**Predicting Customer Churn in a Telecommunications Company**

**Introduction:**

The Telco Customer Churn dataset offers a rich resource for understanding customer behavior in the telecommunication industry. This dataset, encompassing 7,043 customers and 21 features, provides valuable insights into factors that can influence customer loyalty and ultimately lead to churn.

The dataset delves into various customer demographics, service subscriptions, account details, and churn status. Here's a breakdown of some key features:

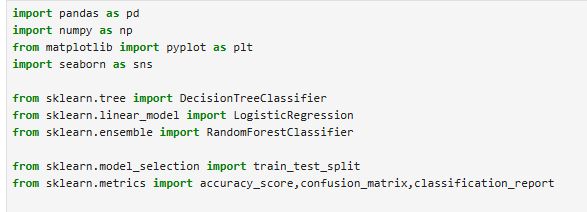
* **Customer Demographics:** Information such as gender, senior citizen status, presence of a partner, and dependents helps paint a picture of the customer base.
* **Service Subscriptions:** Features like phone service, multiple lines, internet service, and subscriptions to additional services like online security, backup, device protection, tech support, streaming TV, and movies reveal the services customers utilize.
* **Account Details:** The dataset includes details on customer tenure Features like contract type (month-to-month, one-year, two-year), billing preferences (paperless vs. paper), and payment method can shed light on customer commitment and potential points of friction.
* **Financial Information:** Monthly charges and total charges show a picture of customer spending habits and their potential value to the company.
* **Churn Status:** This crucial feature indicates whether a customer churned (left the company) within the last month.

Overall, the Telco Customer Churn dataset serves as a powerful tool for telecommunication companies to understand their customer base, predict churn, and ultimately improve customer retention. By leveraging this data, companies can make data-driven decisions to reduce churn, maintain a loyal customer base, and drive business growth.

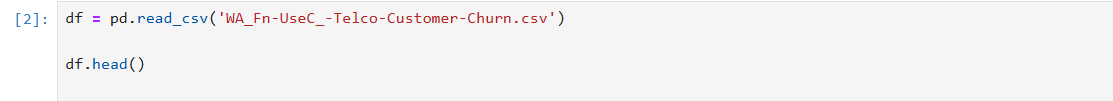
**Data preprocessing :**

The Telco Customer Churn dataset, stored in CSV format, holds valuable customer information.

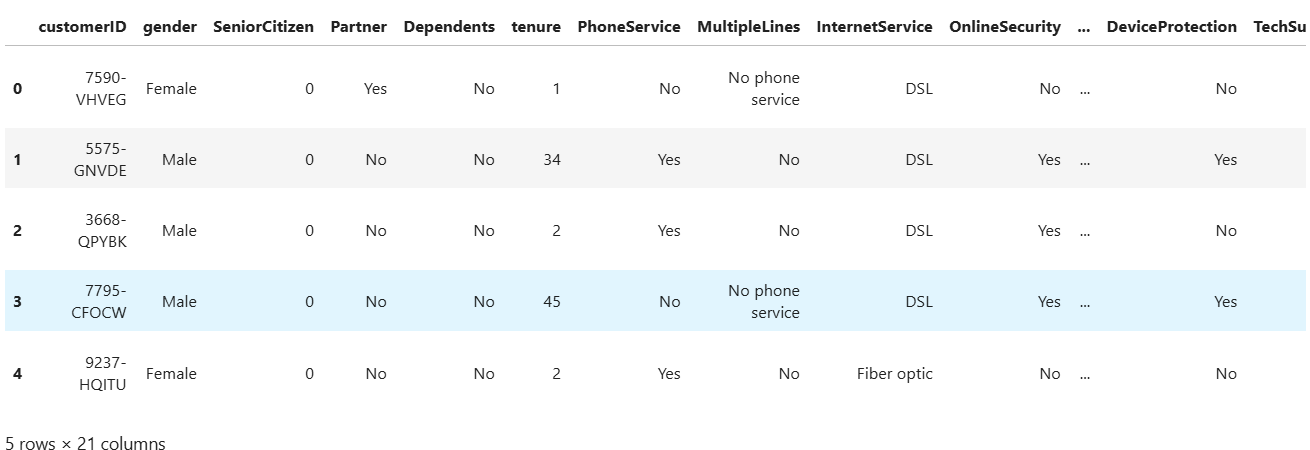
1. **Importing Libraries and Loading Data:**



