Phase 5: Project Documentation & Submission

In this part you will document your project and prepare it for submission.
Document the customer churn prediction project and prepare it for submission.

Documentation

- Outline the project's objective, design thinking process, and development phases.
- Describe the analysis objectives, data collection process, data visualization using IBM Cognos, and predictive modeling.
- Explain how the insights and prediction model can help businesses reduce customer churn.

Submission

- Share the GitHub repository link containing the project's code and files.
- Provide instructions on how to replicate the analysis and generate visualizations using IBM Cognos and build the predictive model using Python.
- Include example outputs of the visualizations and model evaluation.

- IN THIS PHASE WE HAVE
 Outline the project's objective, design thinking process, and development phases.
- Describe the analysis objectives, data collection process, data visualization using IBM Cognos, and predictive modeling.

CUSTOMER CHURN P REDICTION

ABSTRACT

Customer churn prediction refers to the practice of using data analysis and predictive modeling techniques to forecast which customers are likely to stop using a product or service, often referred to as "churning" or "churned customers." Churn prediction is a valuable business strategy, especially for subscription-based services, telecom companies, e-commerce platforms, and other businesses that rely on customer retention and loyalty.

PROBLEM D EFINITION

The project involves using IBM Cognos to predict customer churn and identify f actors influencing customer retention. T he goal is to help businesses reduce cu stomer attrition by understanding the pa tterns and reasons behind customers le aving. This project includes defining an alysis objectives, collecting customer d ata, designing relevant visualizations in IBM   jCognos, and building a predictive model.

DESIGN THINKING

ANALYSIS OBJECTIVE
S

ANALYSIS OBJECTIVES

Define the specific objectives of predicting customer churn, such as identifying potential churners and understanding the key factors contributing to churn.

1.Identify Potential Churners

2.Early Detection

3.Reduce Churn Rate

1.The primary objective of churn prediction is to identify customers who are at risk of churning. This can be done by developing a predictive model that assigns a churn probability score to each customer

2.Aim to detect potential churners as early as possible. Early detection allows for proactive measures to be taken, such as targeted marketing campaigns or personalized incentives, to retain these customers.

3.Set a specific target for reducing the churn rate. This objective could be framed as a percentage reduction in churn over a specified time period (e.g., reduce churn by 10% in the next quarter).

4.Segmentation

5.Feature Analysis

6.Customer Lifetime Value (CLV)

4.Segment the customer base based on churn probabili ty and other relevant factors. This allows for tailored ret ention strategies for different customer groups. For exa mple, high-value customers may receive different retent ion efforts compared to low-value customers.

5.Understand the key factors contributing to churn. Con duct feature importance analysis to identify which custo mer attributes, behaviors, or interactions with the comp any have the most significant impact on churn.

6.Calculate CLV for each customer and analyze how it correlates with churn. The objective may be to increase the CLV of customers at risk of churning

7.Model Performance

8.Actionable Insights

9.Monitoring and Iterat ion

7.Set performance benchmarks for your churn predicti on model. This includes metrics such as accuracy, pre cision, recall, and F1-score. Aim to achieve a certain l evel of model accuracy in predicting churn.

8.The ultimate goal is to provide actionable insights to the business. Ensure that your churn prediction analy sis translates into specific actions that can be taken to retain customers. These actions may include sending targeted offers, improving customer service, or enhan cing product features.

9.Implement a system for continuous monitoring of ch urn and model performance. Establish a process for r egular model retraining and refinement to adapt to ch anging customer behaviors and market conditions.

10.Cost Reduction

11.Customer Feedback Inte gration:

12.Benchmarking

10.Evaluate the cost of customer acquisition compared to the cost of retaining customers. The objective may b e to reduce the cost of retention efforts while maximizin g their effectiveness.

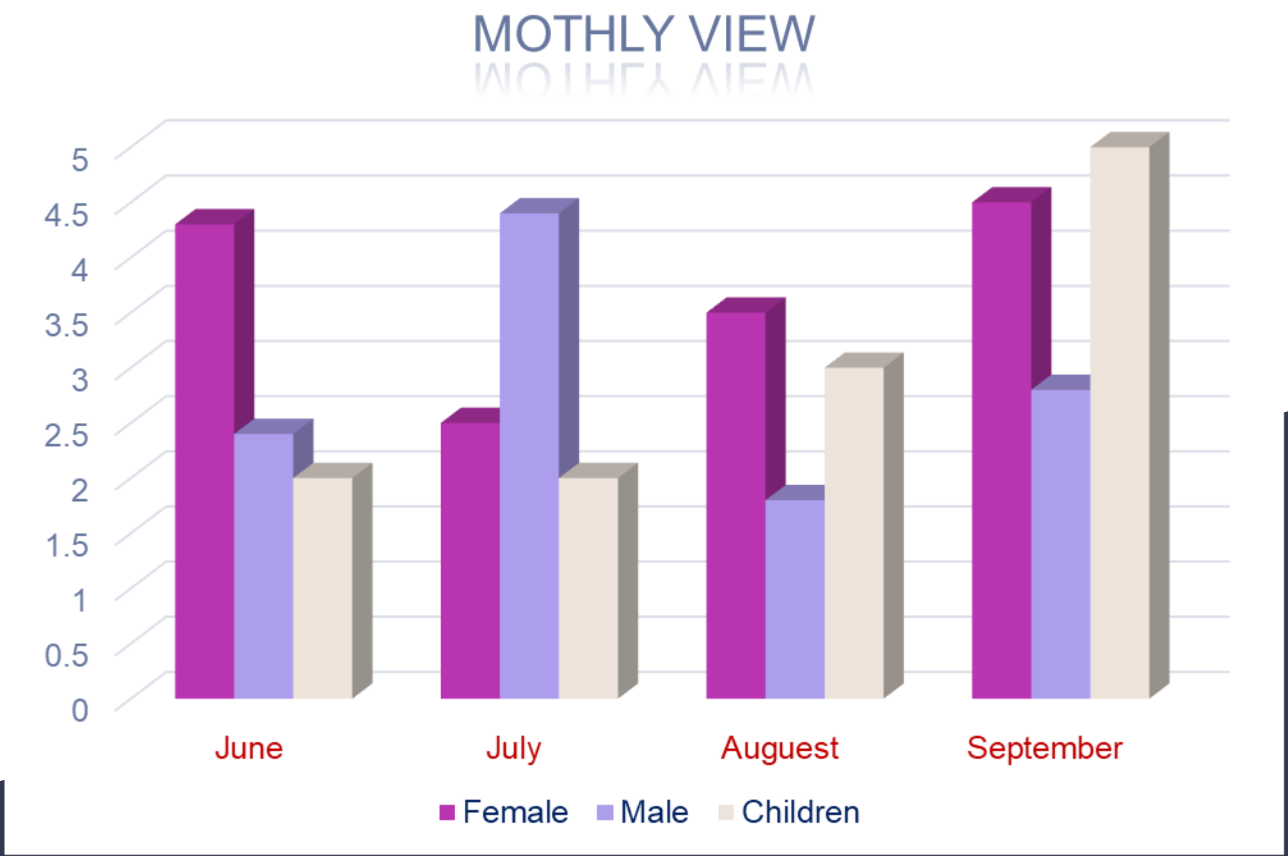
11.Integrate customer feedback into the churn predictio n process. Identify the sentiment of customer feedback from potential churners and use it to refine retention str ategies.

12.Compare your churn prediction and retention efforts with industry benchmarks or competitors to assess your performance and identify areas for improvement

DATA COLLECTION

Data Collection

Determine the sources and methods for collecting customer data, including customer demographics, usage behavior, and historical interactions.


MOTHLY VIEW
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
June
July
Auguest
September
Female
Male
Children

Methods for Collecting Customer Data:

1.Data Mining

2.Machine Learning Models

3.Third-party Data

1.Use data mining techniques to extract valuable insights from large datasets. This can help identify patterns and factors that contribute to customer churn.

2.Implement predictive models like logistic regression, decision trees, or neural networks to analyze historical data and predict future churn based on customer behavior and demographics.

3.Consider using external data sources, such as market data or industry benchmarks, to enhance your analysis and gain a broader perspective on customer behavior.

VISUALIZATION STR
ATEGY

Visualization Strategy

Plan how to visualize the insights using IBM Cognos, showcasing factors affecting churn and retention rates for customer churn prediction project

1.Understand the Data

2.Choose the Right Visualizations

1.Start by thoroughly understanding your dataset and the variables that may customer churn and retention. Identify key features and potential predictors.

2.Select appropriate visualization types for different types of data. For Example: Use line charts to visualize trends in churn and retention rates over Time. Create bar charts or pie charts to represent categorical variables like product usage, demographics, or subscription Type. Scatter plots can be useful to explore relationships between variables.

