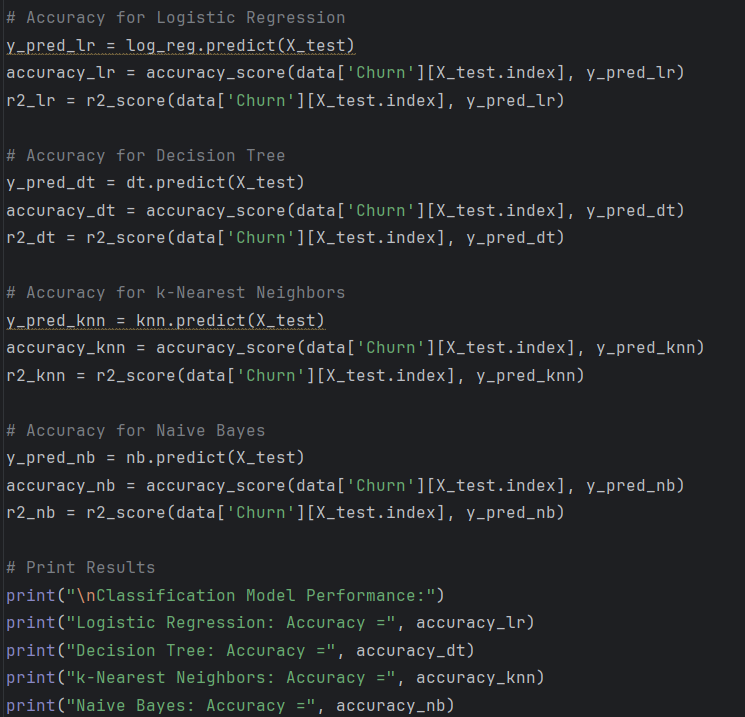


Figure : comparing performance code

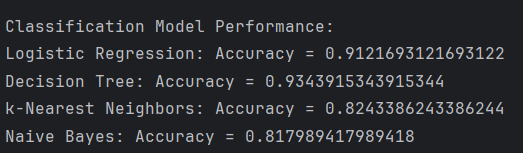


Figure : performance between classification models

## ROC/AUC

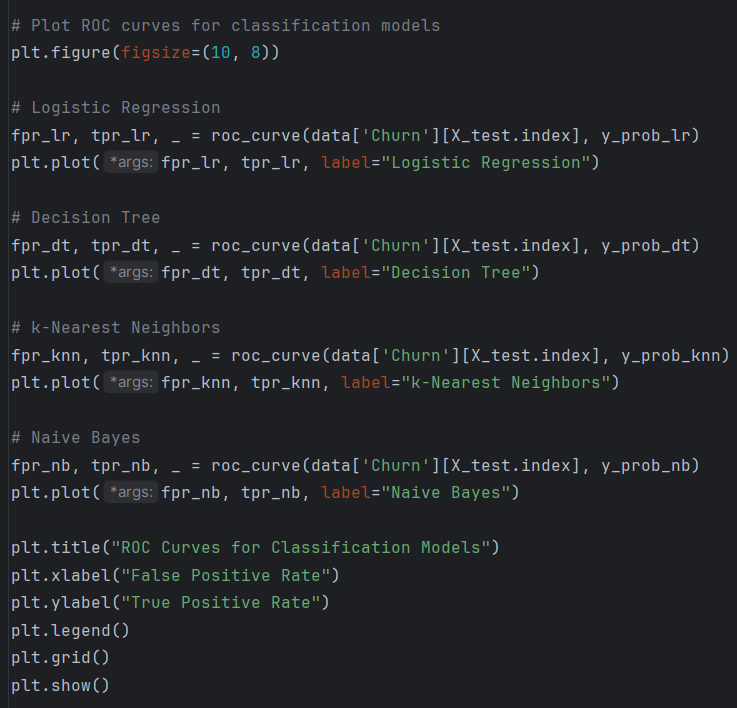


Figure :ROC/AUC code

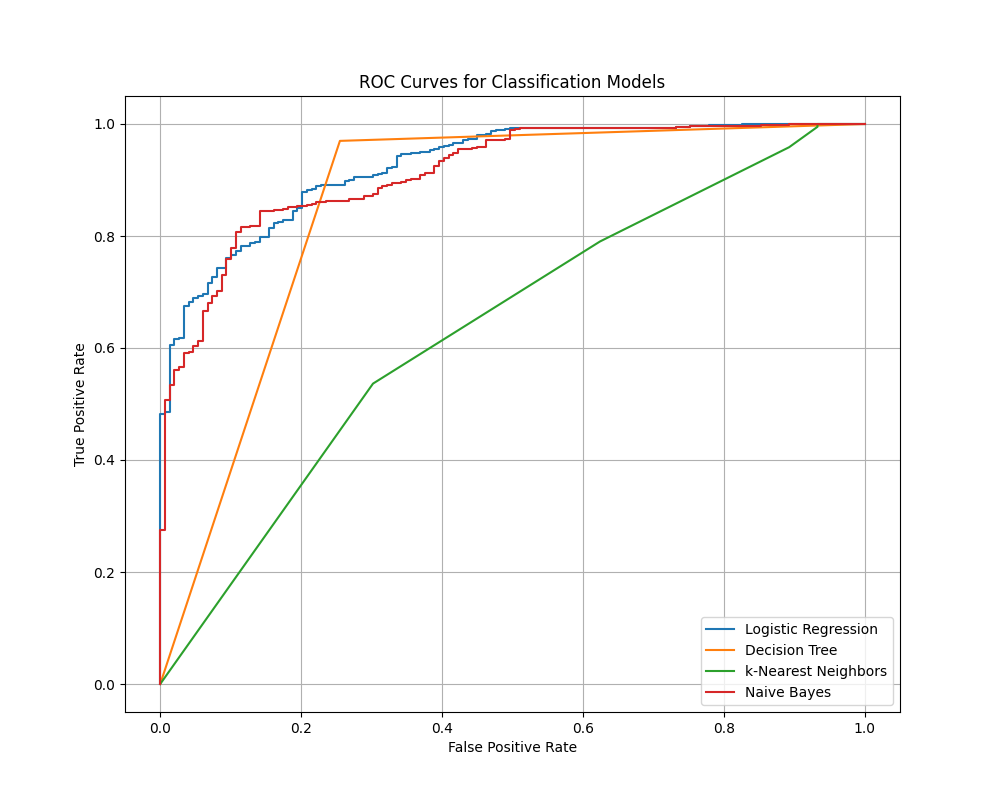


Figure : ROC/AUC figure

## Confusion Matrix

### Naive Bayes

A computer screen shot of text

Description automatically generated

Figure : Confusion matrix Naive Bayes code

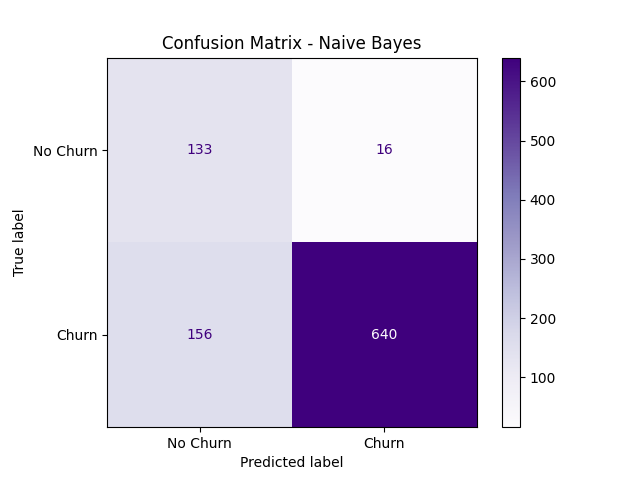


Figure : Confusion matrix Naive Bayes

### Decision Tree

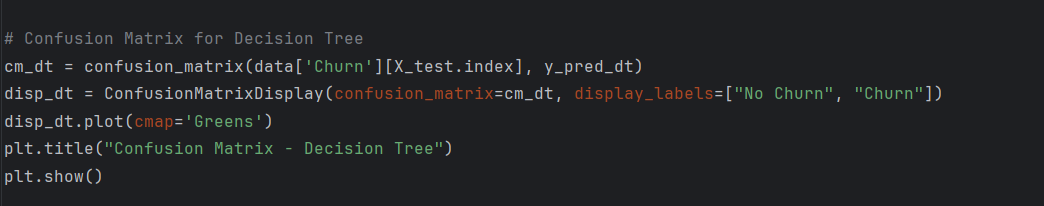


Figure : decision tree matrix code

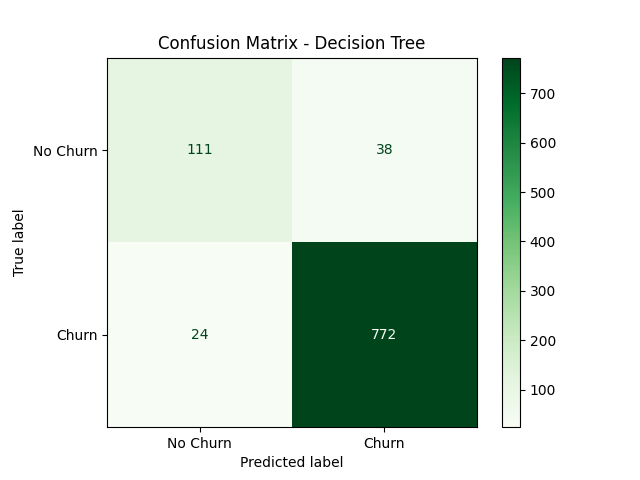


Figure :Decision tree matrix

### KNN

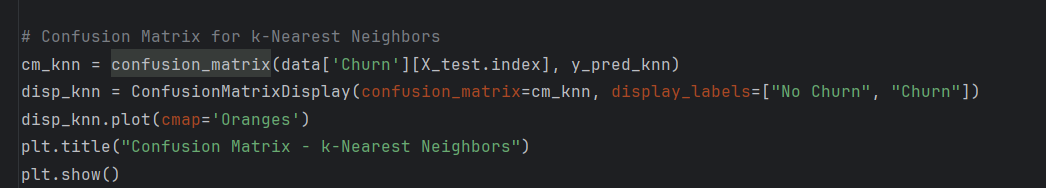


Figure :KNN confusion matrix code

A diagram of a number of individuals

Description automatically generated with medium confidence

Figure : KNN confusion matrix

### Logistic Regression

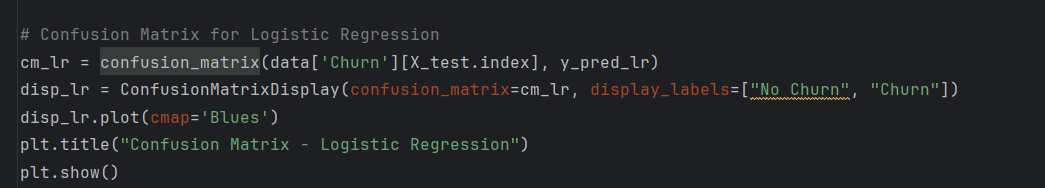


Figure : Logistic regression confusion matrix code

A diagram of a logistic regression

Description automatically generated

Figure :Logistic regression Confusion matrix

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Comments |
