Inspiring Excellence

CSE422: Artificial Intelligence

Project Name: Predicting Customer Churn Using Machine Learning Techniques

***Submitted By*:**

Anindita Dutta (22101031)

Angkon Dutta Joy (22101024)

Group :12

Section 07

Table of Contents

Introduction:

The project, "Predicting Customer Churn Using Machine Learning Techniques," classifies customers based on the topic where they stop using a particular company's product based on their likelihood. It is a key criteria for businesses, especially in service-oriented industries or subscription based ones as it influences revenue and growth of the company. Machine learning can provide reliable solutions with proper data and resources to predict and reduce customer churn, helping improve business and increase customer satisfaction. This project tries to develop a machine learning-based solution that can accurately predict which customers are likely to churn, helping the businesses to take necessary steps which will help the company in long run for regaining those customers.

Aims

1. Develop a machine learning model to predict customer churn with high accuracy.
2. Provide such solutions which can be adapted in different types of industries such as banking, and e-commerce.
3. Customer categorization based on their needs ,likelihood ctc.
4. Identifying the factors which are making the customers choose another business.

Problems Aimed to Solve

1. Failing in identifying potential churners before they leave which causes lost revenue.
2. The cost problem when approaching new customers rather than retaining the previous ones.
3. Inefficient allocation of resources

4. Inability to personalize retention strategies due to a lack of segmentation among customers.

Motivation:

The motivation for predicting customer churn lies in :

1. Financial Implementation: The significance of financial and strategic importance lies in retaining customers as it is more cost-effective than acquiring new ones, as churn leads to revenue loss .
2. Customer Appreciation: Machine learning provides a proactive, data-driven approach to identifying at-risk customers, enabling businesses to address dissatisfaction, improve customer satisfaction, and offer personalized retention strategies.
3. Competitive Marketing : In competitive markets, reducing churn strengthens brand loyalty and provides a competitive edge.
4. Real world Skill: Machine learning allows the practical application of machine learning techniques to solve real-world problems, offering scalable solutions for various industries and contributing to sustainable growth.

Dataset Description:

- **Source Link:**https://www.kaggle.com/datasets/blastchar/telco-customer-churn

- **Dataset Description:**

Number of rows: 7043

Number of columns: 21

1. **Number of Features:**The dataset has **21 features**, including the output feature ("Churn"). These consist of:
    - **Categorical Features:** 17
    - **Quantitative Features (Numerical):** 4

2.Problem Type:

- The output feature is "Churn", which contains discrete values ("Yes" and "No").
- This indicates a classification problem, as the task involves predicting whether a customer will churn or not.