DATASET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	Strea
2	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No
3	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No
4	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No
5	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No
6	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No
7	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes
8	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No
9	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No	No
10	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes
11	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No
12	9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No
13	7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet servic	No internet servic	No internet servic	No internet servic	No internet servic	No in
14	8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes
15	0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes
16	5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes
17	3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes
18	8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet servic	No internet servic	No internet servic	No internet servic	No internet servic	No in
19	9959-WOFKT	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes
20	4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No
21	4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes
22	8779-QRDMV	Male	1	No	No	1	No	No phone service	DSL	No	No	Yes	No	No	Yes
23	1680-VDCWW	Male	0	Yes	No	12	Yes	No	No	No internet servic	No internet servic	No internet servic	No internet servic	No internet servic	No in
24	1066-JKSGK	Male	0	No	No	1	Yes	No	No	No internet servic	No internet servic	No internet servic	No internet servic	No internet servic	No in
25	3638-WEABW	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No
26	6322-HRPFA	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No
27	6865-JZNKO	Female	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No
28	6467-CHFZW	Male	0	Yes	Yes	47	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	Yes
29	8665-UTDHZ	Male	0	Yes	Yes	1	No	No phone service	DSL	No	Yes	No	No	No	No

DATA PREPROCESSING

VISUALIZATION

Check for missing values in each columns and
 decide how to handle them

Handle data types appropriately(eg.convert the
‘date’ column to datetime)

Ensure data consistency and correctness, such
as checking that percentages are within valid
Ranges(0-100%)

Develop informative and visually appealing charts
And graphs

Consider creating interactive visualization for
Online sharing or presentations

Ensure that your visualizations are well labled
And easy to interpret

PREDICTIVE M
ODELING

Algorithms to predict customer churn prediction such as ensemble techniques

1. SVM - SVM or Support Vector Machine
2. Ridge Classifier
3. Random Forest
4. XG boost

About the algorithms

SVM - SVM or Support Vector Machine is a supervised machine learning technique used for classification and regression. Finding a hyperplane in an N-dimensional space that classifies the data points is the goal of the SVM method. The number of features determines the hyperplane's size.

Ridge Classifier - Ridge classification is a metho d used in machine learning to assess linear discri minant models. In order to prevent overfitting, thi s type of normalization limits model coefficients.

Random Forest - Random Forest is a classification algorithm that uses multiple decision trees on smaller sets of the input dataset and averages the results to enhance the dataset's prediction accuracy.

XG Boost - Formally speaking, XGBoost may be described as a decision tree-based ensemble learning framework that uses Gradient Descent as the underlying objective function. It offers excellent flexibility and efficiently uses computation to produce the mandated results.

Conclusion

In conclusion, customer churn prediction plays a pivotal role in helping businesses retain their customers. By leveraging data-driven models a nd analytics, companies can identify potential c hurners and take proactive measures to retain t hem. This not only helps in maintaining revenu e but also enhances customer satisfaction and l oyalty.

Project title: Customer churn prediction

Phase 3: Development

Part 1

In this part you will begin building your project by loading and preprocessing the dataset.

Begin conducting the Customer churn prediction by collecting and preprocessing the data.

Collect and preprocess the Customer data for analysis.

Data Preprocessing:

- **Data preprocessing is a crucial step within the statistics analysis and gadget gaining knowledge of pipeline.**

- **It includes a sequence of strategies and operations finished on uncooked statistics to clean, organize, and transform it right into a layout that is suitable for analysis or device mastering version schooling.**

- **Data preprocessing goals to enhance the first-class of the records, making it greater reliable and conducive to generating accurate consequences.**

Here are some common tasks and techniques involved in data preprocessing:

Data Cleaning:

- **Handling missing values: Deciding how to deal with missing data, whether by imputing values or removing incomplete records.**

- **Outlier detection and treatment: Identifying and handling data points that significantly deviate from the norm.**

Noise reduction:

- Smoothing noisy data through techniques like filtering.

Data Transformation:

- Data normalization: Scaling numerical features to a standard range (e.g., between 0 and 1) to ensure that they have similar influence in the analysis.

- Encoding categorical variables: Converting categorical data into numerical format, such as one-hot encoding or label encoding.

- Feature engineering: Creating new features or modifying existing ones to capture more meaningful information from the data.

- Dimensionality reduction: Reducing the number of features while retaining essential information, using methods like Principal Component Analysis (PCA).

Data Integration:

- Merging or joining datasets: Combining data from multiple sources into a single dataset for analysis.

Aggregation: Summarizing data at a higher level of granularity, such as aggregating daily sales into monthly totals.

Data Reduction:

- Sampling: Reducing the size of a large dataset by randomly selecting a representative subset.

- Binning: Grouping continuous data into discrete bins to simplify analysis.

- Filtering: Selecting a subset of data based on specific criteria.

Data Standardization:

- Ensuring that data follows a consistent format and structure.

- Date and time format conversion: Converting date and time data into a uniform format.

- Currency conversion: Converting monetary values into a common currency.

Data Scaling:

- Scaling numerical data to a common range to prevent some features from dominating the analysis.

Data preprocessing is an iterative process that may involve several of these steps in various orders, depending on the specific dataset and the analysis goals. Proper data preprocessing is essential for improving the accuracy and effectiveness of machine learning models, as well as for making data more accessible for traditional statistical analysis.

Here is the data preprocessing codes along with the output of the given dataset:

Importing the libraries:

Import three basic libraries which are very common in machine learning and will be used every time you train a model

- NumPy: it is a library that allows us to work with arrays and as most machine learning models work on arrays NumPy makes it easier

- Matplotlib: this library helps in plotting graphs and charts, which are very useful while showing the result of your model

- Pandas: pandas allows us to import our dataset and also creates a matrix of features containing the dependent and independent variable.

```
#Connect the google drive for reading the
dataset # Connect the google drive
from google.colab import drive
drive.mount("/content/drive")
```

```
Mounted at /content/drive
```

```
# Preparing Dataset
# Import the dataset
import pandas as
pd
dataset = pd.read_csv("/content/drive/MyDrive/BIT/Customer-churn.csv")
```

```
print(dataset)
```

```
     customerID  gender  SeniorCitizen  Partner  Dependents  tenure
     PhoneService
0    7590-VHVEG  Female              0      Yes          No       1          No
1    5575-GNVDE    Male              0       No          No      34         Yes
2    3668-QPYBK    Male              0       No          No       2         Yes
3    7795-CFOCW    Male              0       No          No      45          No
4    9237-HQITU  Female              0       No          No       2         Yes
5    9305-CDSKC  Female              0       No          No       8         Yes
6    1452-KIOVK    Male              0       No         Yes      22         Yes
7    6713-OKOMC  Female              0       No          No      10          No
8    7892-POOKP  Female              0      Yes          No      28         Yes
9    6388-TABGU    Male              0       No         Yes      62         Yes
10   9763-GRSKD    Male              0      Yes         Yes      13         Yes
11   7469-LKBCI    Male              0       No          No      16         Yes
12   8091-TTVAX    Male              0      Yes          No      58         Yes
13   0280-XJGEX    Male              0       No          No      49         Yes
14   5129-JLPIS    Male              0       No          No      25         Yes
15   3655-SNQYZ  Female              0      Yes         Yes      69         Yes
16   8191-XWSZG  Female              0       No          No      52         Yes
17   9959-WOFKT    Male              0       No         Yes      71         Yes
18   4190-MFLUW  Female              0      Yes         Yes      10         Yes
19   4183-MYFRB  Female              0       No          No      21         Yes
20   8779-QRDMV    Male              1       No          No       1          No
21   1680-VDCWW    Male              0      Yes          No      12         Yes
22   1066-JKSGK    Male              0       No          No       1         Yes
23   3638-WEABW  Female              0      Yes          No      58         Yes
24   6322-HRPFA    Male              0      Yes         Yes      49         Yes
25   6865-JZNKO  Female              0       No          No      30         Yes
26   6467-CHFZW    Male              0      Yes         Yes      47         Yes
27   8665-UTDHZ    Male              0      Yes         Yes       1          No
28   5248-YGIJN    Male              0      Yes          No      72         Yes

        MultipleLines InternetService   OnlineSecurity  ...  \
0    No phone service             DSL               No  ...
1                  No             DSL              Yes  ...
2                  No             DSL              Yes  ...
3    No phone service             DSL              Yes  ...
4                  No     Fiber optic               No  ...
```

5	Yes	Fiber optic	No	...
6	Yes	Fiber optic	No	...
7	No phone service	DSL	Yes	...
8	Yes	Fiber optic	No	...
9	No	DSL	Yes	...
10	No	DSL	Yes	...
11	No	No	No internet service	...
12	Yes	Fiber optic	No	...
13	Yes	Fiber optic	No	...
14	No	Fiber optic	Yes	...
15	Yes	Fiber optic	Yes	...
16	No	No	No internet service	...
17	Yes	Fiber optic	Yes	...
18	No	DSL	No	...
19	No	Fiber optic	No	...
20	No phone service	DSL	No	...
21	No	No	No internet service	...
22	No	No	No internet service	...
23	Yes	DSL	No	...
24	No	DSL	Yes	...

dataset.dropna

<bound method DataFrame.dropna of customerID gender SeniorCitizen Partner Dependents tenure PhoneService \

