Artificial Intelligence Project Report

Nawaz Uddin Tamim- 22241061

Department of Computer Science and Enginnering

Course: CSE422

Submission Date: 8th September, 2025

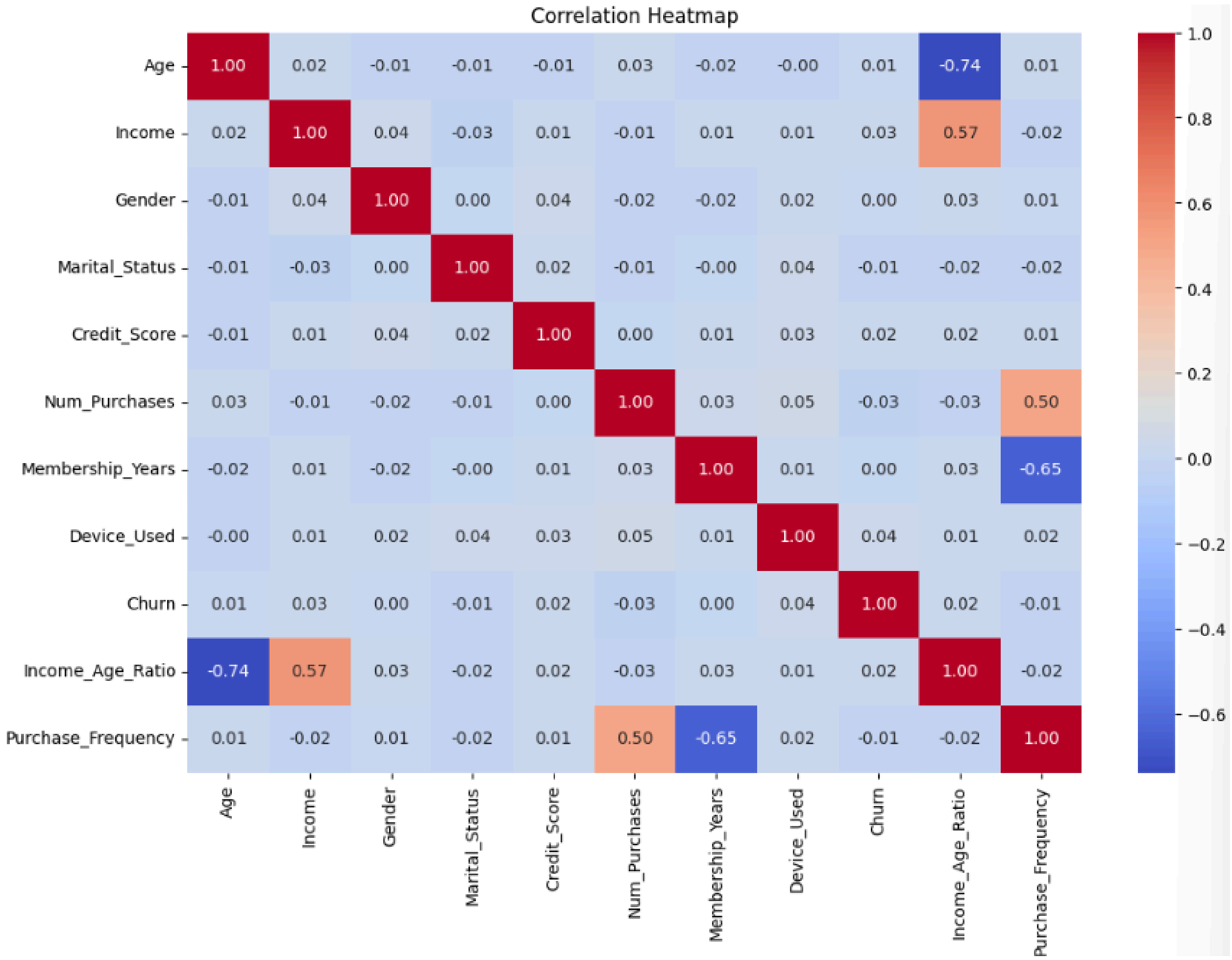
Table of Contents

1. Introduction:

This project focuses on predicting customer churn, a critical challenge that directly affects business revenue and growth. Using demographic, financial, and behavioral data such as age, income, credit score, marital status, and service usage, the goal is to identify patterns that indicate whether a customer is likely to leave. The motivation behind this work is to help organizations take proactive steps; such as targeted offers or service improvements to retain customers and reduce churn. To achieve this, multiple supervised learning models (KNN, Logistic Regression, Neural Networks) are applied, along with clustering (K-Means) for deeper insights. This study ultimately addresses a real-world classification problem with practical implications for customer retention strategies.