| |  | | --- | | Logistic Regression |  |  | | --- | |  | | |  | | --- | | 0.912 |  |  | | --- | |  | | |  | | --- | | Consistently high accuracy with good generalization capabilities. |  |  | | --- | |  | |
| |  | | --- | | Decision Tree |  |  | | --- | |  | | |  | | --- | | 0.934 |  |  | | --- | |  | | |  | | --- | | The highest accuracy but could be prone to overfitting. |  |  | | --- | |  | |
| k-Nearest Neighbors | |  | | --- | | 0.824 |  |  | | --- | |  | | |  | | --- | | Moderate performance; sensitive to noisy data and scaling. |  |  | | --- | |  | |
| Naive Bayes | |  | | --- | | 0.818 |  |  | | --- | |  | | Lowest accuracy; limited by its assumption of feature independence. |

**Decision Tree (Accuracy = 0.934):**

* **outperforms:**
  + Decision trees can capture complex patterns in the data, leading to the highest accuracy.
  + It handles categorical and numerical features well and does not require feature scaling.

While the high accuracy indicates strong performance, Decision Trees are prone to overfitting, especially if the tree grows too deep. This can cause poor generalization to unseen data unless techniques such as pruning, cross-validation, or ensemble methods (e.g., Random Forest) are applied.

**Logistic Regression (Accuracy = 0.912)**:

Logistic Regression provides robust performance when the dataset has a linear relationship between features and the target variable. The use of feature scaling (via StandardScaler) further improves its efficiency and accuracy.

It is less complex and less prone to overfitting than a Decision Tree. However, it struggles to capture non-linear relationships, which can limit its applicability to datasets with intricate patterns.

**k-Nearest Neighbors (Accuracy = 0.824)**:

kNN reliance on the proximity of data points makes it sensitive to noise and scaling. Additionally, its performance heavily depends on the choice of hyperparameters (k) and can degrade for larger datasets or those with overlapping classes.

kNN works well for smaller, well-separated datasets but is less effective for datasets with complex or noisy distributions.

**Naive Bayes (Accuracy = 0.818)**:

Naive Bayes assumes independence between features, which is rarely the case in real-world datasets. This oversimplification can lead to suboptimal performance when feature dependencies exist.

It is computationally efficient and suitable for high-dimensional datasets. Despite its limitations, it can serve as a baseline model for comparison.