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
0	7590-VHVEG	Female	0	Yes	No	1	No
1	5575-GNVDE	Male	0	No	No	34	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes
3	7795-CFOCW	Male	0	No	No	45	No
4	9237-HQITU	Female	0	No	No	2	Yes
5	9305-CDSKC	Female	0	No	No	8	Yes
6	1452-KIOVK	Male	0	No	Yes	22	Yes
7	6713-OKOMC	Female	0	No	No	10	No
8	7892-POOKP	Female	0	Yes	No	28	Yes
9	6388-TABGU	Male	0	No	Yes	62	Yes
10	9763-GRSKD	Male	0	Yes	Yes	13	Yes
11	7469-LKBCI	Male	0	No	No	16	Yes
12	8091-TTVAX	Male	0	Yes	No	58	Yes
13	0280-XJGEX	Male	0	No	No	49	Yes
14	5129-JLPIS	Male	0	No	No	25	Yes
15	3655-SNQYZ	Female	0	Yes	Yes	69	Yes
16	8191-XWSZG	Female	0	No	No	52	Yes
17	9959-WOFKT	Male	0	No	Yes	71	Yes
18	4190-MFLUW	Female	0	Yes	Yes	10	Yes
19	4183-MYFRB	Female	0	No	No	21	Yes
20	8779-QRDMV	Male	1	No	No	1	No
21	1680-VDCWW	Male	0	Yes	No	12	Yes
22	1066-JKSGK	Male	0	No	No	1	Yes
23	3638-WEABW	Female	0	Yes	No	58	Yes
24	6322-HRPFA	Male	0	Yes	Yes	49	Yes
25	6865-JZNKO	Female	0	No	No	30	Yes
26	6467-CHFZW	Male	0	Yes	Yes	47	Yes
27	8665-UTDHZ	Male	0	Yes	Yes	1	No
28	5248-YGIJN	Male	0	Yes	No	72	Yes

	MultipleLines	InternetService	OnlineSecurity	...	\
0	No phone service	DSL	No	...	
1	No	DSL	Yes	...	
2	No	DSL	Yes	...	

3	No phone service	DSL	Yes	...
4	No	Fiber optic	No	...
5	Yes	Fiber optic	No	...
6	Yes	Fiber optic	No	...
7	No phone service	DSL	Yes	...
8	Yes	Fiber optic	No	...
9	No	DSL	Yes	...
10	No	DSL	Yes	...
11	No	No	No internet service	...
12	Yes	Fiber optic	No	...
13	Yes	Fiber optic	No	...
14	No	Fiber optic	Yes	...
15	Yes	Fiber optic	Yes	...
16	No	No	No internet service	...
17	Yes	Fiber optic	Yes	...
18	No	DSL	No	...
19	No	Fiber optic	No	...
20	No phone service	DSL	No	...
21	No	No	No internet service	...
22	No	No	No internet service	...
23	Yes	DSL	No	...
24	No	DSL	Yes	

dataset.isnull()

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneServi
0	False	False	False	False	False	False	Fal
1	False	False	False	False	False	False	Fal
2	False	False	False	False	False	False	Fal
3	False	False	False	False	False	False	Fal
4	False	False	False	False	False	False	Fal
5	False	False	False	False	False	False	Fal
6	False	False	False	False	False	False	Fal
7	False	False	False	False	False	False	Fal
8	False	False	False	False	False	False	Fal
9	False	False	False	False	False	False	Fal
10	False	False	False	False	False	False	Fal
11	False	False	False	False	False	False	Fal

dataset.info

```
<bound method DataFrame.info of                customerID      gender
          SeniorCitizen Partner Dependents    tenure PhoneService \
0    7590-VHVEG  Female          0     Yes         No       1           No
1    5575-GNVDE    Male          0      No         No      34          Yes
2    3668-QPYBK    Male          0      No         No       2          Yes
3    7795-CFOCW    Male          0      No         No      45           No
4    9237-HQITU  Female          0      No         No       2          Yes
5    9305-CDSKC  Female          0      No         No       8          Yes
6    1452-KIOVK    Male          0      No        Yes      22          Yes
7    6713-OKOMC  Female          0      No         No      10           No
8    7892-POOKP  Female          0     Yes         No      28          Yes
9    6388-TABGU    Male          0      No        Yes      62          Yes
10   9763-GRSKD    Male          0     Yes        Yes      13          Yes
11   7469-LKBCI    Male          0      No         No      16          Yes
12   8091-TTVAX    Male          0     Yes         No      58          Yes
13   0280-XJGEX    Male          0      No         No      49          Yes
14   5129-JLPIS    Male          0      No         No      25          Yes
15   3655-SNQYZ  Female          0     Yes        Yes      69          Yes
16   8191-XWSZG  Female          0      No         No      52          Yes
17   9959-WOFKT    Male          0      No        Yes      71          Yes
18   4190-MFLUW  Female          0     Yes        Yes      10          Yes
19   4183-MYFRB  Female          0      No         No      21          Yes
20   8779-QRDMV    Male          1      No         No       1           No
21   1680-VDCWW    Male          0     Yes         No      12          Yes
22   1066-JKSGK    Male          0      No         No       1          Yes
23   3638-WEABW  Female          0     Yes         No      58          Yes
24   6322-HRPFA    Male          0     Yes        Yes      49          Yes
25   6865-JZNKO  Female          0      No         No      30          Yes
26   6467-CHFZW    Male          0     Yes        Yes      47          Yes
27   8665-UTDHZ    Male          0     Yes        Yes       1           No
28   5248-YGIJN    Male          0     Yes         No      72          Yes

        MultipleLines InternetService      OnlineSecurity    ...   \
0    No phone service             DSL                  No  ...
```

1	No	DSL	Yes	...
2	No	DSL	Yes	...
3	No phone service	DSL	Yes	...
4	No	Fiber optic	No	...
5	Yes	Fiber optic	No	...
6	Yes	Fiber optic	No	...
7	No phone service	DSL	Yes	...
8	Yes	Fiber optic	No	...
9	No	DSL	Yes	...
10	No	DSL	Yes	...
11	No	No	No internet service	...
12	Yes	Fiber optic	No	...
13	Yes	Fiber optic	No	...
14	No	Fiber optic	Yes	...
15	Yes	Fiber optic	Yes	...
16	No	No	No internet service	...
17	Yes	Fiber optic	Yes	...
18	No	DSL	No	...
19	No	Fiber optic	No	...
20	No phone service	DSL	No	...
21	No	No	No internet service	...
22	No	No	No internet service	...
23	Yes	DSL	No	...
24	No	DSL	Yes	...

dataset.describe

<bound method NDFrame.describe of customerID gender SeniorCitizen Partner Dependents tenure PhoneService \

0	7590-VHVEG	Female	0	Yes	No	1	No
1	5575-GNVDE	Male	0	No	No	34	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes
3	7795-CFOCW	Male	0	No	No	45	No
4	9237-HQITU	Female	0	No	No	2	Yes
5	9305-CDSKC	Female	0	No	No	8	Yes
6	1452-KIOVK	Male	0	No	Yes	22	Yes
7	6713-OKOMC	Female	0	No	No	10	No
8	7892-POOKP	Female	0	Yes	No	28	Yes
9	6388-TABGU	Male	0	No	Yes	62	Yes
10	9763-GRSKD	Male	0	Yes	Yes	13	Yes
11	7469-LKBCI	Male	0	No	No	16	Yes
12	8091-TTVAX	Male	0	Yes	No	58	Yes
13	0280-XJGEX	Male	0	No	No	49	Yes
14	5129-JLPIS	Male	0	No	No	25	Yes
15	3655-SNQYZ	Female	0	Yes	Yes	69	Yes
16	8191-XWSZG	Female	0	No	No	52	Yes
17	9959-WOFKT	Male	0	No	Yes	71	Yes
18	4190-MFLUW	Female	0	Yes	Yes	10	Yes
19	4183-MYFRB	Female	0	No	No	21	Yes
20	8779-QRDMV	Male	1	No	No	1	No
21	1680-VDCWW	Male	0	Yes	No	12	Yes
22	1066-JKSGK	Male	0	No	No	1	Yes
23	3638-WEABW	Female	0	Yes	No	58	Yes
24	6322-HRPFA	Male	0	Yes	Yes	49	Yes
25	6865-JZNKO	Female	0	No	No	30	Yes

26	6467-CHFZW	Male	0	Yes	Yes	47	Yes
27	8665-UTDHZ	Male	0	Yes	Yes	1	No
28	5248-YGIJN	Male	0	Yes	No	72	Yes

```
      MultipleLines InternetService       OnlineSecurity  ...  \
0  No phone service             DSL                   No  ...
1                No             DSL                  Yes  ...
2                No             DSL                  Yes  ...
3  No phone service             DSL                  Yes  ...
4                No     Fiber optic                   No  ...
5               Yes     Fiber optic                   No  ...
6               Yes     Fiber optic                   No  ...
7  No phone service             DSL                  Yes  ...
8               Yes     Fiber optic                   No  ...
9                No             DSL                  Yes  ...
10               No             DSL                  Yes  ...
11               No              No  No internet service  ...
12              Yes     Fiber optic                   No  ...
13              Yes     Fiber optic                   No  ...
14               No     Fiber optic                  Yes  ...
15              Yes     Fiber optic                  Yes  ...
16               No              No  No internet service  ...
17              Yes     Fiber optic                  Yes  ...
18               No             DSL                   No  ...
19               No     Fiber optic                   No  ...
20 No phone service             DSL                   No  ...
21               No              No  No internet service  ...
22               No              No  No internet service  ...
23              Yes             DSL                   No  ...
```

```
import matplotlib.pyplot as plt

X=dataset.MonthlyCharges

Y=dataset.TotalCharges

Xtrain = dataset[['gender','PaymentMethod','OnlineBackup','PaperlessBilling']]

Ytrain = dataset[['Churn']]

print(Xtrain)
```

```
    gender              PaymentMethod OnlineBackup PaperlessBilling
0   Female           Electronic check          Yes              Yes
1     Male               Mailed check           No               No
2     Male               Mailed check          Yes              Yes
3     Male  Bank transfer (automatic)           No               No
4   Female           Electronic check           No              Yes
5   Female           Electronic check           No              Yes
6     Male    Credit card (automatic)          Yes              Yes
7   Female               Mailed check           No               No
8   Female           Electronic check           No              Yes
9     Male  Bank transfer (automatic)          Yes               No
10    Male               Mailed check           No              Yes
```