2. Dataset Description:

- ➔ The dataset used is a sample of the Consumer Classification Dataset, containing 13,500 datapoints and 8 input features.
- ➔ There are 8 features and 1 output feature.
- ➔ This is a binary classification problem. Because we are predicting two labels from the features.
- ➔ There are 13500 data points.
- ➔ The dataset contains both quantitative (e.g., Age, Income, Credit_Score, Num_Purchases, Membership_Years, Churn) and categorical (e.g., Gender, Marital_Status, Device_Used) variables.
- ➔ Since machine learning models require numerical input, categorical features such as *Gender, Marital_Status & Device_Used* were encoded using Label Encoding to convert them into numeric form.
- ➔ Below is a graph to visualize correlations between numeric features using a heatmap (Seaborn library).



	Age	Income	Gender	Marital_Status	Credit_Score	Num_Purchases	Membership_Years	Device_Used	Churn	Income_Age_Ratio	Purchase_Frequency
Age	1.00	0.02	-0.01	-0.01	-0.01	0.03	-0.02	-0.00	0.01	-0.74	0.01
Income	0.02	1.00	0.04	-0.03	0.01	-0.01	0.01	0.01	0.03	0.57	-0.02
Gender	-0.01	0.04	1.00	0.00	0.04	-0.02	-0.02	0.02	0.00	0.03	0.01
Marital_Status	-0.01	-0.03	0.00	1.00	0.02	-0.01	-0.00	0.04	-0.01	-0.02	-0.02
Credit_Score	-0.01	0.01	0.04	0.02	1.00	0.00	0.01	0.03	0.02	0.02	0.01
Num_Purchases	0.03	-0.01	-0.02	-0.01	0.00	1.00	0.03	0.05	-0.03	-0.03	0.50
Membership_Years	-0.02	0.01	-0.02	-0.00	0.01	0.03	1.00	0.01	0.00	0.03	-0.65
Device_Used	-0.00	0.01	0.02	0.04	0.03	0.05	0.01	1.00	0.04	0.01	0.02
Churn	0.01	0.03	0.00	-0.01	0.02	-0.03	0.00	0.04	1.00	0.02	-0.01
Income_Age_Ratio	-0.74	0.57	0.03	-0.02	0.02	-0.03	0.03	0.01	0.02	1.00	-0.02
Purchase_Frequency	0.01	-0.02	0.01	-0.02	0.01	0.50	-0.65	0.02	-0.01	-0.02	1.00

➔ From this heatmap, we can observe the following:

As we have created, Income_Age_Ratio by feature engineering has a strong negative correlation with Age and a moderate positive correlation with Income, which is expected given how this feature was engineered.

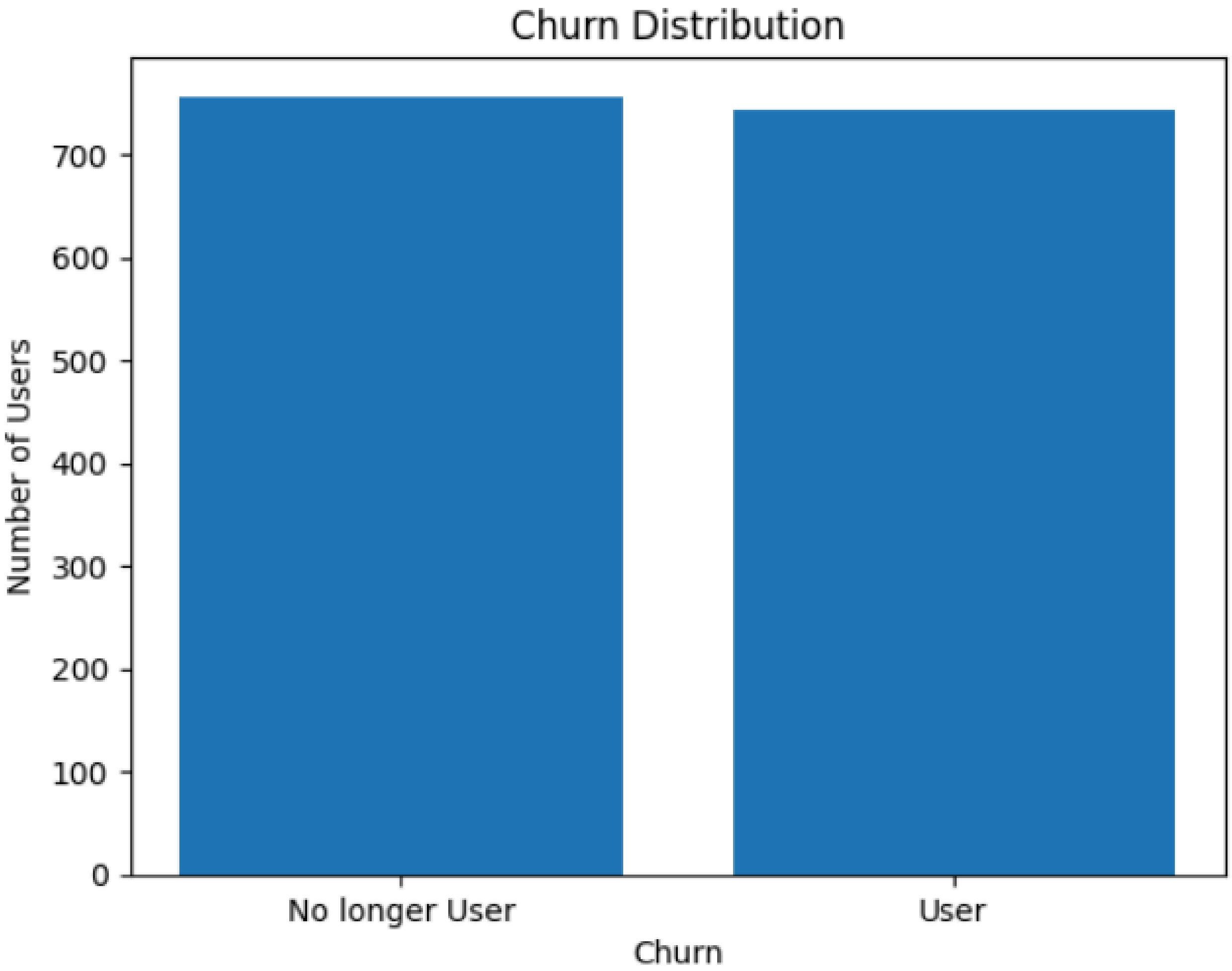
Purchase_Frequency has a moderate negative correlation with Membership_Years and a moderate positive correlation with Num_Purchases, which is also expected.