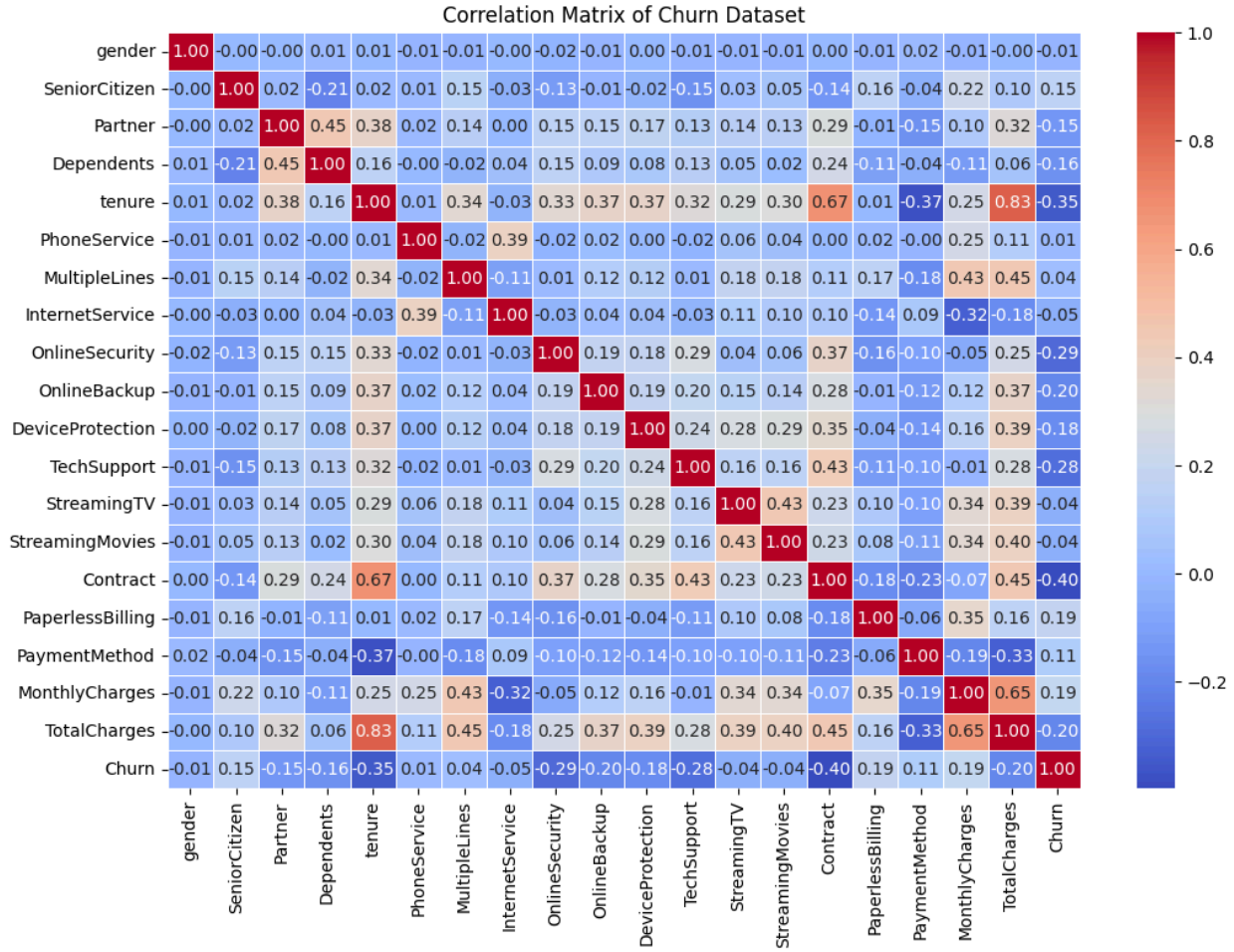
3.**Number of Data Points:**

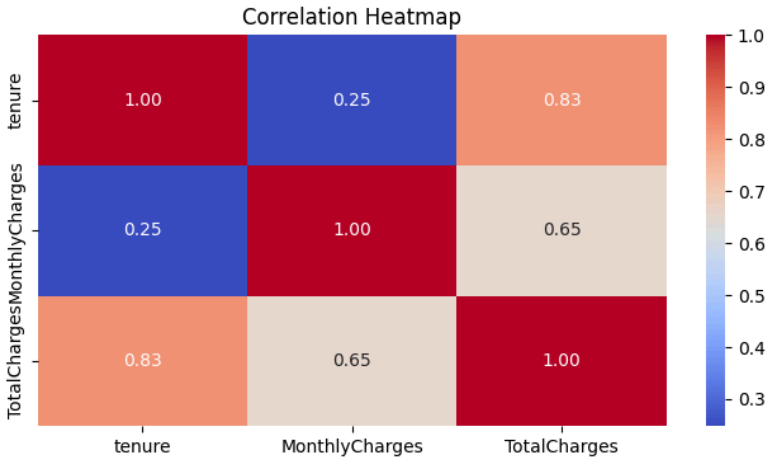
- The dataset contains **7,043 rows**, each representing a unique customer.