11	Male	Credit card (automatic)	No internet service	No
12	Male	Credit card (automatic)	No	No
13	Male	Bank transfer (automatic)	Yes	Yes
14	Male	Electronic check	No	Yes

```
15  Female    Credit card (automatic)                  Yes          No
16  Female               Mailed check  No internet service          No
17    Male  Bank transfer (automatic)                   No          No
18  Female    Credit card (automatic)                   No          No
19  Female           Electronic check                  Yes         Yes
20    Male           Electronic check                   No         Yes
21    Male  Bank transfer (automatic)  No internet service          No
22    Male               Mailed check  No internet service          No
23  Female    Credit card (automatic)                  Yes         Yes
24    Male    Credit card (automatic)                  Yes          No
25  Female  Bank transfer (automatic)                  Yes         Yes
26    Male           Electronic check                  Yes         Yes
27    Male           Electronic check                  Yes          No
28    Male    Credit card (automatic)                  Yes         Yes
```

```
print(Ytrain)
```

```
   Churn
0     No
1     No
2    Yes
3     No
4    Yes
5    Yes
6     No
7     No
8    Yes
9     No
10    No
11    No
12    No
13   Yes
14    No
15    No
16    No
17    No
18   Yes
19    No
20   Yes
21    No
22   Yes
23    No
24    No
25    No
26   Yes
27   Yes
28    No
```

```
from sklearn.preprocessing import OrdinalEncoder

enc = OrdinalEncoder()

enc.fit(Xtrain)
```

▾ OrdinalEncoder

```
Xtrain_encoded=enc.transform(Xtrain)

print(Xtrain_encoded)
```

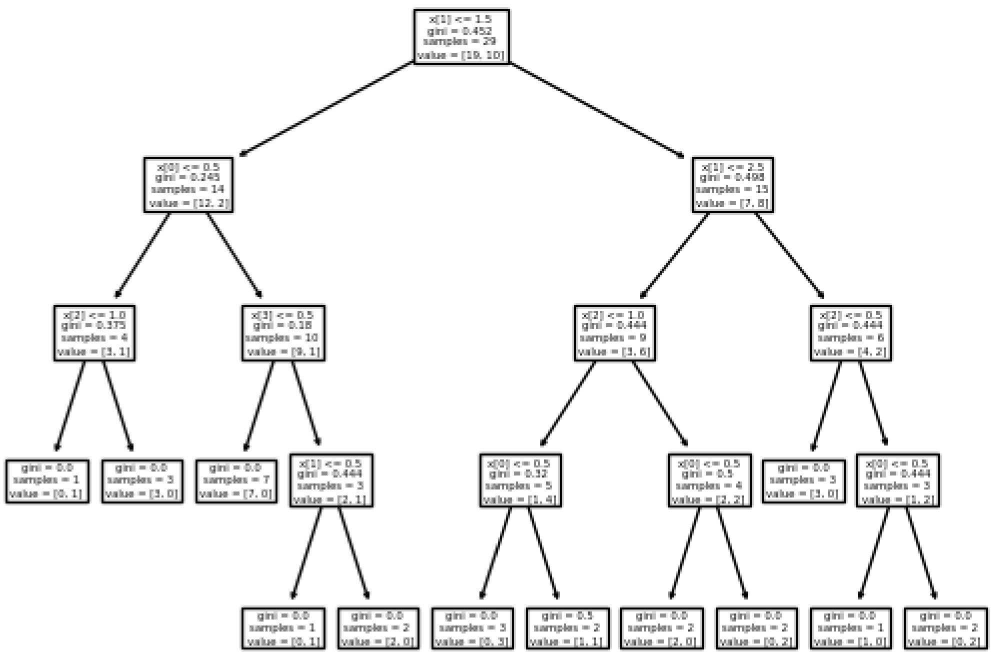
```
[[0. 2. 2. 1.]
 [1. 3. 0. 0.]
 [1. 3. 2. 1.]
 [1. 0. 0. 0.]
 [0. 2. 0. 1.]
 [0. 2. 0. 1.]
 [1. 1. 2. 1.]
 [0. 3. 0. 0.]
 [0. 2. 0. 1.]
 [1. 0. 2. 0.]
 [1. 3. 0. 1.]
 [1. 1. 1. 0.]
 [1. 1. 0. 0.]
 [1. 0. 2. 1.]
 [1. 2. 0. 1.]
 [0. 1. 2. 0.]
 [0. 3. 1. 0.]
 [1. 0. 0. 0.]
 [0. 1. 0. 0.]
 [0. 2. 2. 1.]
 [1. 2. 0. 1.]
 [1. 0. 1. 0.]
 [1. 3. 1. 0.]
 [0. 1. 2. 1.]
 [1. 1. 2. 0.]
 [0. 0. 2. 1.]
 [1. 2. 2. 1.]
 [1. 2. 2. 0.]
 [1. 1. 2. 1.]]
```

```
from sklearn import tree

clf = tree.DecisionTreeClassifier()

clf.fit(Xtrain_encoded,Ytrain)
```

▾ DecisionTreeClassifier
DecisionTreeClassifier()

```
tree.plot_tree(clf)
```

```
[Text(0.4642857142857143, 0.9, 'x[1] <= 1.5\ngini = 0.452\nsamples =
29\nvalue = [19, 10]'),
 Text(0.19047619047619047, 0.7, 'x[0] <= 0.5\ngini = 0.245\nsamples =
14\nvalue = [12, 2]'),
 Text(0.09523809523809523, 0.5, 'x[2] <= 1.0\ngini = 0.375\nsamples =
4\nvalue = [3, 1]'),
 Text(0.047619047619047616, 0.3, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(0.14285714285714285, 0.3, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
 Text(0.2857142857142857, 0.5, 'x[3] <= 0.5\ngini = 0.18\nsamples = 10\nvalue
= [9, 1]'),
 Text(0.23809523809523808, 0.3, 'gini = 0.0\nsamples = 7\nvalue = [7, 0]'),
 Text(0.3333333333333333, 0.3, 'x[1] <= 0.5\ngini = 0.444\nsamples = 3\nvalue
= [2, 1]'),
 Text(0.2857142857142857, 0.1, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(0.38095238095238093, 0.1, 'gini = 0.0\nsamples = 2\nvalue = [2, 0]'),
 Text(0.7380952380952381, 0.7, 'x[1] <= 2.5\ngini = 0.498\nsamples =
15\nvalue = [7, 8]'),
 Text(0.6190476190476191, 0.5, 'x[2] <= 1.0\ngini = 0.444\nsamples = 9\nvalue
= [3, 6]'),
 Text(0.5238095238095238, 0.3, 'x[0] <= 0.5\ngini = 0.32\nsamples = 5\nvalue
= [1, 4]'),
 Text(0.47619047619047616, 0.1, 'gini = 0.0\nsamples = 3\nvalue = [0, 3]'),
 Text(0.5714285714285714, 0.1, 'gini = 0.5\nsamples = 2\nvalue = [1, 1]'),
 Text(0.7142857142857143, 0.3, 'x[0] <= 0.5\ngini = 0.5\nsamples = 4\nvalue =
[2, 2]'),
 Text(0.6666666666666666, 0.1, 'gini = 0.0\nsamples = 2\nvalue = [2, 0]'),
 Text(0.7619047619047619, 0.1, 'gini = 0.0\nsamples = 2\nvalue = [0, 2]'),
 Text(0.8571428571428571, 0.5, 'x[2] <= 0.5\ngini = 0.444\nsamples = 6\nvalue
= [4, 2]'),
 Text(0.8095238095238095, 0.3, 'gini = 0.0\nsamples = 3\nvalue = [3, 0]'),
 Text(0.9047619047619048, 0.3, 'x[0] <= 0.5\ngini = 0.444\nsamples = 3\nvalue
= [1, 2]'),
 Text(0.8571428571428571, 0.1, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
 Text(0.9523809523809523, 0.1, 'gini = 0.0\nsamples = 2\nvalue = [0, 2]')]
```



```
from sklearn.ensemble import
RandomForestClassifier clf =
RandomForestClassifier(n_estimators = 100)
clf.fit(Xtrain_encoded,Ytrain)
```

```
<ipython-input-20-b6cd1249641e>:3: DataConversionWarning: A column-vector y
  w clf.fit(Xtrain_encoded,Ytrain)
```

```
▾ RandomForestClassifier
RandomForestClassifier()
```

```
import numpy as np
arr = np.array([[1, 1, 2, 1]])
print(clf.predict(arr))
```

```
['No']
```