The correlations between the features and the target variable Churn are relatively weak, suggesting that no single feature has a strong linear relationship with churn.

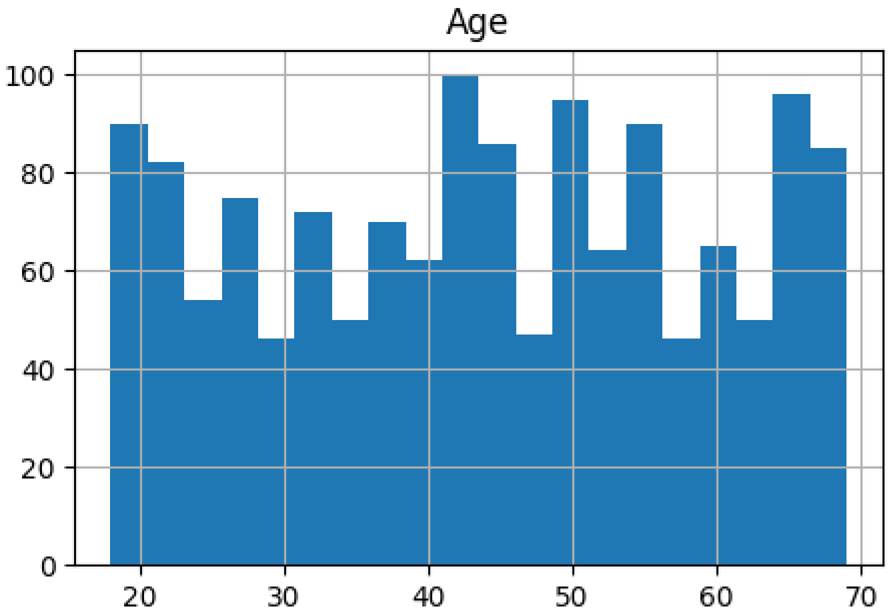
The correlations among most of the original features (Age, Income, Gender, Marital_Status, Credit_Score, Num_Purchases, Membership_Years, Device_Used) are generally low, indicating that they are relatively independent of each other.

Imbalance Dataset:

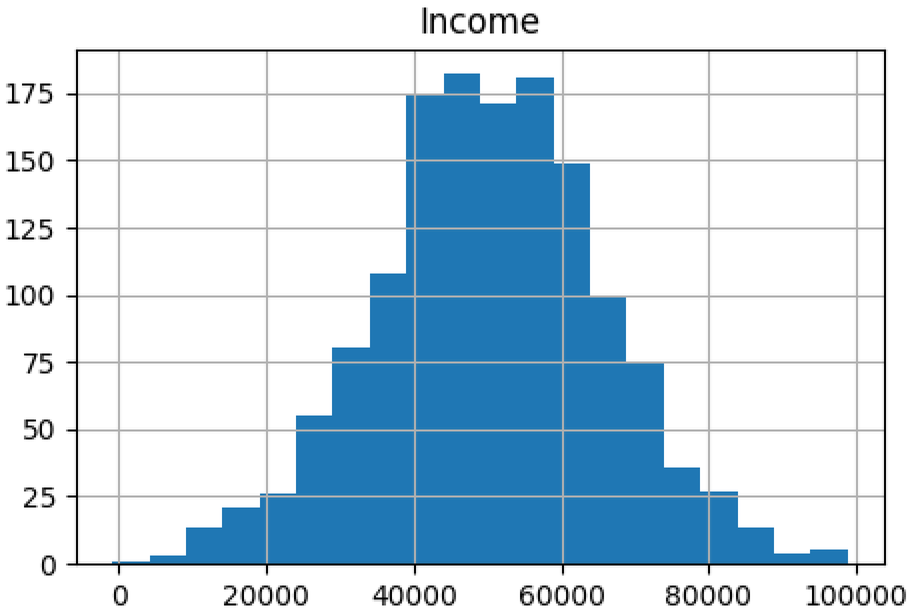
The output or target churn is not balanced as it has 757 instances of no longer user and 743 instances of users.