4.**Feature Types:**

- **Quantitative Features:**These features contain numeric values . Examples:
    - **Tenure:** The number of months a customer has stayed with the company.
    - **MonthlyCharges:** Monthly billing amount for the customer.
    - **TotalCharges:** Total amount of the customer's tenure.
- **Categorical Features:** These features contain nominal or ordinal data, which needs to be labeled for machine learning models. Examples:
    - **Gender:** Male or Female.
    - **InternetService:** DSL, Fiber optic, or None.
    - **Contract:** Month-to-month, One year, or Two years.

Correlation Analysis:



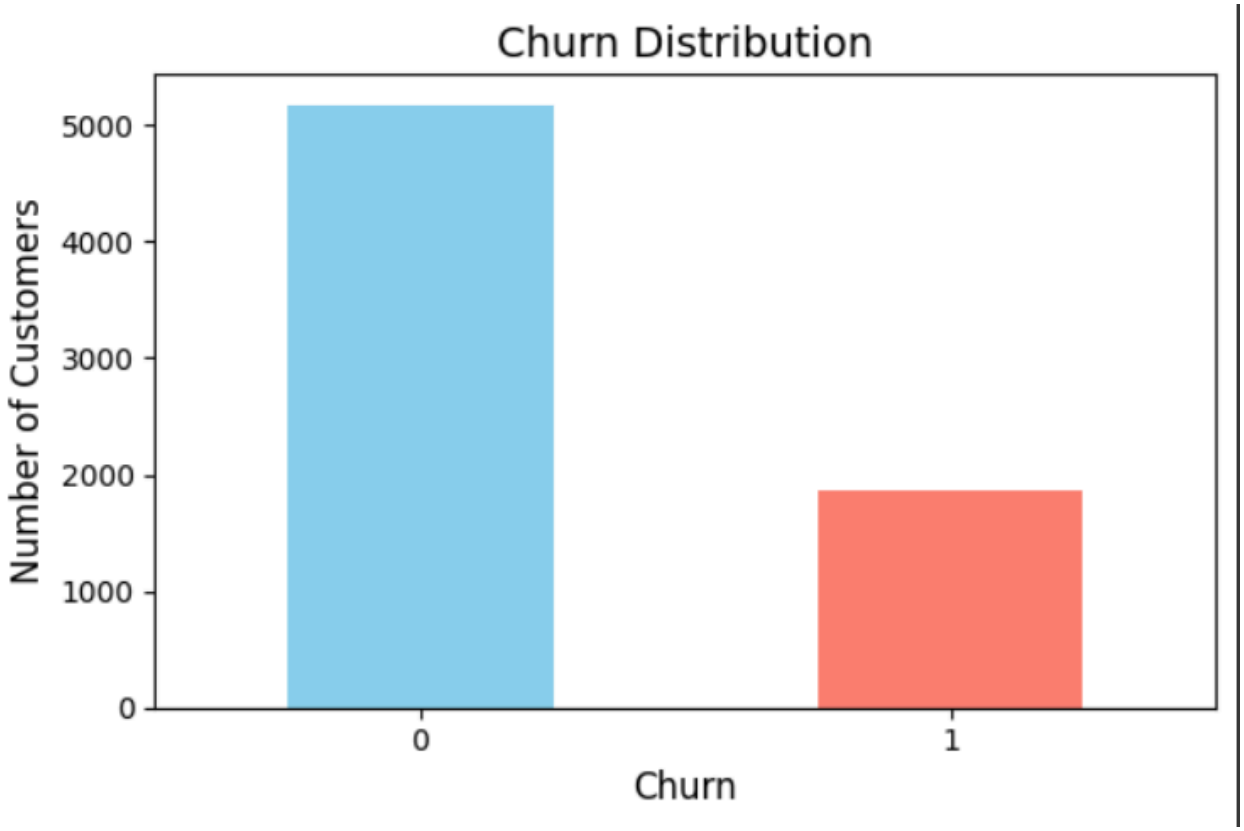


Correlational studies show that tenure is 83% of the variation in total charges so that strengthens the relationship. Tenure has a 67% correlation with long term contracts which means the type of customers who have stable income are those who stay for long periods. As with churn, an extended period of stay reduces the chances of churn by 35% and a shift to long term contracts reduces the risk of churn by 40%. On the other hand, an increase in monthly charges increases the chances of churn by a small margin of 20% which means cost can be a factor for the customers. OnlineSecurity, TechSupport and others also help in customer retention although their effect value is lesser on their own. Some features like gender, phone service and others which have correlations approaching 0% of the variable in question are of very little importance and thus their influence on churn is minimal hence less preferred to be used in predictive models. So tenure, contract type and monthly charges are very important in customer retention.

Imbalanced Data:

The output feature **"Churn"** is imbalanced:

Because Class **"No"** has significantly more instances compared to class **"Yes"**. This is a critica; situation as it can create biases in machine learning models to favor the majority class. If we present it by Bar Chart, the chart shows a higher count of customers who did not churn compared to those who did.This imbalance highlights the need for techniques like oversampling (e.g., SMOTE) or class weighting during model training to ensure balanced performance.



Dataset pre-processing:

1.Customer id was causing biases and the correlation value was too minor ,so we Deleted Customer ID column as it will not affect other data.

2. This dataset does not have any null values.

3. Label encoding of target column “Churn” is done .The Yes, NO values of “Churn” column were replaced by numerical values 1 and 0 as Machine learning models can only work with numerical values.

4. Label encoding of categorical features is also done . Dictionary was used to save the encoders. Like { gender:Label Encoder()

partner:Label Encoder()}

5. Missing values in the TotalCharges column were replaced with 0.0. (In the dataset it was shown as an object but we had to turn it into float 64.

6. The imbalanced data issue is handled indirectly by applying **Class Weighting** during hyperparameter tuning with Random Search CV.

Feature scaling:

StandardScaler feature scaling is used in this project to scale the features. It is one of the crucial steps using algorithms that can be sensitive to feature magnitudes like K-Nearest Neighbours, SVM. One of the main purposes of using the scale is so that features with bigger magnitudes can dominate distance based algorithms like KNN and SVM and scaling make sure all the features can contribute equally. Also some machine learning methods perform better after scaling.The training data was scaled using the fit_transform() method to compute the necessary scaling parameters (e.g., mean and standard deviation for standardization). These same parameters were then applied to the test data using the transform() method, ensuring consistent scaling across both datasets.

Dataset Splitting:

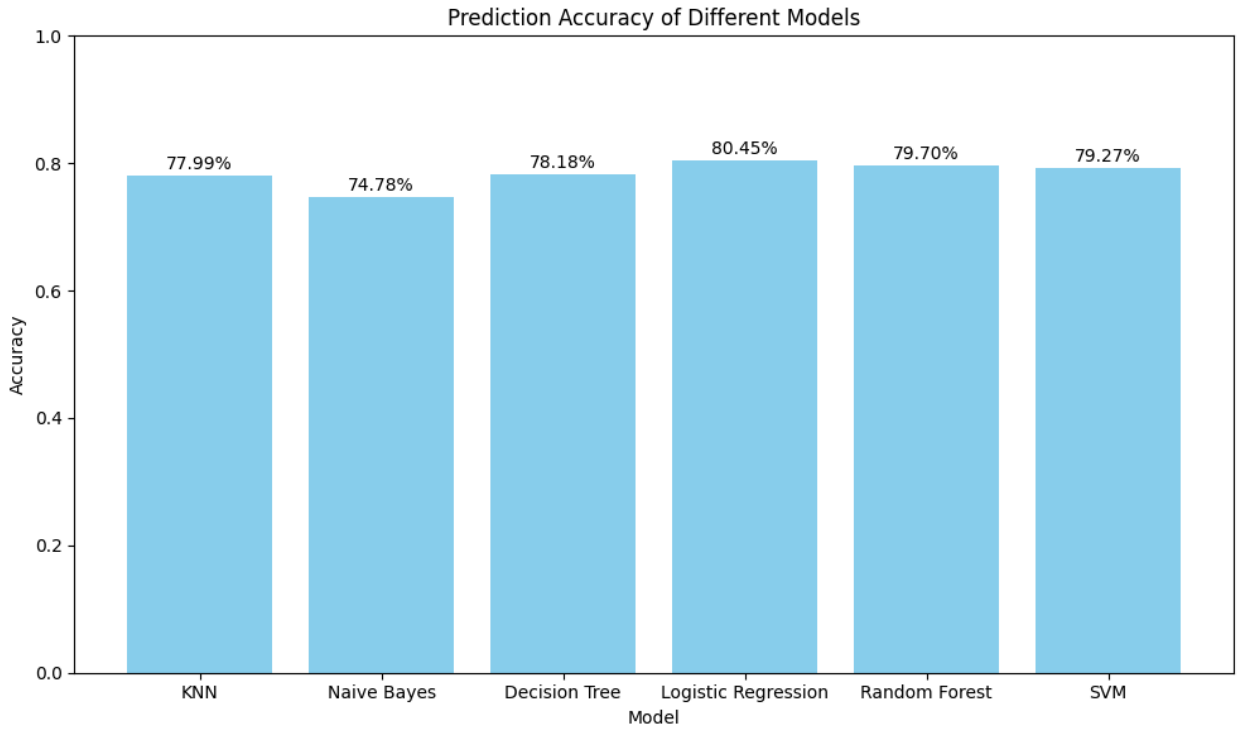
Dataset splitting was Random. Test set was 30% and the training set was 70% which means among the 7043 dataset , the amount of testing dataset was 2113 and training dataset was 4930.

Model Training and Testing:

We used six different models (K-Nearest Neighbors (KNN), Naive Bayes Decision Tree, Logistic Regression, Random Forest, Support Vector Machine (SVM)) to predict churn and see the outcome. We have used StandardScaler for all of them and used RandomizedSearchCV for optimization and tuning purposes.Accuracy, Confusion Matrix, and Classification Report are computed for each model. Results are visualized with a bar chart comparing model accuracies.

From the project, we can see that Logistic Regression came first, with the highest accuracy at 80.45%, showing very reliable performance. Then comes Random Forest with 79.70%, slightly above SVM, which had an accuracy of 79.27%. The Decision Tree showed a good score of 78.18% accuracy, while KNN followed closely with 77.99%, and lastly, Naive Bayes ranked at the bottom with poor accuracy of 74.78%. Therefore, Logistic Regression is the best choice to satisfy the task at hand but is followed by Random Forest and SVM, in that order.

Model selection/Comparison analysis:



Output of K-Nearest Neighbors Classifier (KNN):

```
================ K-Nearest Neighbors Classifier - Test Data ================
Best Parameters: {'metric': 'manhattan', 'n_neighbors': 14, 'weights': 'uniform'}
Accuracy of Best KNN Classifier on Test Data: 77.99%

Confusion Matrix on Test Data:
[[1364  188]
 [ 277  284]]

Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.83      0.88      0.85      1552
           1       0.60      0.51      0.55       561

    accuracy                           0.78      2113
   macro avg       0.72      0.69      0.70      2113
weighted avg       0.77      0.78      0.77      2113
```

Output of Naive Bayes:

```
================= Gaussian Naive Bayes Classifier - Test Data ================
Accuracy of Gaussian Naive Bayes Classifier on Test Data: 74.78%

Confusion Matrix on Test Data:
[[1167  385]
 [ 148  413]]

Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.89      0.75      0.81      1552
           1       0.52      0.74      0.61       561

    accuracy                           0.75      2113
   macro avg       0.70      0.74      0.71      2113
weighted avg       0.79      0.75      0.76      2113
```