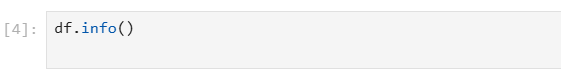
1. **Exploring the Data:**

****

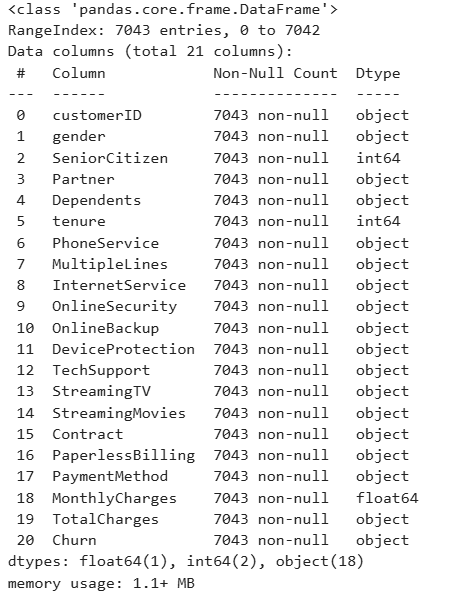
**Output:**

****

1. **Handling Missing Values:**

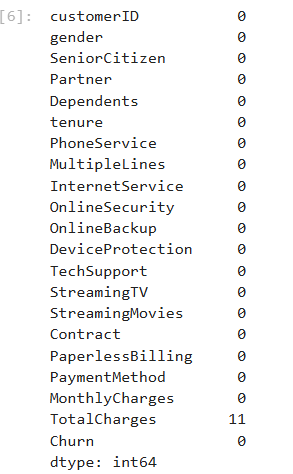
****

**Output:**

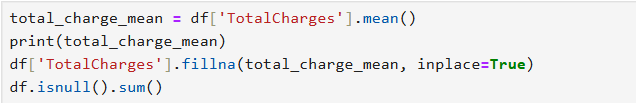
****



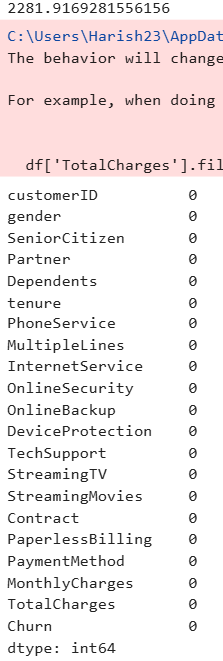
**Output**:



**using the Impute method to fill the missing values:**

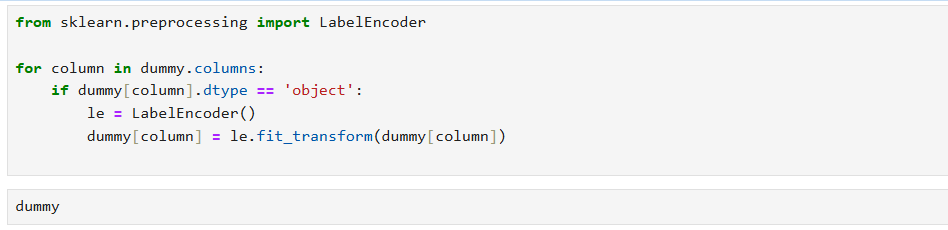
****

**Output:**

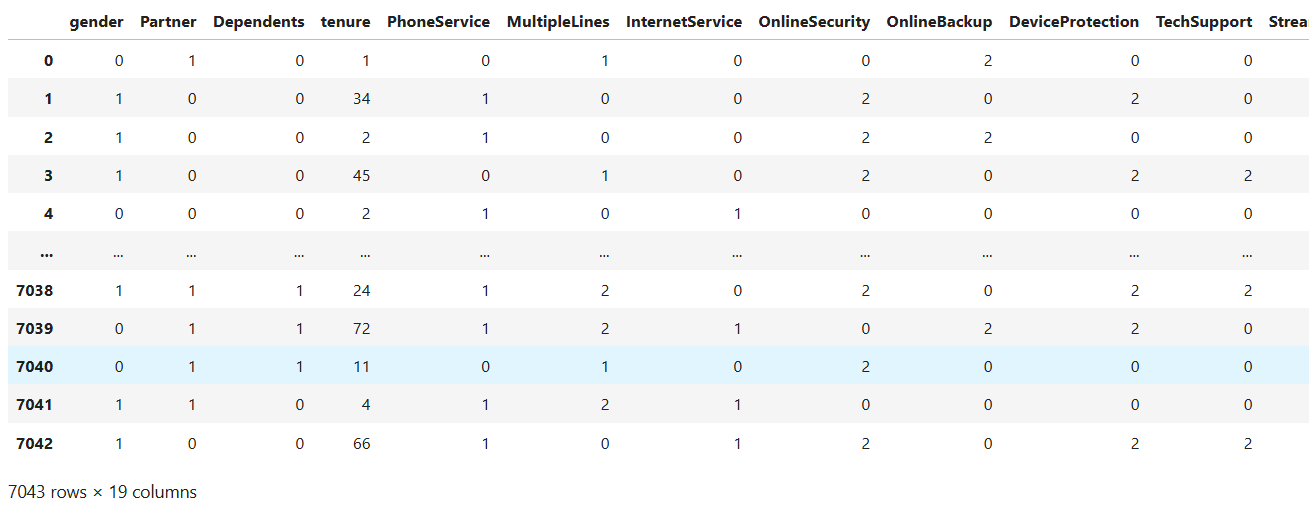
****

1. **Encoding Categorical Features:**

The Categorical Features are the LabelEncoder library used to change the object data type to Numeric data type