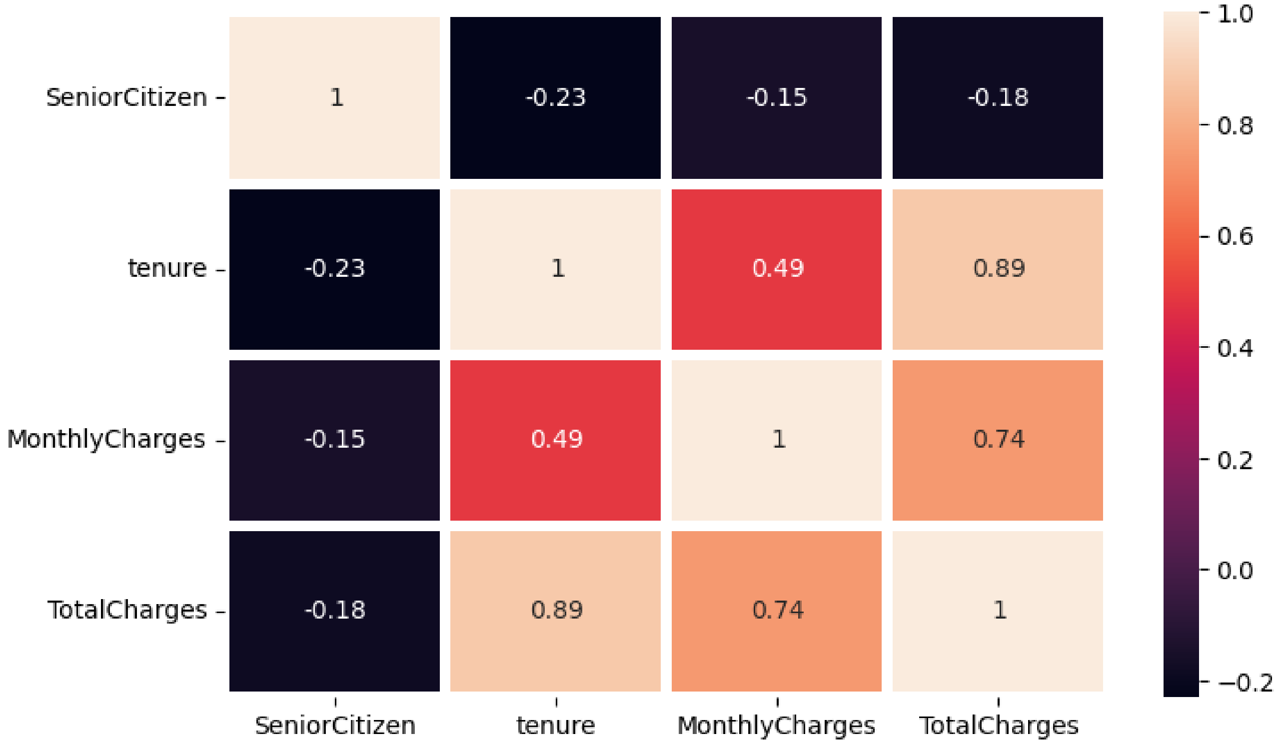
```
import numpy as np
arr1 = np.array([[1, 3, 2, 1]])
print(clf.predict(arr1))
```

```
['Yes']
```

```
import seaborn as sns
```

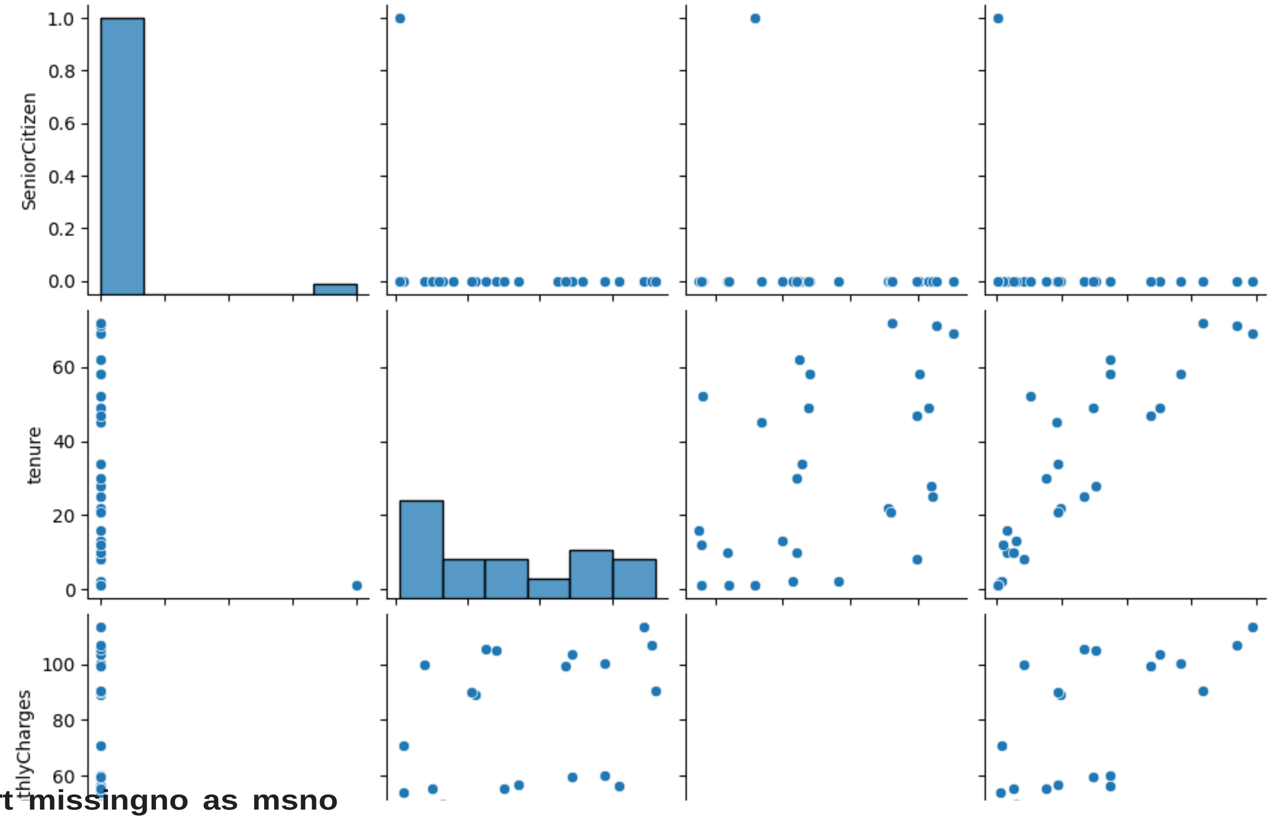
```
plt.figure(figsize=(8,5))
sns.heatmap(dataset.corr(),annot=True,linewidth=
3) plt.show
```

```
<ipython-input-25-095c9657f905>:2: FutureWarning: The default value of numeri
  sns.heatmap(dataset.corr(),annot=True,linewidth=3)
<function matplotlib.pyplot.show(close=None, block=None)>
```