Output of Decision Tree:

```
================= Decision Tree Classifier - Test Data =================
Accuracy of Decision Tree Classifier on Test Data: 78.18%

Confusion Matrix on Test Data:
[[1408  144]
 [ 317  244]]

Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.82      0.91      0.86      1552
           1       0.63      0.43      0.51       561

    accuracy                           0.78      2113
   macro avg       0.72      0.67      0.69      2113
weighted avg       0.77      0.78      0.77      2113
```

Output of Logistic Regression:

```
================= Logistic Regression Classifier - Test Data =================
Accuracy of Logistic Regression Classifier on Test Data: 80.45%

Confusion Matrix on Test Data:
[[1380  172]
 [ 241  320]]

Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.85      0.89      0.87      1552
           1       0.65      0.57      0.61       561

    accuracy                           0.80      2113
   macro avg       0.75      0.73      0.74      2113
weighted avg       0.80      0.80      0.80      2113
```

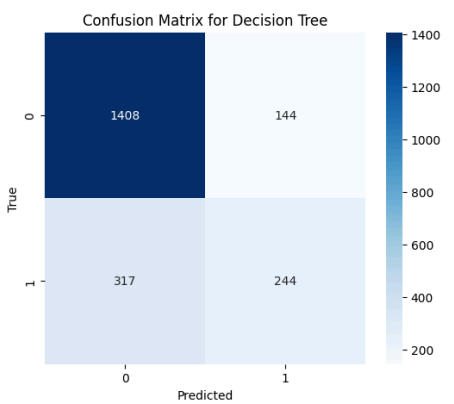
Output of Random Forest:

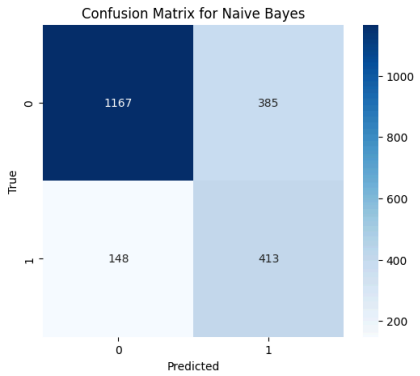
```
================= Random Forest Classifier - Test Data =================
Accuracy of Random Forest Classifier on Test Data: 79.70%

Confusion Matrix on Test Data:
[[1408  144]
 [ 285  276]]

Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1552
           1       0.66      0.49      0.56       561

    accuracy                           0.80      2113
   macro avg       0.74      0.70      0.72      2113
weighted avg       0.79      0.80      0.79      2113
```

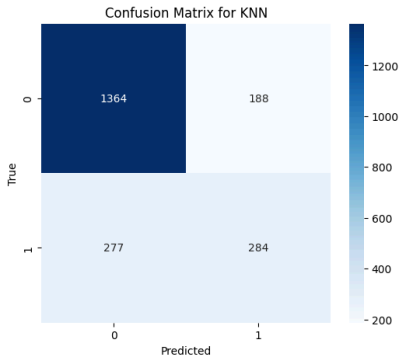
Output of Support Vector Machine (SVM):