Explanatory Data Analysis:



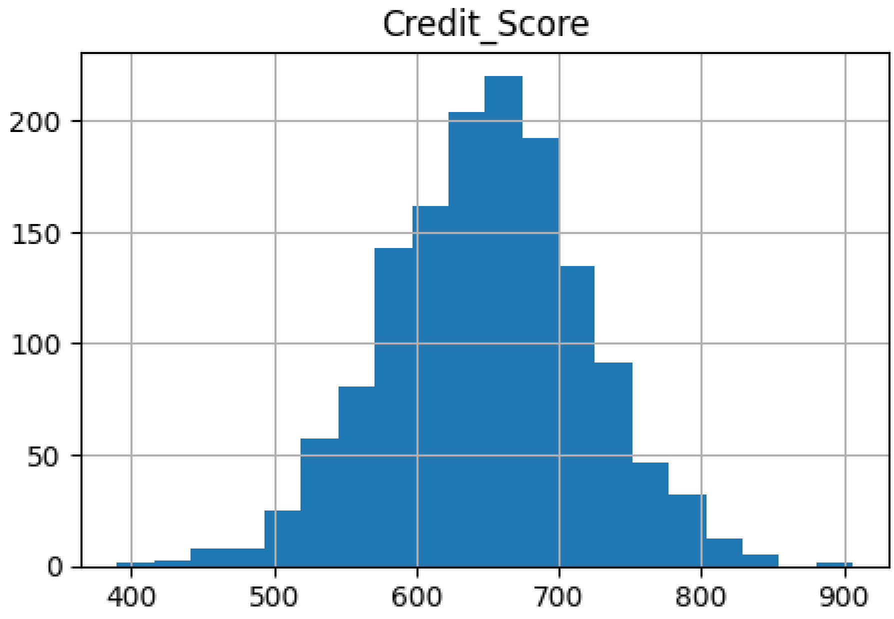
- ➔ The age distribution looks almost uniform between 20 to 70 years.
- ➔ No strong skewness; customers are evenly spread across different age groups.
- ➔ Age might not be a strong predictor of churn or purchases since distribution is balanced.



- ➔ Income follows a normal-like distribution, centered around 40k–60k.

- ➔ Very few customers earn below 20k or above 90k.

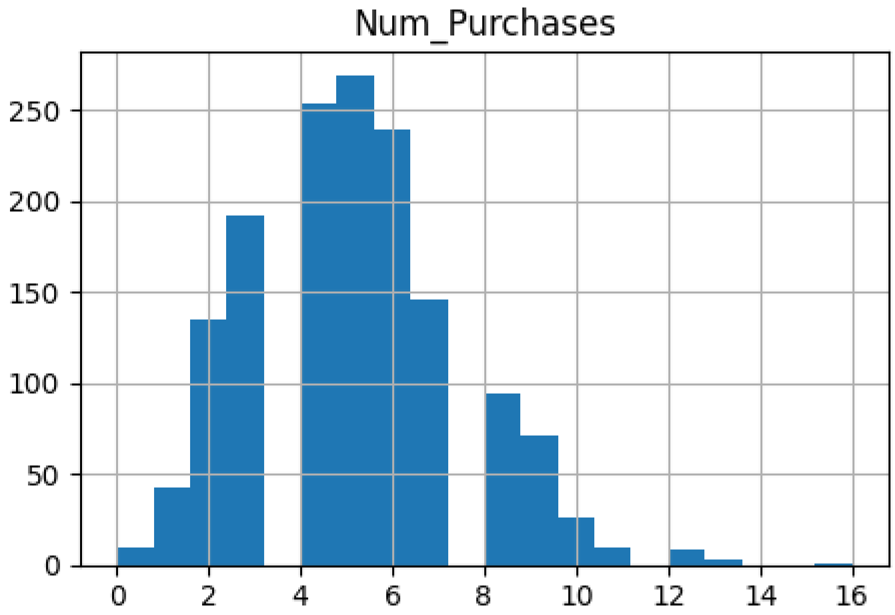
- ➔ Most customers belong to a middle-income group. Extreme incomes are rare and may represent outliers.



- ➔ Distribution is normal, centered around 650–700.

- ➔ Few customers have very low (<500) or very high (>850) scores.