**Output:**

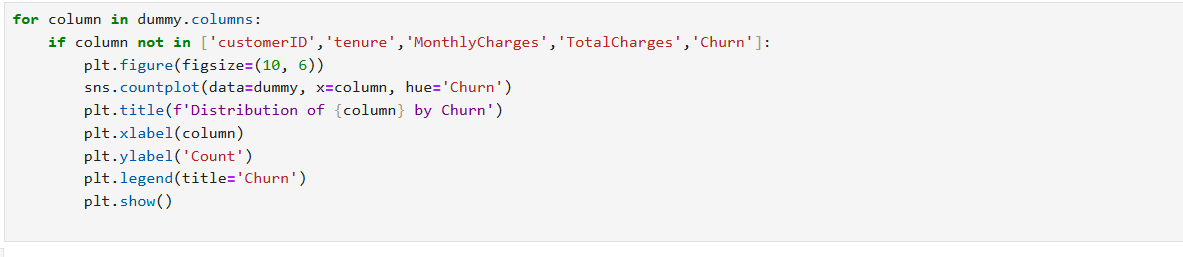


1. **Data visualization:**

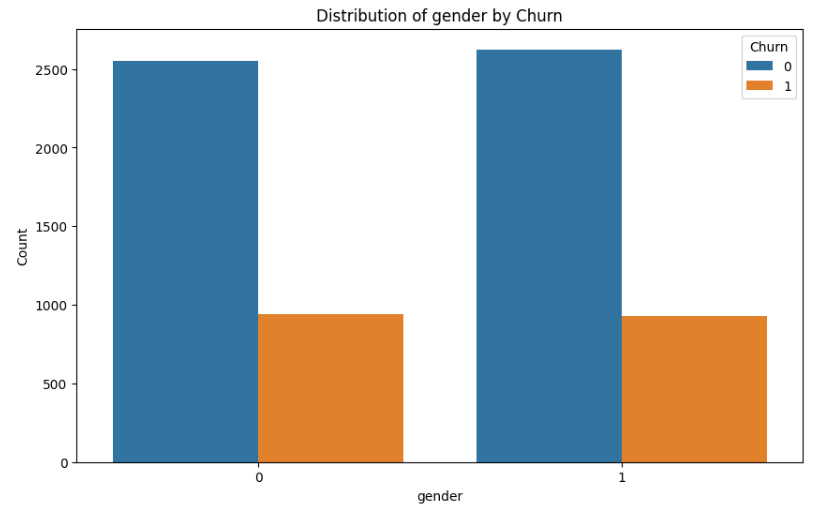
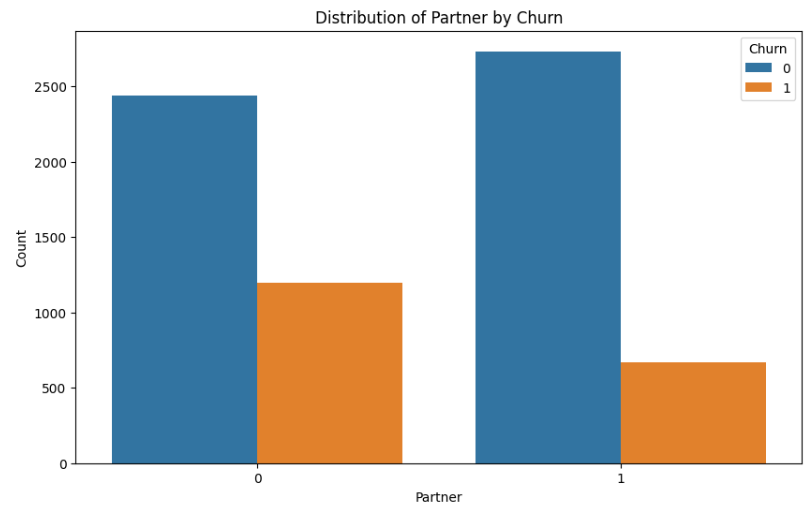
After preprocessing the Telco Customer Churn dataset and converting categorical features using LabelEncoder, we can leverage Matplotlib and Seaborn for data visualization to gain insights into customer churn.

**Visualizing Customer Distribution:**

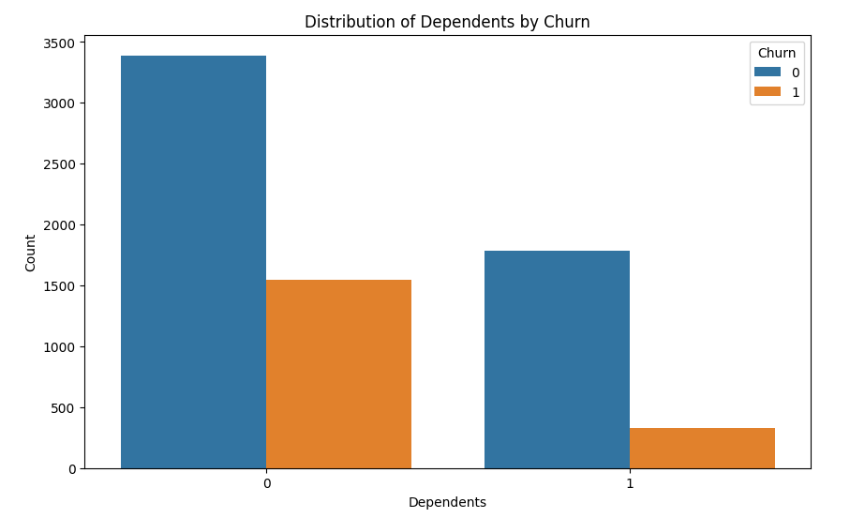
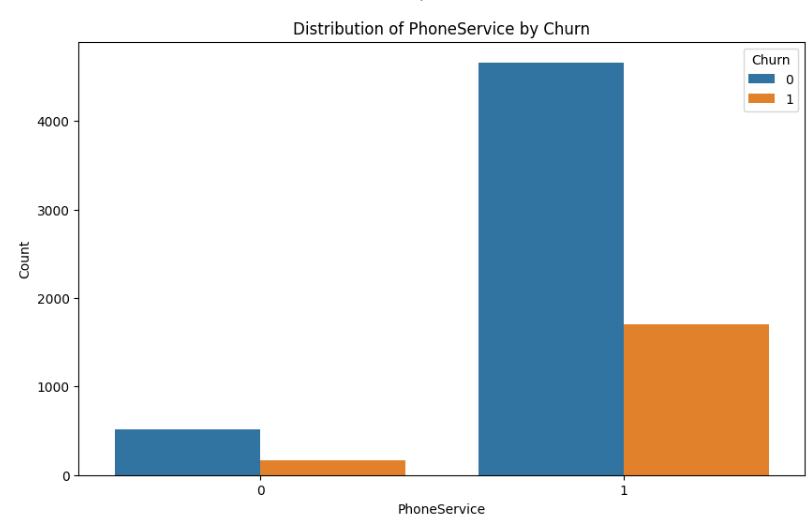
* **Distribution of Churn:** Using a bar chart (Matplotlib) or a count plot (Seaborn) to visualize the distribution of churn (Yes vs. No). This reveals the overall churn rate and any potential imbalance in the data.
  1. **Categorical feature:**



