***Customer Churn Prediction***

***1. Data Preprocessing:***

In this step, the program begins by loading the provided dataset from an Excel file using the Pandas library. This dataset contains information about customers including their age, gender, location, subscription details, monthly bill, total usage, and whether they have churned or not. After loading the data, it is saved as a CSV file for easier access and to avoid reloading from the Excel file in the future. This initial data exploration step helps in understanding the structure of the dataset. We explore our dataset and we use various method like describe .we take a look of head and tail of dataset.

***2. Handle Missing Data and Outliers:***

Here, the program checks for missing values in the dataset. It uses the `isnull().sum()` function to determine the number of missing values in each column. In our case, there are no missing values in any column. After that, the program calculates correlations between different numerical features using the `data.corr()` function and displays the correlation matrix using a heatmap. This provides insights into how features are related to each other. Non-informative columns like 'CustomerID' and 'Name' are dropped as they don't contribute significantly to the churn prediction.

***3. Encode Categorical Variables:***

The program visualizes the distribution of categorical features using count plots. This gives an overview of how many instances belong to each category. To prepare the data for machine learning, categorical variables like 'Gender' and 'Location' are encoded using one-hot encoding. This process converts categorical values into binary columns, which the model can use for predictions.

***4. Data Visualization:***

The program uses various plots to visualize important aspects of the dataset. It creates a pie chart to display the distribution of churned and not churned customers. Count plots are used to show gender and location distributions. Additionally, a heatmap is generated to visualize the correlations between numerical features, helping to identify potential relationships here we draw a conclusion that age and subscription length month , total used gb and monthly bill are correlated and we will use these correlation later on. Finally, a histogram is used to illustrate the distribution of customer ages.

***5. Split Data:***

Data splitting is crucial for training and evaluating machine learning models. The program separates the dataset into features (X) and the target variable (y), where 'X' contains the features used for prediction and 'y' contains the target variable indicating churn status.

***6. Feature Engineering:***

The program introduces a new feature called 'Bill\_Usage\_Ratio' by dividing the 'Monthly\_Bill' by the 'Total\_Usage\_GB' as we prior observe that they are highly correlated .This ratio represents how much a customer spends relative to their data usage. StandardScaler is applied to scale the features, ensuring that they have similar scales, which is important for many machine learning algorithms to work effectively.

***7. Model Building:***

Two classification models, Random Forest and Gradient Boosting, are initialized and trained using the scaled training data. These models are then used to make predictions on the scaled testing data. To improve predictive performance, an ensemble model is created using a Voting Classifier, which combines the predictions of both the Random Forest and Gradient Boosting models.

***8. Model Optimization:***

The program uses GridSearchCV to perform a hyperparameter search for the Gradient Boosting model as gradient boosting model gives highly accuracy then random forest. Hyperparameters are tunable parameters that affect the model's performance. By testing different combinations of hyperparameters, the program finds the best set of values that optimize the model's accuracy than we put these values in model building part to increase accuracy.

***9. Model Evaluation:***

Both the ensemble model and the optimized Gradient Boosting model are evaluated using the accuracy metric. Accuracy measures how many predictions the model got correct out of the total predictions.

***10. Model Deployment:***

The program defines a function called 'predict\_churn' that takes new customer data as input, preprocesses it, and makes churn predictions using the ensemble model.

*Thankyou*

*Shriya Srivastava*