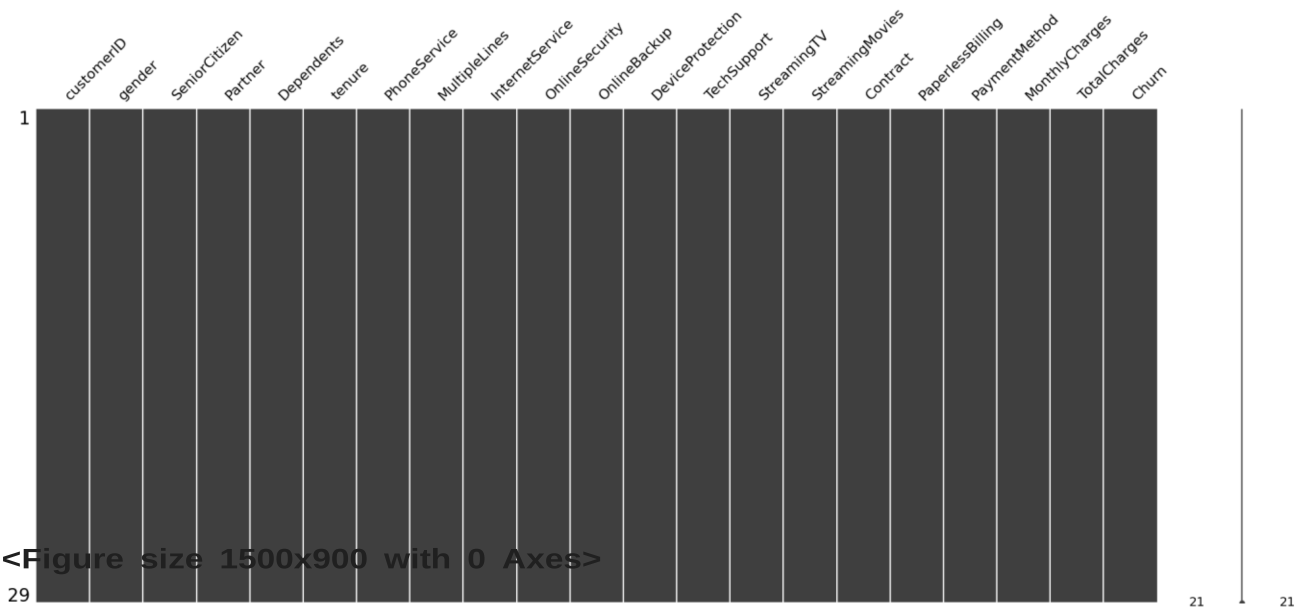



```
sns.pairplot(dataset)
```

<seaborn.axisgrid.PairGrid at 0x7813753a3e80>



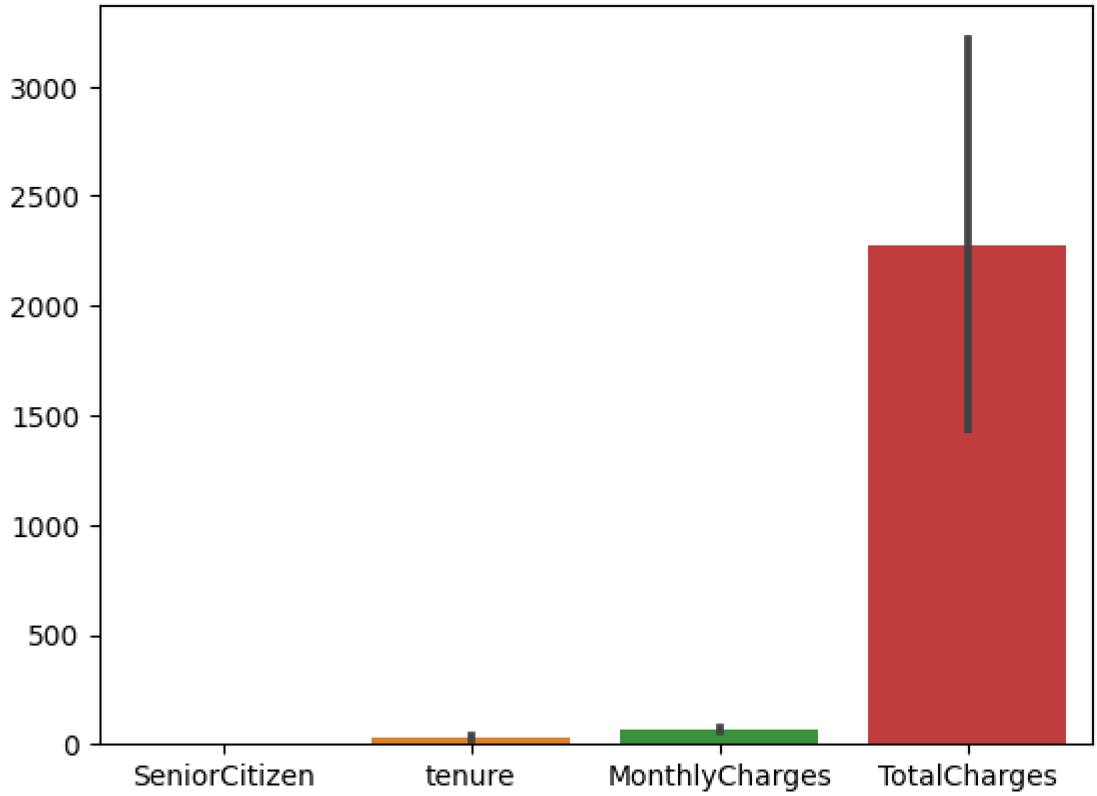
```
import missingno as msno
msno.matrix(dataset)
plt.figure(figsize=(15,9))
plt.show()
```



<Figure size 1500x900 with 0 Axes>

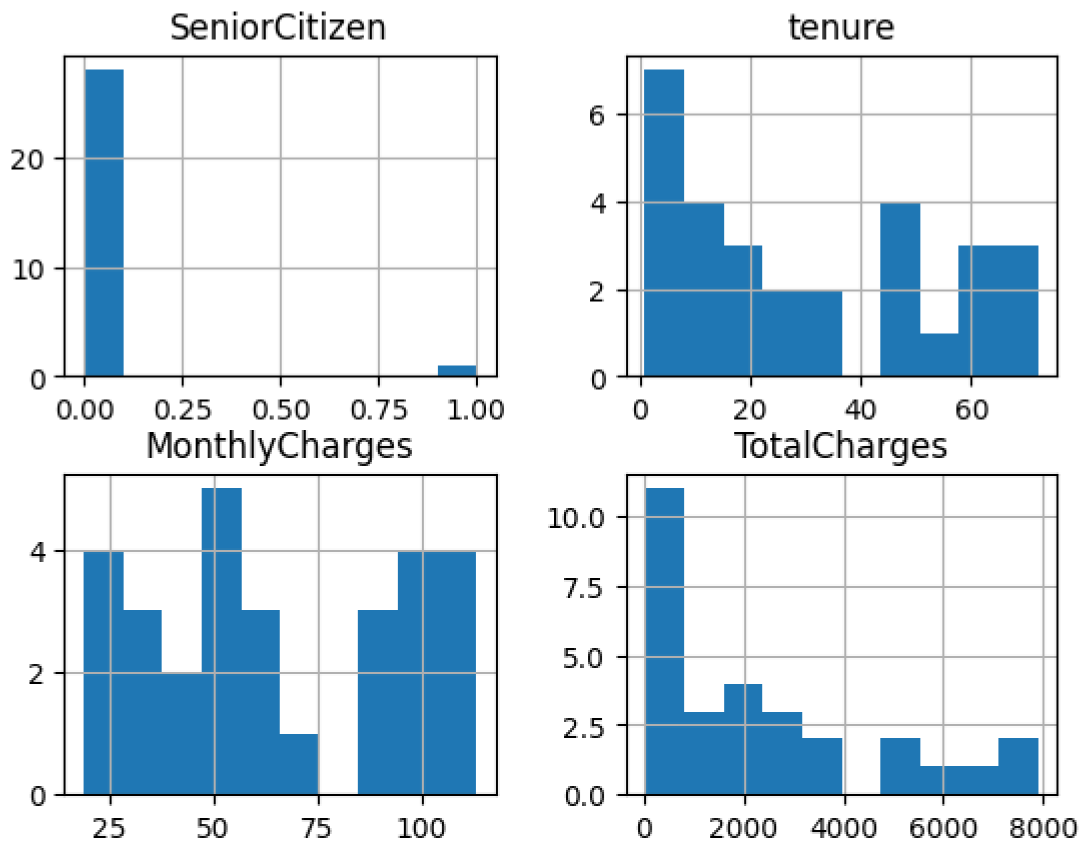
```
sns.barplot(dataset)
```

<Axes: >



dataset.hist()

```
array([[<Axes: title={'center': 'SeniorCitizen'}>,
        <Axes: title={'center': 'tenure'}>],
       [<Axes: title={'center': 'MonthlyCharges'}>,
        <Axes: title={'center': 'TotalCharges'}>]], dtype=object)
```