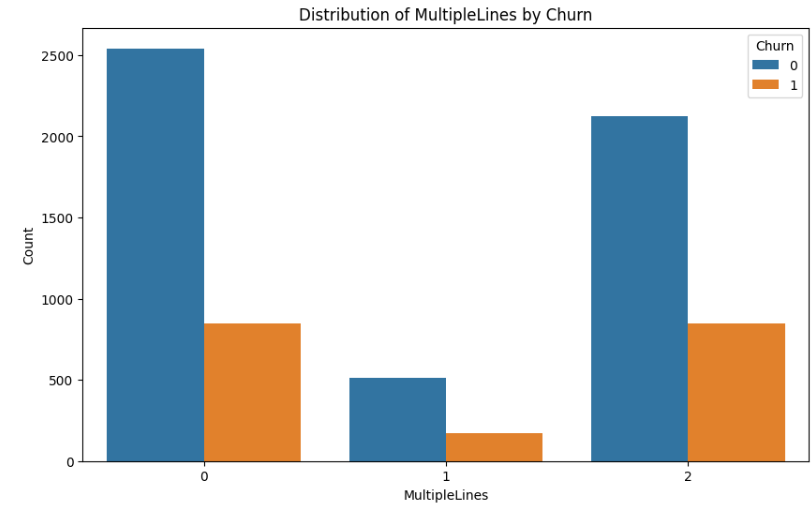
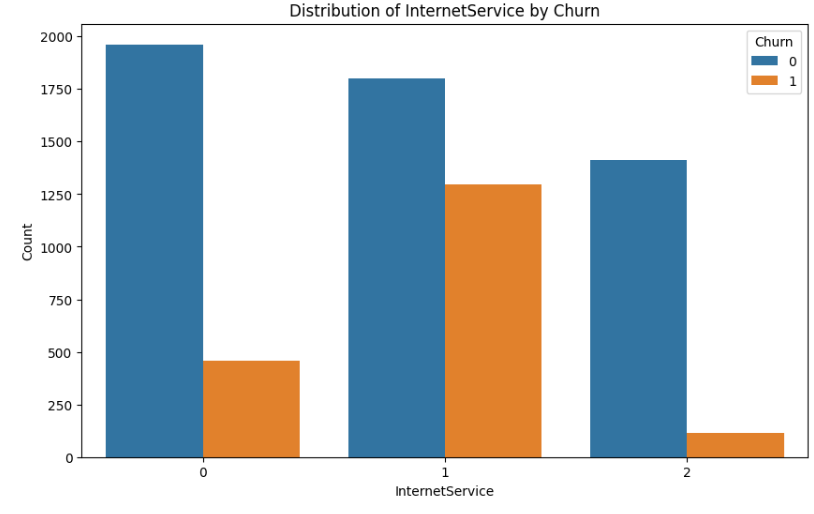
**Output**:

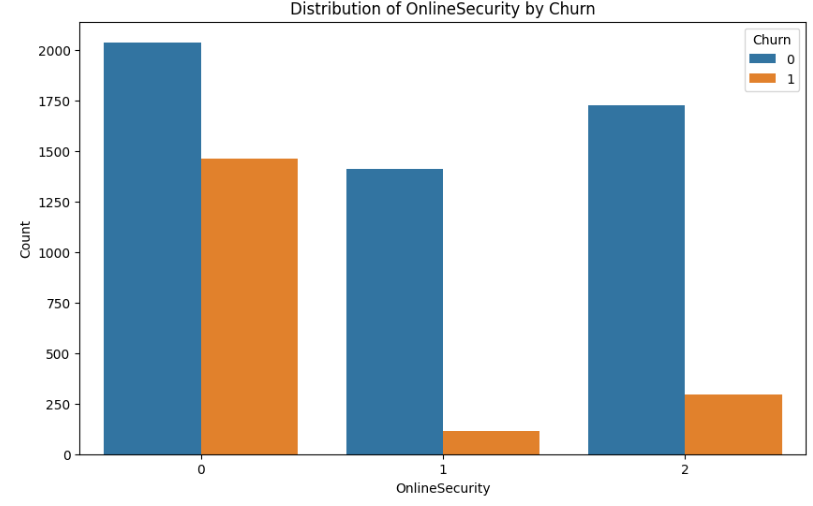
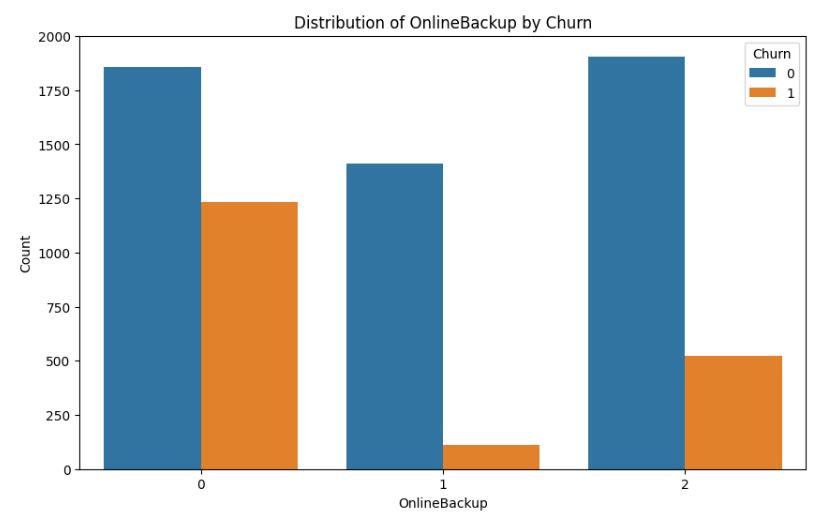
**Gender Partner**

