**Machine Learning Churn Prediction Project Documentation**

1. **Introduction:**

**Project Overview :**

The goal of this project is to build a machine learning model for predicting customer churn in a subscription-based service. Customer churn refers to the situation where customers discontinue their subscription or leave the service. By accurately predicting churn, companies can take proactive measures to retain customers and maintain their subscription base.

**2. Data Preprocessing :**

**Loading the Dataset :**

The project begins by loading the provided dataset containing customer information, subscription details, and usage data. The dataset includes columns such as CustomerID, Age, Gender, Location, Subscription\_Length\_Months, Monthly\_Bill, Total\_Usage\_GB, and Churn.

**Initial Data Exploration**

An initial exploration of the dataset is performed to understand its structure and characteristics. This involves checking for the number of rows and columns, data types, and summary statistics of numerical variables.

**Handling Missing Data and Outliers**

The dataset is examined for missing data and outliers. Although no Outliers are found Techniques like imputation or removal of missing values and outlier detection are applied to ensure the quality of the data.

**Encoding Categorical Variables**

Categorical variables like Gender and Location are encoded using one-hot encoding. This process converts categorical variables into numerical format suitable for machine learning algorithms.

**Train-Test Split**

The dataset is split into training and testing sets. This division ensures that the model is trained on one subset and evaluated on another, enabling an unbiased assessment of its performance.

**3. Feature Engineering :**

**Creating New Features:**

Relevant features are generated from the dataset to enhance the model's predictive capabilities. Features such as

**1) "Average Monthly Usage" (Total\_Usage\_GB divided by Subscription\_Length\_Months) are created to capture additional insights from the data,**

**2) Interaction between Usage and Bill**

**3)Categorizing age**

**Feature Scaling or Normalization**

feature scaling or normalization is applied to ensure that numerical features are on the same scale. This helps prevent certain features from dominating the learning process.

**4. Model Building :**

**Choosing Machine Learning Algorithms :**

Several machine learning algorithms, including Logistic Regression, Random Forest,SVM,Naïve Bayes,Gradient Boosting,KNN and Neural Network, are considered for the churn prediction task. These algorithms offer different trade-offs between interpretability and complexity.

**Training and Validating Models**

Models are trained using the training dataset and validated using metrics such as accuracy, precision, recall, and F1-score. Cross-validation is employed to obtain a more accurate estimate of how well the models will generalize to new data.

The best outputs received are through KNN :

KNN:

Accuracy: 0.50

Precision: 0.49

Recall: 0.50

F1-score: 0.50

**5. Model Optimization**

**Cross-Validation**

Cross-validation is performed to assess the models' performance across multiple folds of the training data. This helps to identify any overfitting issues and provides a more reliable measure of their effectiveness.

The acquired accuracy is nearly 0.5 after applying cross validation

**7. Conclusion**

**Summary of the Project**

The project focused on building a machine learning model to predict customer churn in a subscription-based service. It covered data preprocessing, feature engineering, model building, and optimization.