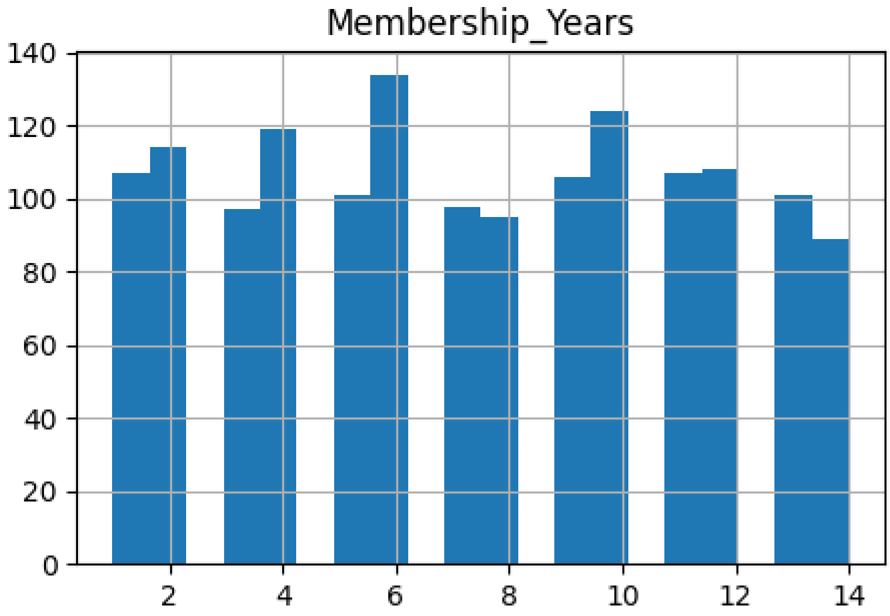
- ➔ Credit score variation is moderate, most customers are in the “good credit” range.



➔ Distribution is right-skewed (positively skewed).

➔ Most customers make 3–6 purchases, while a few make over 10 purchases.

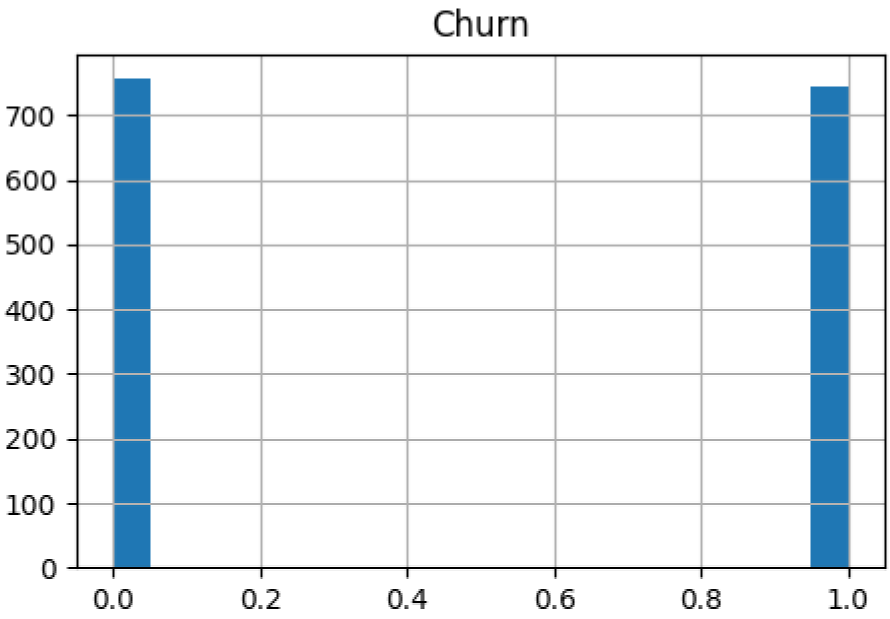
➔ The majority are average buyers, with a small group of "high spenders" (could be loyal customers).



➔ Distribution is almost uniform from 1 to 14 years.

- Customers are spread across all membership durations, with slight spikes around 6 and 9 years.

- Customer retention is consistent across years, no single group dominates.

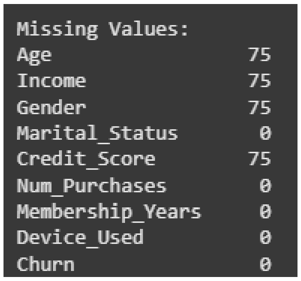


- Churn is binary (0 or 1).

- The dataset is not totally balanced (roughly equal numbers of churned and non-churned customers).

3. Dataset Pre-processing

Problem 1: There are some missing values in the dataset



```
Missing Values:
Age                75
Income             75
Gender             75
Marital_Status      0
Credit_Score       75
Num_Purchases       0
Membership_Years    0
Device_Used         0
Churn               0
```

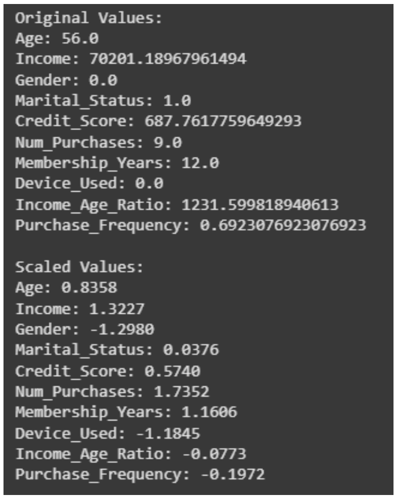
Solution: We identified missing values in 'Age', 'Income', 'Gender', and 'Credit_Score'. These were handled using median imputation for numerical features and mode imputation for categorical features.