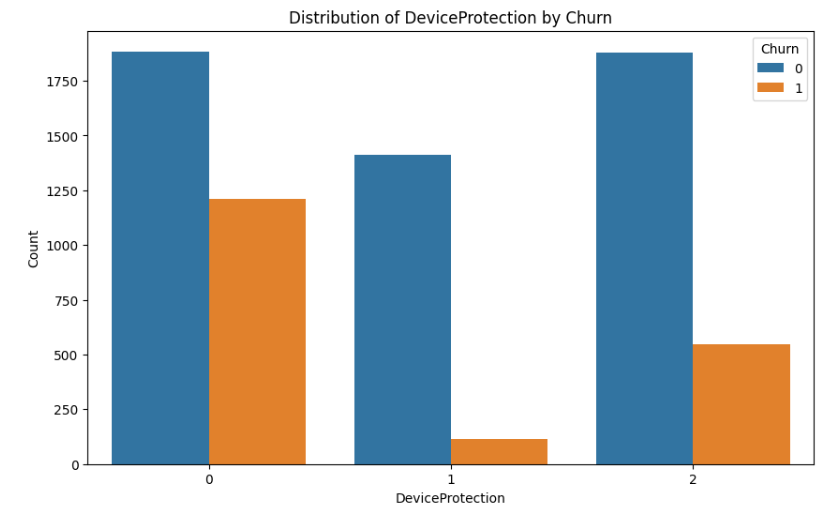
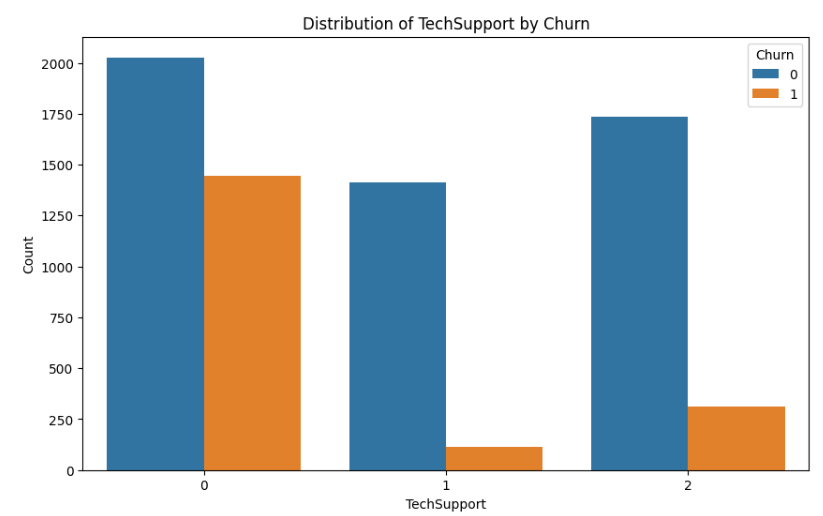
**Dependents PhoneService**

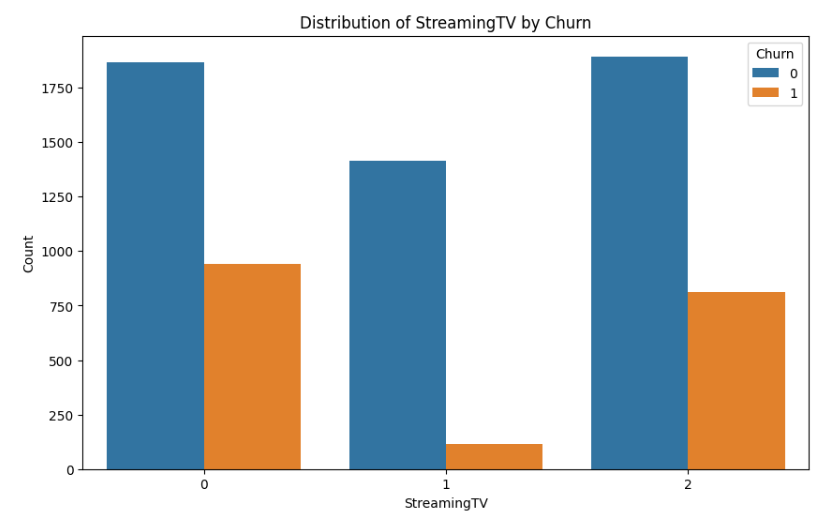
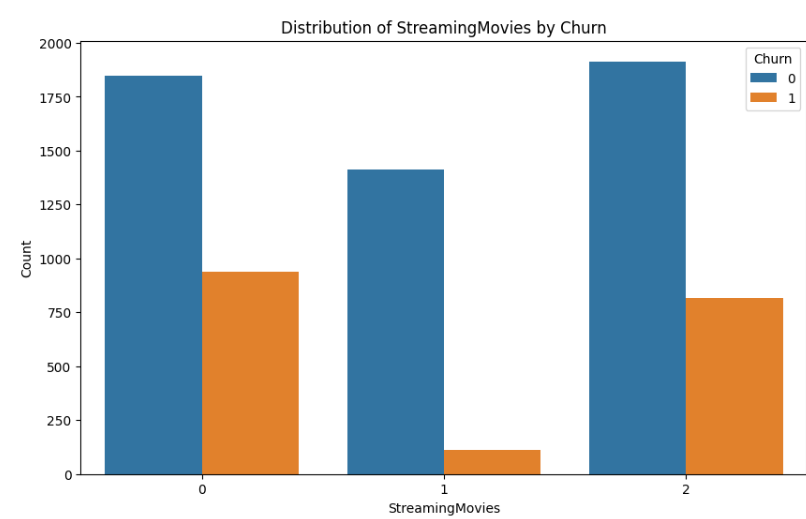
**MultipleLines InternetService**