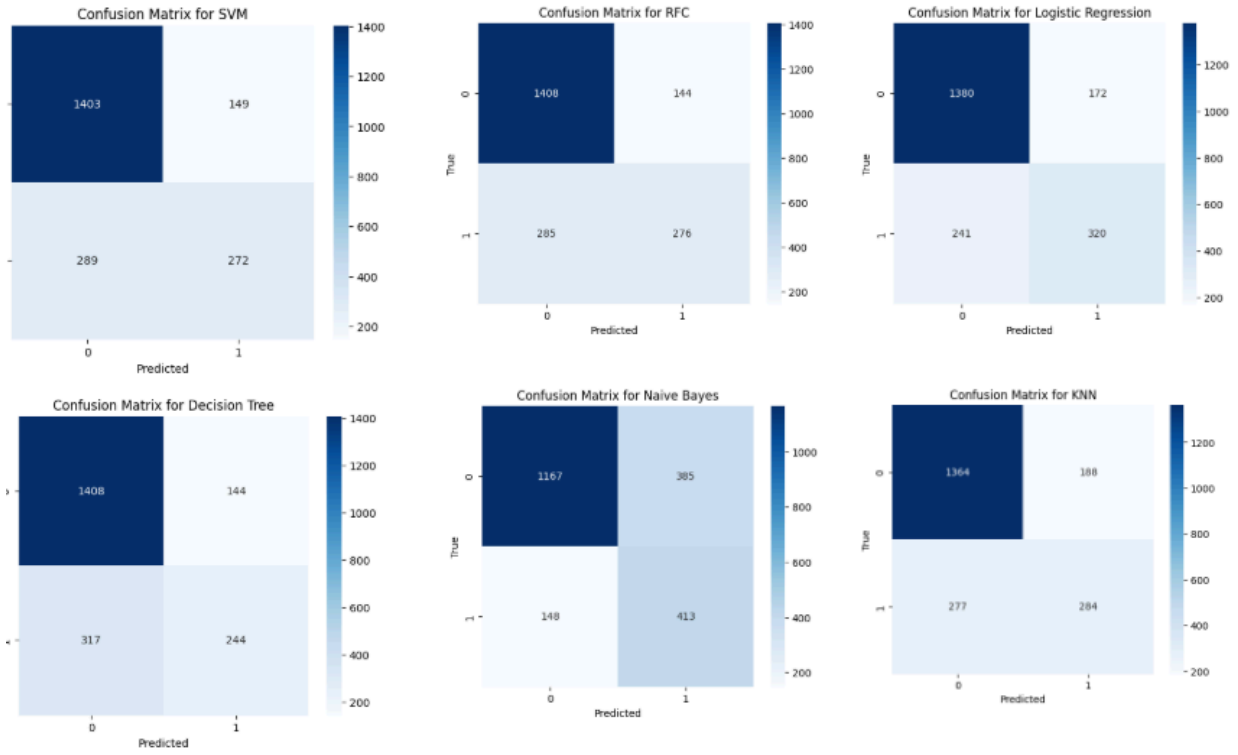
```
================= Support Vector Machine Classifier - Test Data =================
Accuracy of SVM Classifier on Test Data: 79.27%

Confusion Matrix on Test Data:
[[1403  149]
 [ 289  272]]

Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.83      0.90      0.86      1552
           1       0.65      0.48      0.55       561

    accuracy                           0.79      2113
   macro avg       0.74      0.69      0.71      2113
weighted avg       0.78      0.79      0.78      2113
```

Confusion Matrix:









Conclusion:

The project "Predicting Customer Churn Using Machine Learning Techniques" shows the risk of losing customers of a business. . With 7,043 rows and 21 columns including 4 numerical columns and 17 categorical columns this project works on preprocessing steps included deleting the CustomerID column to avoid bias, using class weight for handling imbalanced and scaling features using StandardScaler to keep the consistency in algorithms like KNN and SVM. The dataset was split into 70% training data (4,930 rows) and 30% testing data (2,113 rows) for model evaluation.

Six machine learning models were tested: Logistic Regression, Random Forest, SVM, Decision Tree, K-Nearest Neighbors (KNN), and Naive Bayes. Among these, Logistic Regression gave the highest accuracy of 80.45%, followed by Random Forest (79.70%) and SVM (79.27%). Other models were slightly lower, Decision Tree 78.18%, KNN 77.99%, Naive Bayes 74.78%. These results were visualized using bar charts below, Logistic Regression is the most reliable and efficient for this task.

The project also found the key factors of churn such as tenure, payment methods and contract types so targeted strategies can be applied to retain customers. The methodology is scalable and can be applied to industries like banking, telecommunications and e-commerce. This practical application of machine learning shows how businesses can use predictive analytics to increase customer satisfaction, reduce churn and maximize long term profitability. Future work can include ensemble models, customer sentiment data or advanced optimization techniques to get better performance.