# Documentation and Report

## Customer Churn Prediction Machine Learning Project

### 1. Introduction:

Customer churn is a critical challenge for businesses, especially in subscription-based services. This project focuses on developing a machine learning model to predict customer churn using a dataset containing various customer attributes.

### 2. Data Collection and Preprocessing:

The dataset contains the following columns:
Name: Customer's name (categorical)
Age: Customer's age (numerical)
Gender: Customer's gender (categorical)
Location: Customer's location (categorical)
Subscription_Length_Months: Length of subscription in months (numerical)
Monthly_Bill: Monthly bill amount (numerical)
Total_Usage_GB: Total usage in gigabytes (numerical)
Churn: Target variable indicating churn (0 for not churned, 1 for churned)

Data preprocessing steps include handling missing values, encoding categorical variables (One-Hot or Label Encoding), and splitting the data into training and testing sets.
There is no any null and duplicate values presents in given dataset.

## 3. Exploratory Data Analysis (EDA):

EDA involves visualizing and analyzing the dataset to gain insights. This includes distribution plots, correlation matrices, and churn rate calculations. EDA helps in understanding relationships between features and their impact on churn. There is 100000 rows and 9 columns are present.

## 4. Feature Engineering:

Feature engineering enhances model performance by creating new features or transforming existing ones. For instance, creating a "Usage_Per_Subscription" feature by dividing Total_Usage_GB by Subscription_Length_Months. Feature scaling might also be necessary for some models.

## 5. Model Selection:

Different machine learning algorithms will be considered, such as Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting. Model selection depends on the dataset size, complexity, and interpretability.

## 6. Model Training and Evaluation:

The dataset will be split into training and testing sets. Models will be trained on the training set and evaluated on the testing set using metrics like accuracy, precision, recall, F1-score. Cross-validation can be employed to tune hyperparameters and assess model robustness.

- The Logistic Regression model accuracy is 49.99% on the test set and 50.34% on the train set
- The Decision Tree model accuracy is 49.99% on the test set and 100% on the train set.
- The Random Forest model accuracy is 49.29% on the test set and 100% on the train set.
- The SVC model accuracy is 49.29% on the test set and 100% on the train set.
- The KNN model accuracy is 49.29% on the test set and 100% on the train set.
- The GradientBoostingClassifier model accuracy is 49.29% on the test set and 100% on the train set

## 7. Results and Discussion:

Results will be presented through confusion matrices, and relevant metrics. Models will be compared based on their performance. Discussion will revolve around features contributing to churn prediction and the business implications. .
we use RandomizedSearchCV for increase the accuracy of the model.

**The RandomizedSearchCVmodel accuracy is 93.88%** .

## 8. Conclusion:

Summarize the findings, highlighting the model with the best performance. Emphasize the importance of churn prediction in reducing customer attrition and its potential financial impact.

After implementing feature selection and optimizing hyperparameters, the Random Forest model demonstrated a notable improvement in accuracy.
Now we got accuracy **93.88 %** by using RandomizedSearchCV model

## 9. Future Recommendations:

Suggest potential actions the business can take based on the model's predictions. For instance, targeting high-risk customers with special offers or improving services for segments prone to churn.