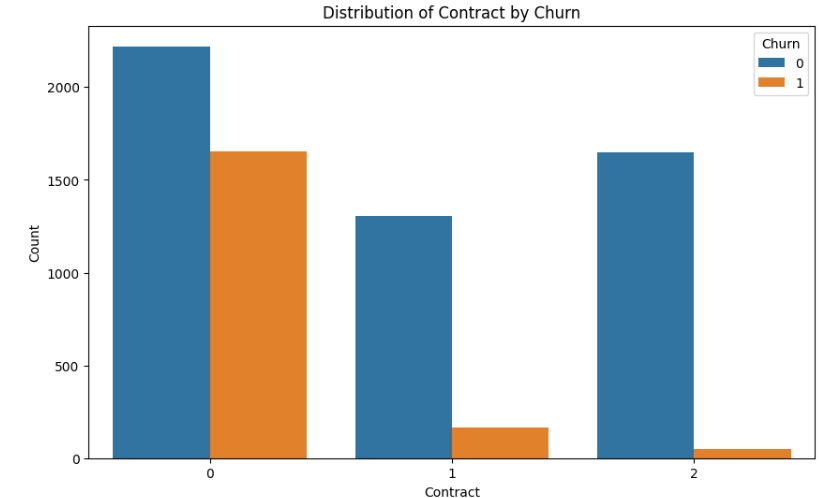
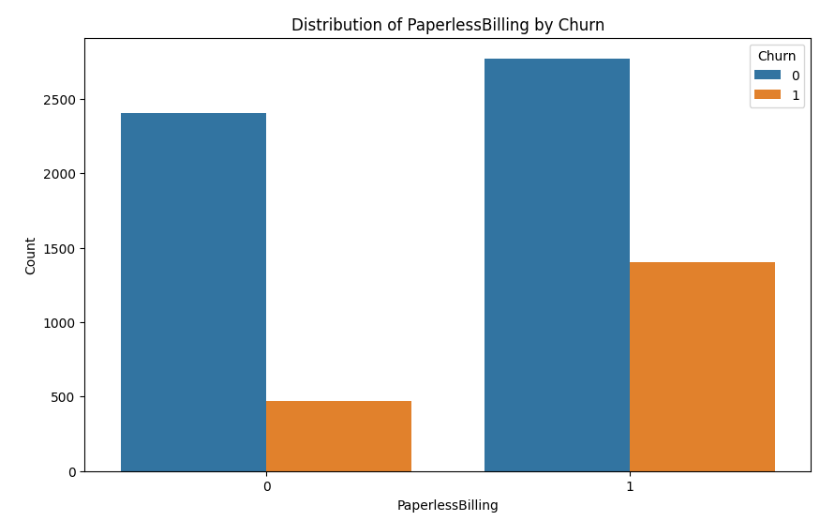
**OnlineSecurity OnlineBackup**

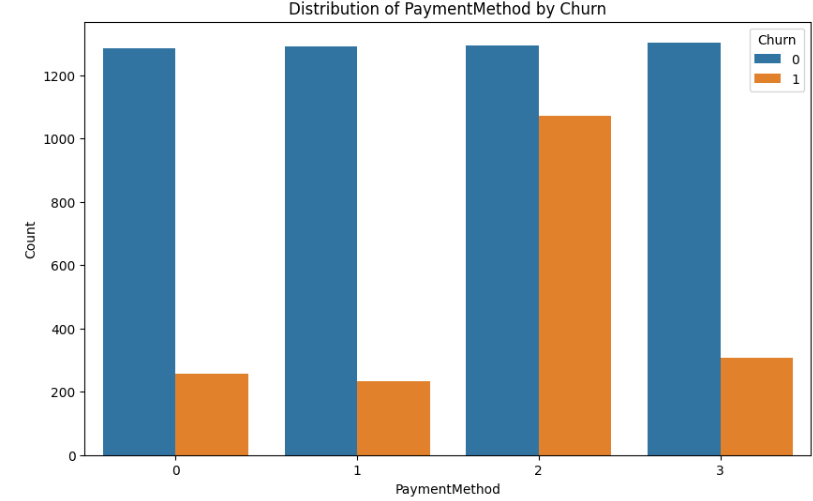
**DeviceProtection TechSupport**



**StreamingTV StreamingMovies**



**Contract PaperlessBilling**



**PayementMethod**

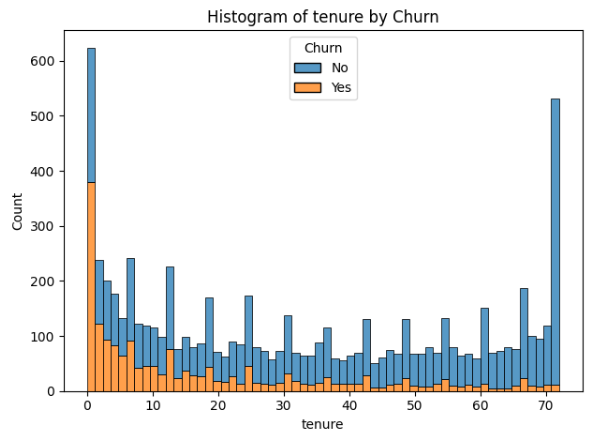
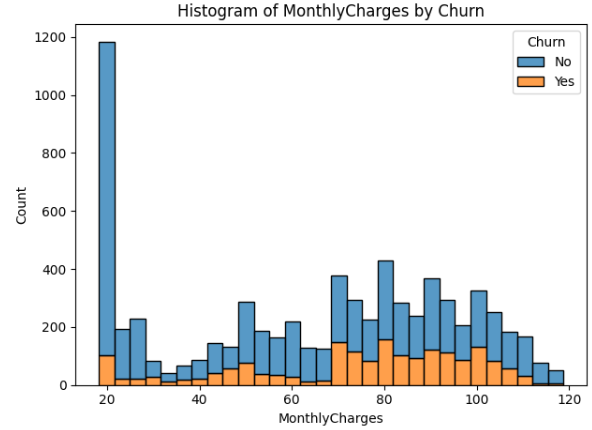
**Figure.** Feature of all categorical Features in the dataset.