PROJECT TITLE :CUSTOMER CHURN PREDICTION

PHASE 4: DEVELOPMENT PART 2

PROBLEM STATEMENT

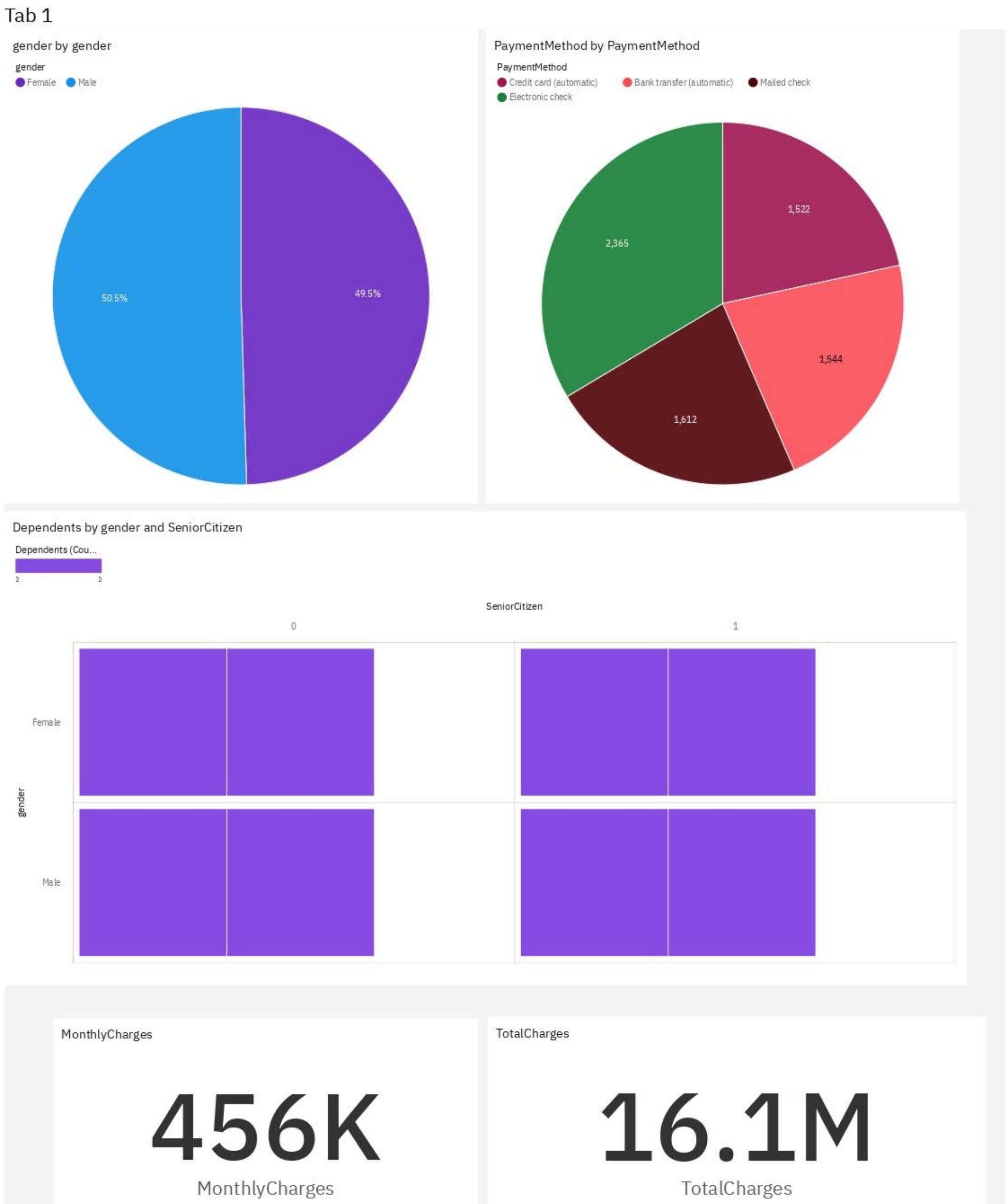
Phase 4: Development Part 2

In this part you will continue building your project.

- Continue building the analysis by creating visualizations using IBM Cognos and developing a predictive model.
- Create interactive dashboards and reports in IBM Cognos to visualize churn patterns, retention rates, and key factors influencing churn.
- Use machine learning algorithms to build a predictive model that identifies potential churners based on historical data and relevant features

.

WE HAVE CREATED DASHBOARD USING IBM COGNOS

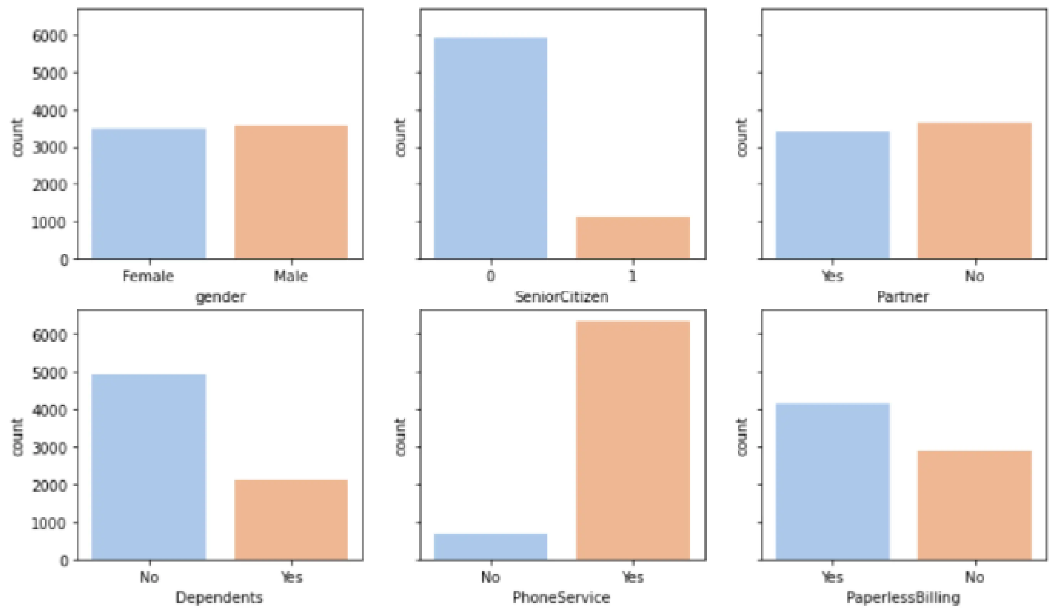


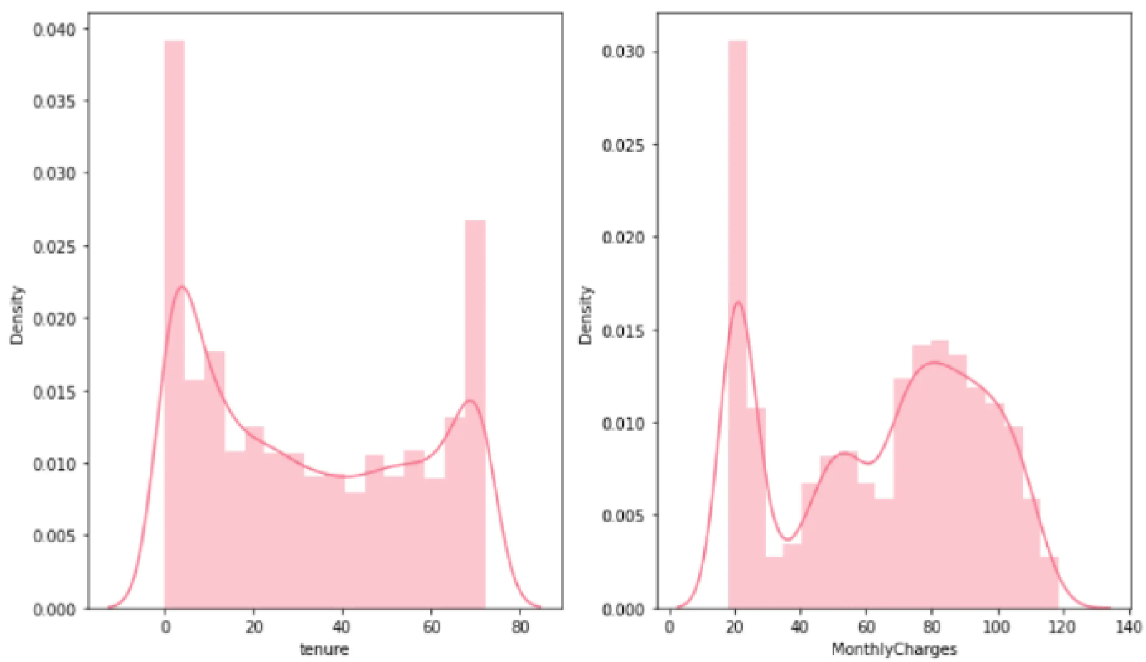
We have used SVM algorithm to build predictive modeling

```
# Loading Data

# Importing Dataset
data =
pd.read_csv("/kaggle/input/telco-customer-churn/WA_Fn-UseC_-Telco-Customer-Churn.c
sv")
# Printing Data
data.head()
```

DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

```
with sns.color_palette("pastel"):
   fig, axes = plt.subplots(2, 3, figsize=(12, 7), sharey=True)
   sns.countplot("gender", data=data, ax=axes[0,0])
   sns.countplot("SeniorCitizen", data=data, ax=axes[0,1])
   sns.countplot("Partner", data=data, ax=axes[0,2])
   sns.countplot("Dependents", data=data, ax=axes[1,0])
   sns.countplot("PhoneService", data=data, ax=axes[1,1])
   sns.countplot("PaperlessBilling", data=data, ax=axes[1,2])


with sns.color_palette("husl"):
   fig, axes = plt.subplots(1,2, figsize=(12, 7))
   sns.distplot(data["tenure"], ax=axes[0])
   sns.distplot(data["MonthlyCharges"], ax=axes[1])
```