Problem 2: Some features (Gender, Marital_Status, Device_Used) are categorical. Categorical Variables can not be used in Machine learning.

<u>Solution:</u> All categorical features were converted into numerical form using Label Encoding. This step was necessary to allow the KNN predictive imputation algorithm to function, as it requires all input data to be numerical

Problem 3: The numerical features in the dataset are vastly different scales. For example, Age ranges from 20 to 70 years, while Membership_years are distributed from 1-14 years. KNN and Neural Networks are highly sensitive to the scale of features.

<u>Solution:</u> To mitigate the problem we applied feature scaling Standard Scaler to range the scale between 0 to 1. This ensures no single feature is biased due to its scale.



```
Original Values:
Age: 56.0
Income: 70201.18967961494
Gender: 0.0
Marital_Status: 1.0
Credit_Score: 687.7617759649293
Num_Purchases: 9.0
Membership_Years: 12.0
Device_Used: 0.0
Income_Age_Ratio: 1231.599818940613
Purchase_Frequency: 0.6923076923076923

Scaled Values:
Age: 0.8358
Income: 1.3227
Gender: -1.2980
Marital_Status: 0.0376
Credit_Score: 0.5740
Num_Purchases: 1.7352
Membership_Years: 1.1606
Device_Used: -1.1845
Income_Age_Ratio: -0.0773
Purchase_Frequency: -0.1972
```

4. Dataset Splitting:

The dataset was split into 70% training (1050 samples) and 30% testing (450 samples) using stratified sampling to maintain the class distribution of the target variable.

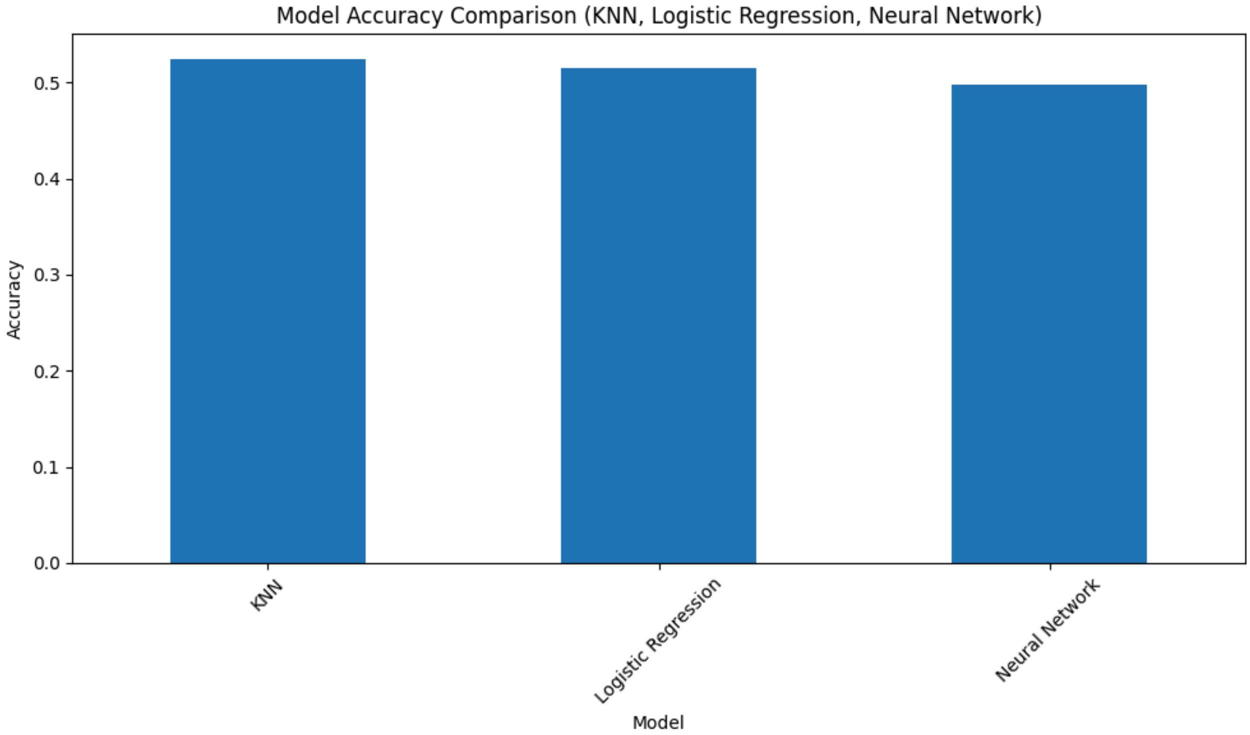
5. Model training & testing (Supervised):

We applied KNN, Logistic Regression and Neural Network models. All models were trained on the training set.