1. **Numeric Features:**

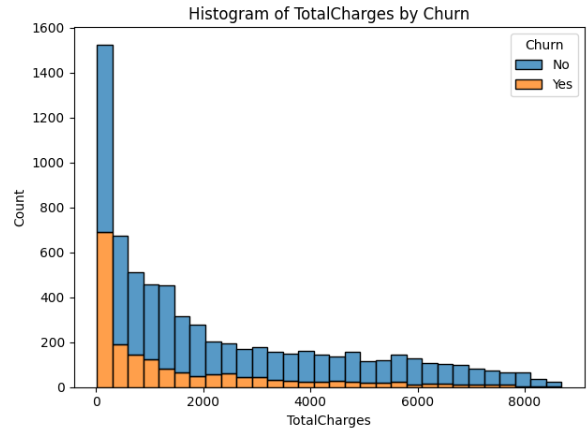
The distribution of numerical features within different churn categories (Yes vs. No) is crucial role to analyzing customer churn using a histogram graph



**Output:**

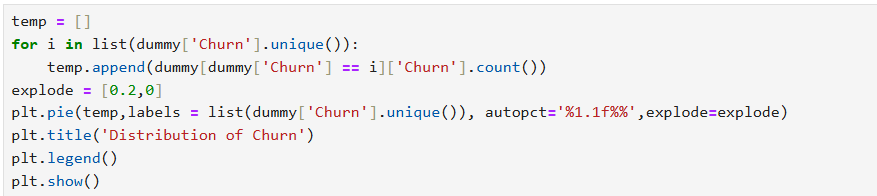
**Tenure MonthlyCharges**



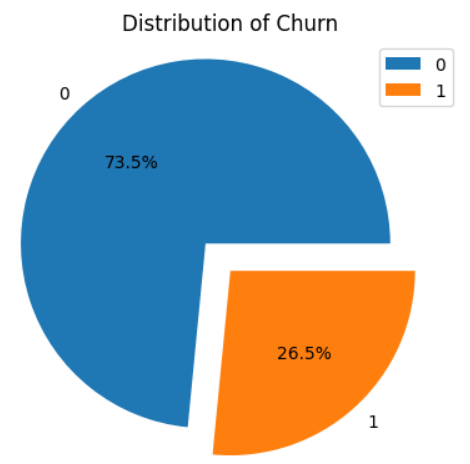
**TotalCharges**

**Figure.** Feature of all Numerical Features in the dataset.

1. **Overall Distribution of Target:**

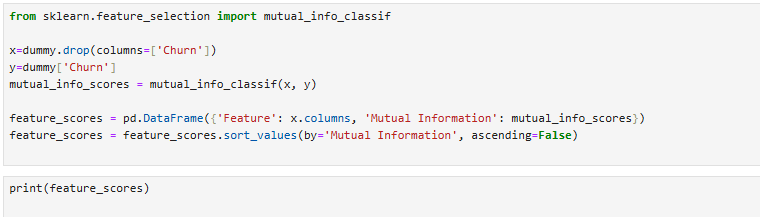


**Output:**

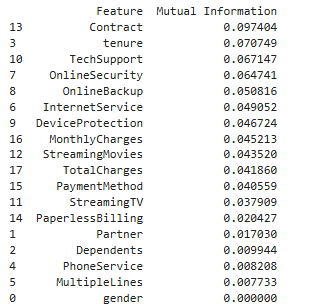


1. **Feature Selection:**

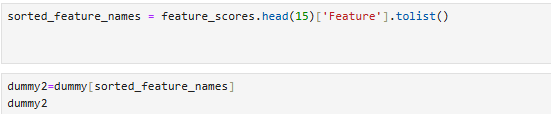
Mutual Information for feature selection is a powerful technique to identify and retain the most informative features for classification tasks. It helps in reducing dimensionality and improving model performance



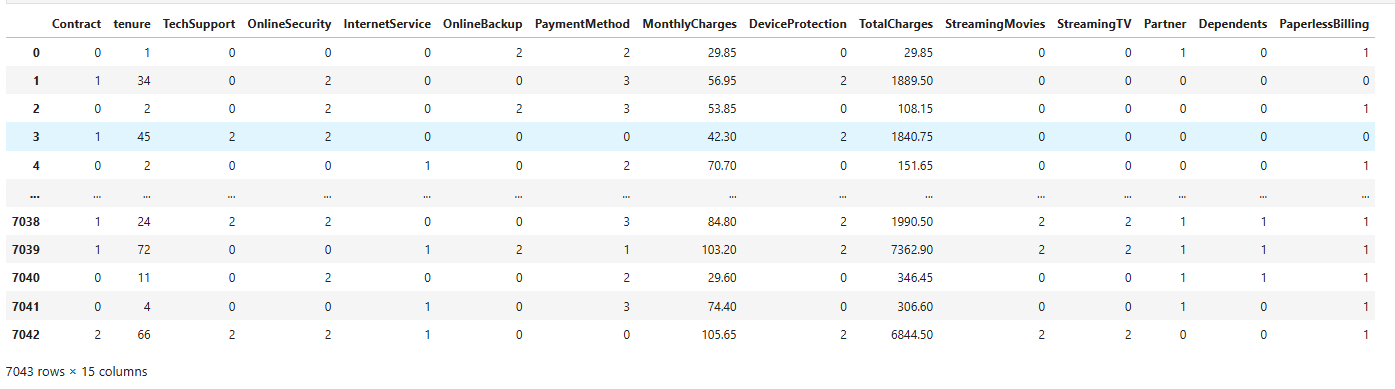
**Output:**



1. **Top 15 features to get:**



**Output:**



**7. Splitting Data into Training and Testing Sets:**



**Methodology:**

The goal of this study is to predict customer churn in a telecommunications company using three machine learning algorithms: Logistic Regression, Random Forest Classifier, and XGBoost Classifier. The methodology involves data preprocessing, feature selection, model training, evaluation, and comparison.