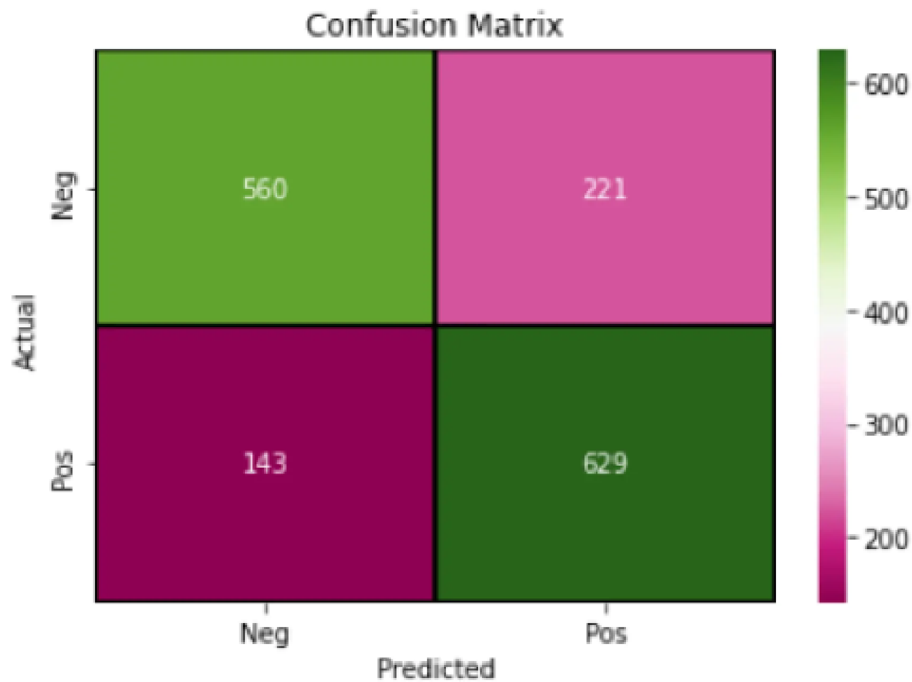
SVM CLASSIFIER

1. **SVM - SVM or Support Vector Machine is a supervised machine learning technique used for classification and regression. Finding a hyperplane in**

an N-dimensional space that classifies the data points is the goal of the SVM method. The number of features determines the hyperplane's size.

```
# Training the model using the optimal parameters discovered with SVM Classifier
svmclf =  SVC(C=3,class_weight='balanced', random_state=43)
svmclf.fit(X_train,y_train)

result2 = ["2.","SVM","Balanced using class weights"]
y_pred_tr = svmclf.predict(X_train)
print('Train accuracy SVM: ',accuracy_score(y_train,y_pred_tr))
result2.append(round(accuracy_score(y_train,y_pred_tr),2))

y_pred_test = svmclf.predict(X_test)
print('Test accuracy SVM: ',accuracy_score(y_test,y_pred_test))
result2.append(round(accuracy_score(y_test,y_pred_test),2))

recall = recall_score(y_test,y_pred_test)
print("Recall Score: ",recall)
result2.append(round(recall,2))

# Building a confusion matrix
matrix = confusion_matrix(y_test,y_pred_test)
ax=plt.subplot();
sns.heatmap(matrix, annot=True, fmt='d', linewidths=2, linecolor='black',
cmap='YlGnBu',ax=ax)
ax.set_xlabel('Predicted')
ax.set_ylabel('Actual')
ax.set_ylim(2.0,0)
ax.set_title('Confusion Matrix')
ax.xaxis.set_ticklabels(['Neg','Pos'])
ax.yaxis.set_ticklabels(['Neg','Pos'])
plt.show()
```

```
OUTPUT
Train accuracy SVM:  0.8186469584991473
Test accuracy SVM:  0.7656149388280747
Recall Score:  0.8147668393782384
```



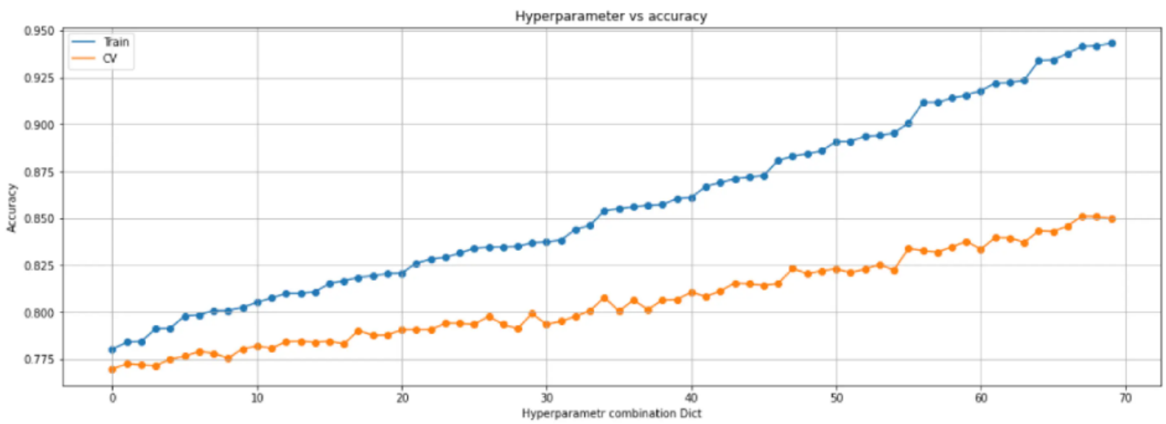
1. XG Boost - Formally speaking, XGBoost may be described as a decision tree-based ensemble learning framework that uses Gradient Descent as the underlying objective function. It offers excellent flexibility and efficiently uses computation to produce the mandated results.

```
# Grid Search To Get Best Hyperparameters

parameters = {"learning_rate"    : [0.10,0.20,0.30 ],\

            "max_depth"       : [ 3,5,10,20],\

            "n_estimators" : [ 100, 200, 300, 500],\

            "colsample_bytree" : [ 0.3, 0.5, 0.7 ] }

clf_xgb = XGBClassifier(scale_pos_weight=scale, eval_metric ='mlogloss')

grid = GridSearchCV(estimator=clf_xgb, param_grid=parameters,
scoring='accuracy',return_train_score=True,verbose=1)

grid.fit(X_train,y_train)


# plotting only the first 70 train scores
```

```
cv_result =
pd.DataFrame(grid.cv_results_).sort_values(by='mean_train_score',ascending=True)[:70]

param_list = list(cv_result['params'])

param_index = np.arange(70)

plt.figure(figsize=(18,6))

plt.scatter(param_index,cv_result['mean_train_score'])

plt.plot(param_index,cv_result['mean_train_score'],label='Train')

plt.scatter(param_index,cv_result['mean_test_score'])

plt.plot(param_index,cv_result['mean_test_score'],label="CV")

plt.title('Hyperparameter vs accuracy')

plt.grid()

plt.legend()

plt.xlabel('Hyperparametr combination Dict')

plt.ylabel('Accuracy')

plt.show()
```

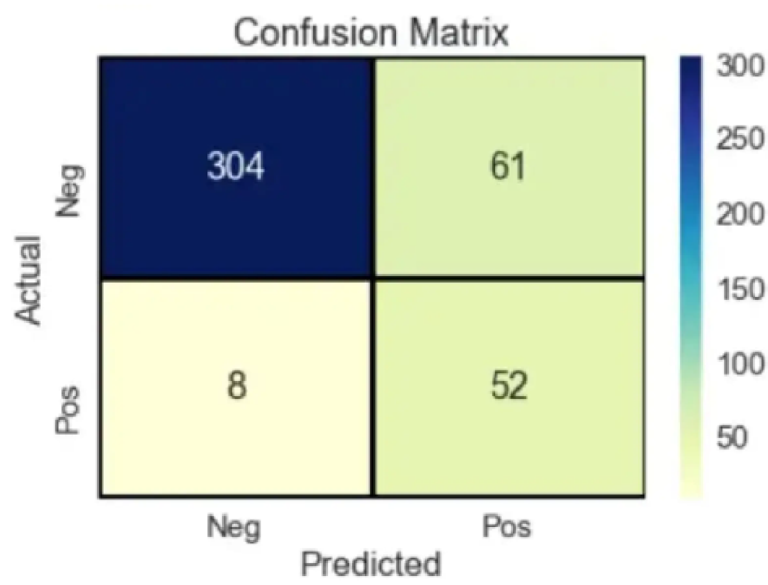
OUTPUT

Fitting 5 folds for each of 144 candidates, totaling 720 fits




```
# Using XG Boost

clf_xgb = XGBClassifier(learning_rate= best_parameters['learning_rate']
,max_depth=best_parameters ['max_depth'],
n_estimators=best_parameters['n_estimators'],
colsample_bytree=best_parameters['colsample_bytree'],
eval_metric='mlogloss',scale_pos_weight=scale)

clf_xgb.fit(X_train,y_train)


xgbresult = ["4.","XGBClassifier","Balanced using scale_pos_weight"]

y_pred_tr = clf_xgb.predict(X_train)

print('Train accuracy XGB: ',accuracy_score(y_train,y_pred_tr))

xgbresult.append(round(accuracy_score(y_train,y_pred_tr),2))


y_pred_test = clf_xgb.predict(X_train)

print('Test accuracy XGB: ',accuracy_score(y_test,y_pred_test))

xgbresult.append(round(accuracy_score(y_test,y_pred_test),2))


recall = recall_score(y_test,y_pred_test)

print("Recall Score: ",recall)
```

```
xgbresult.append(round(recall,2))


# Building confusion matrix
cm = confusion_matrix(y_test,y_pred_test)
ax=plt.subplot();
sns.heatmap(cm, annot=True, fmt='d', linewidths=2, linecolor='black',
cmap='YlGnBu',ax=ax)
ax.set_xlabel('Predicted')
ax.set_ylabel('Actual')
ax.set_ylim(2.0,0)
ax.set_title('Confusion Matrix')
ax.xaxis.set_ticklabels(['Neg','Pos'])
ax.yaxis.set_ticklabels(['Neg','Pos'])
plt.show()


Train accuracy XGB:  0.8543490619670268
Test accuracy: 0.80
Recall Score: 0.75
```



CONCLUSION

IN THIS PHASE WE HAVE CREATED DASHBOARD USING IBM COGNOS

AND WE USED MACHINE LEARNING ALGORITHM TO BUILD PREDICTIVE MODELING FOR CUSTOMER DATA AND WE USED SVM AND XG BOOST