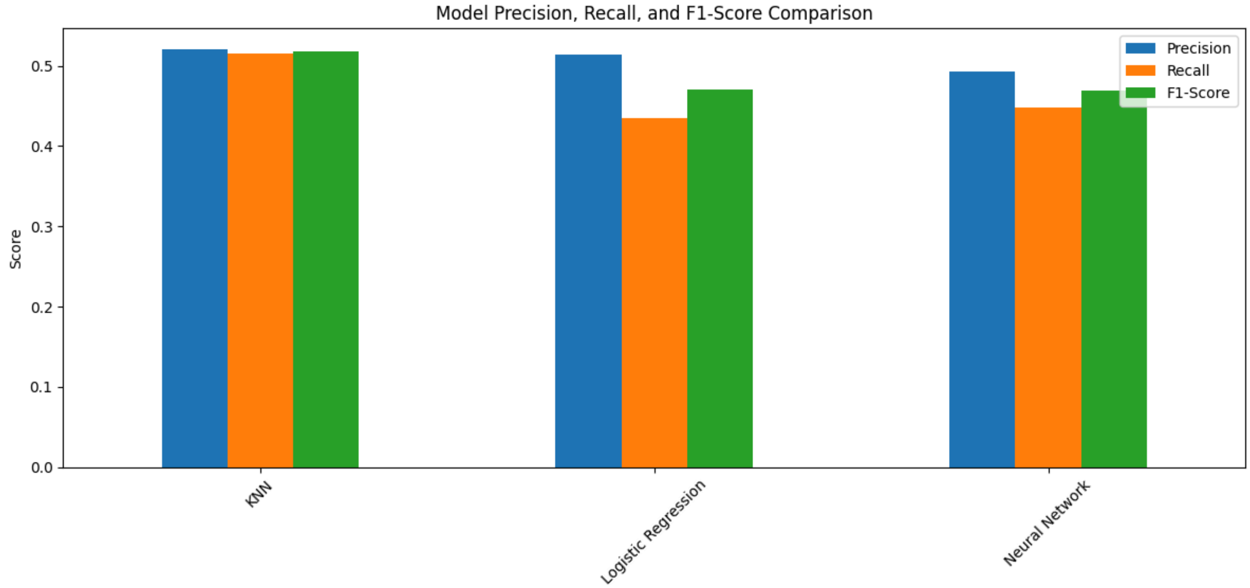
6. Model Selection/Comparison Analysis

Bar Plot

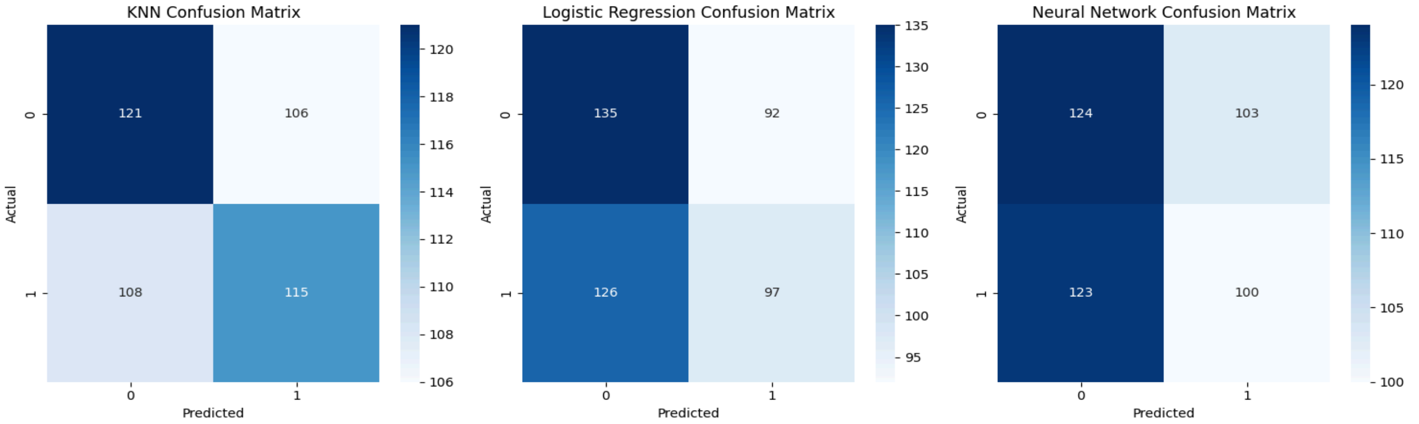
Accuracy



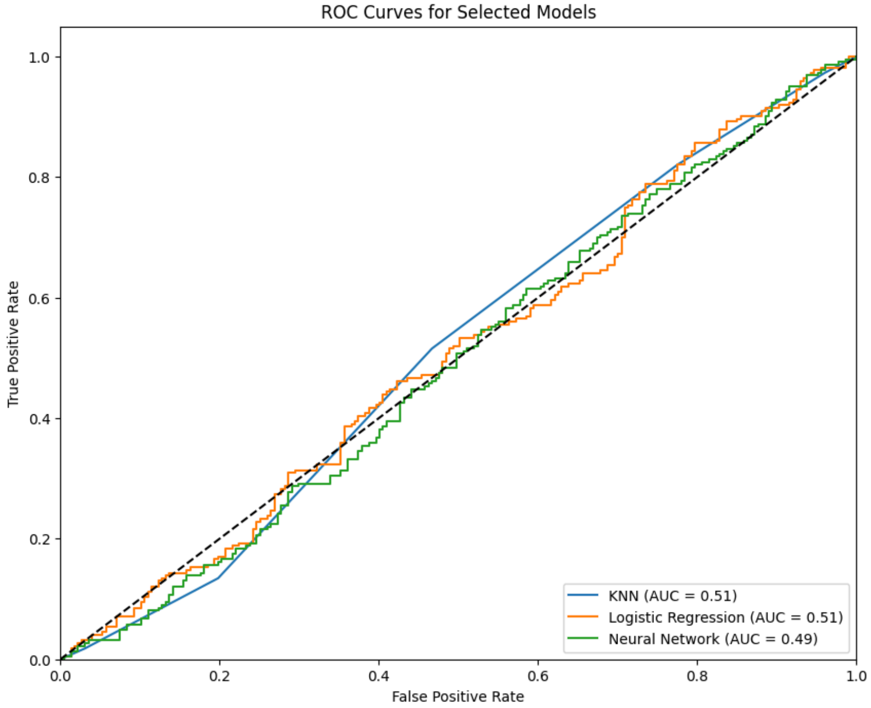
Precision, Recall and F1-Score Comparison



Confusion Matrix

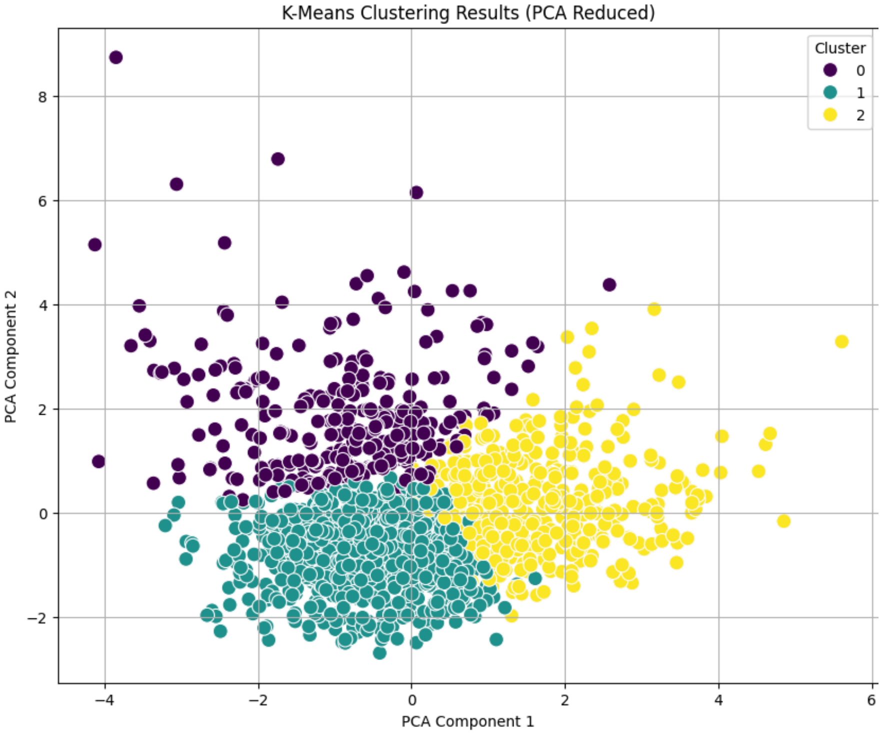


ROC Curves for Selected Models



<u>Unsupervised Learning: K-Means Clustering:</u>

In this step, K-Means clustering was applied to group customers into **three clusters (k=3)** for demonstration. The goal was to identify hidden patterns in the data without using the churn label, allowing us to explore natural customer segments. To make the results easier to visualize, **PCA** was used to reduce the dataset into two components, and the clusters were plotted in 2D space. This helps to see how customers are distributed across clusters and whether meaningful groups exist. Finally, the **cluster distribution** was examined to understand how many customers fall into each group.



7. Conclusion:

From the results, we learned that predicting customer churn is challenging since no single feature strongly relates to churn. The models gave reasonable performance, but accuracy was limited because churn depends on many subtle factors and some external ones not captured in the dataset.The accuracies of the KNN, Logistic Regression, and Neural Network models are 0.5244, 0.5156, and 0.4978, respectively. So after analyzing we can say KNN model performs the best in this case as the accuracy rate is highest. We faced challenges with missing values, categorical features, and scaling, but preprocessing helped to overcome them. Overall, this project showed the importance of proper data preparation and model comparison, while also providing useful insights into customer segments and churn behavior.