#### Model Training

**a. Logistic Regression:**

* Train a logistic regression model on the selected features. Logistic regression is a linear model suitable for binary classification tasks and provides interpretability of feature coefficients.

**b. Random Forest Classifier:**

* Train a random forest classifier, an ensemble learning method that combines multiple decision trees to improve prediction accuracy and control overfitting.

**c. XGBoost Classifier:**

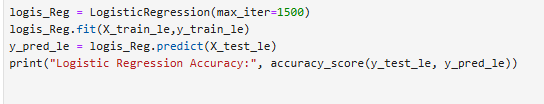
* Train an XGBoost classifier, a powerful gradient boosting algorithm known for its efficiency and performance in various machine learning competitions.

**Model Evaluation**

**a. Performance Metrics:**

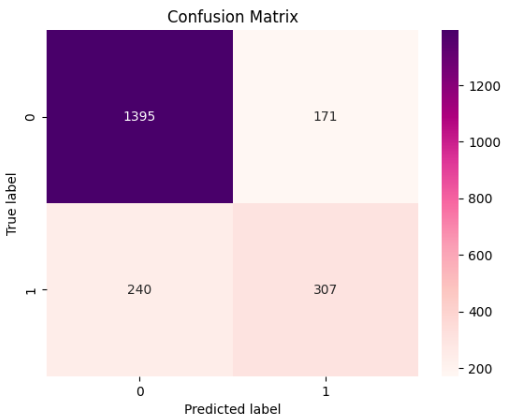
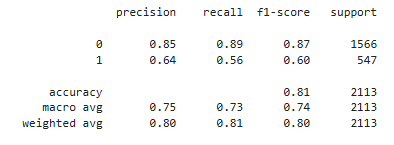
* Evaluate the performance of each model using the following metrics:
  + **Accuracy**: The ratio of correctly predicted instances to the total instances.
  + **Precision**: The ratio of true positive predictions to the total predicted positives.
  + **Recall**: The ratio of true positive predictions to the total actual positives.
  + **F1 Score**: The harmonic mean of precision and recall.

1. **Logistic regression:**

****

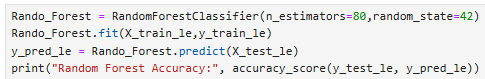
**Output:**

****

** **

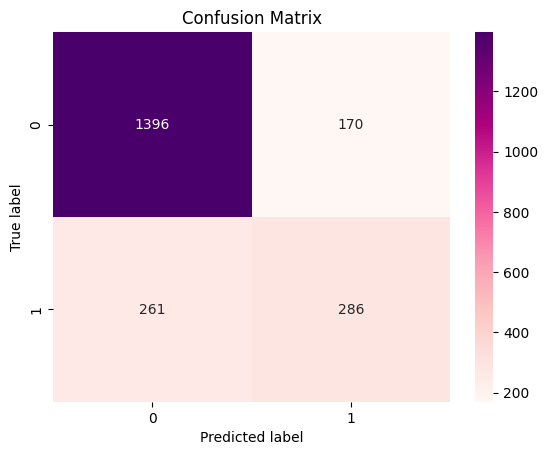
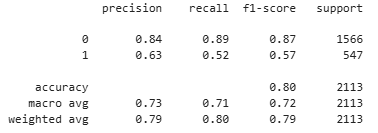
**Confusion Matrix Classification Report**

1. **Random Forest Classifier**

****

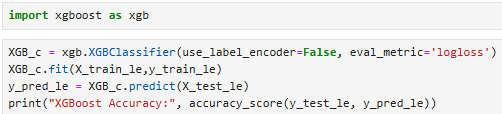
**Output:**

****

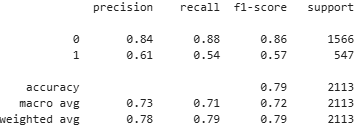
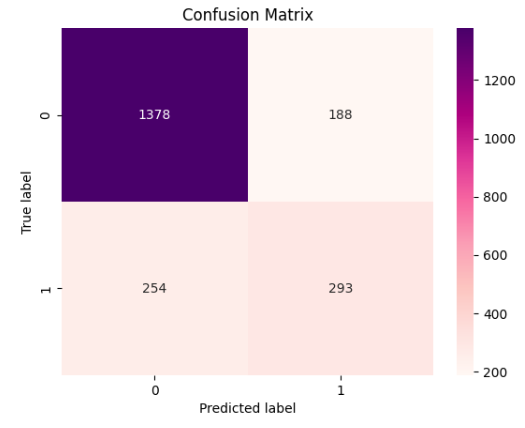
**Confusion Matrix** **Classification Report**

1. XgBoost Classifier:



**Output:**



   
 **Confusion Matrix Classification Report**

**Result:**

In this project, I have passed the telco customer churn dataset to the three different machine learning models and analyzed them with various metrics. The accuracy of these three different models was Logistic Regression 81%, Random Forest Classifier 79.6%, and XG Boost Classifier 79.08%. Among these models, the Logistic Regression model has produced higher accuracy with the80.5%. And also determined it with the confusion matrix and classification report. In future work, the dataset should be increased because the distribution of the target with the “NO” class is higher with a percentage of 75%. This may introduce biases in the data. So, the distribution of the dataset is equalized nearly to make the model more accurate.