Computer Science Department

Machine Learning COMP4388

Student Name: Raheeq Mousa

Student ID: 1220515

Instructor: Mr. Radi Jarrar

First of all, before starting to work on the data, we must check if there are any empty values.

```
PS C:\Users\DELL\Desktop\Raheeq\5th_smstr_RaheeqMousa\Machine learning\Assignment\Assignment python> python main.py
ID                          0
Call Failure                0
Complains                   0
Charge Amount               0
Freq. of use                0
Freq. of SMS                0
Distinct Called Numbers     0
Age Group                   0
Plan                        0
Status                      0
Age                         0
Customer Value              0
Churn                       0
dtype: int64
```

Figure 1

From Figure1, we can see that there are no missing data in the data set.

1. Summary statistics of all attributes in the dataset

```
[9]: read_data.describe()
```

[9]:

	ID	Call Failure	Charge Amount	Freq. of use	Freq. of SMS	Distinct Called Numbers	Age Group	Age	Customer Value
count	3150.000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000
mean	1575.500	7.627937	129.882540	69.460635	73.174921	23.509841	2.826032	30.998413	470.972916
std	909.471	7.263886	102.790931	57.413308	112.237560	17.217337	0.892555	8.831095	517.015433
min	1.000	0.000000	20.000000	0.000000	0.000000	0.000000	1.000000	15.000000	0.000000
25%	788.250	1.000000	50.000000	27.000000	6.000000	10.000000	2.000000	25.000000	113.801250
50%	1575.500	6.000000	100.000000	54.000000	21.000000	21.000000	3.000000	30.000000	228.480000
75%	2362.750	12.000000	200.000000	95.000000	87.000000	34.000000	3.000000	30.000000	788.388750
max	3150.000	36.000000	400.000000	255.000000	522.000000	97.000000	5.000000	55.000000	2165.280000

Figure 2

From the count row → The number of each label is equal to the number of rows in the file, this means that there are no empty values.

Count: the number of non-null entries in the column.

Mean: the average of each column.

Std: the standard deviation, the amount of variation of the values of a variable about its mean.

Min: minimum value in the column.

Max: maximum value in the column.

25%: the 25th percentile (first quartile), represents the value below which 25% of the data falls.

50%: the 50th percentile (second quartile), represents the value below which 50% of the data falls.

75%: the 75th percentile (third quartile), represents the value below which 75% of the data falls.

Since that it is not useful to summary catergorical classes then we will summary only the numerical data.

We can say from the minimum number of "distinct called numbers" which equals zero means that there exists a customer whos

From the age column, The average age is 30.998, with a standard deviation of 8.831, indicating relatively young customers.

Note: in this task, I have included rows that have Freq. of Use = 0, Freq. of SMS = 0, Status = Not-active, and Churn = No, so provide useful insights into subscriber declines or increased inactive users.

2. Show the distribution of the class label (churn) and indicate any highlights in the distribution of the class label.

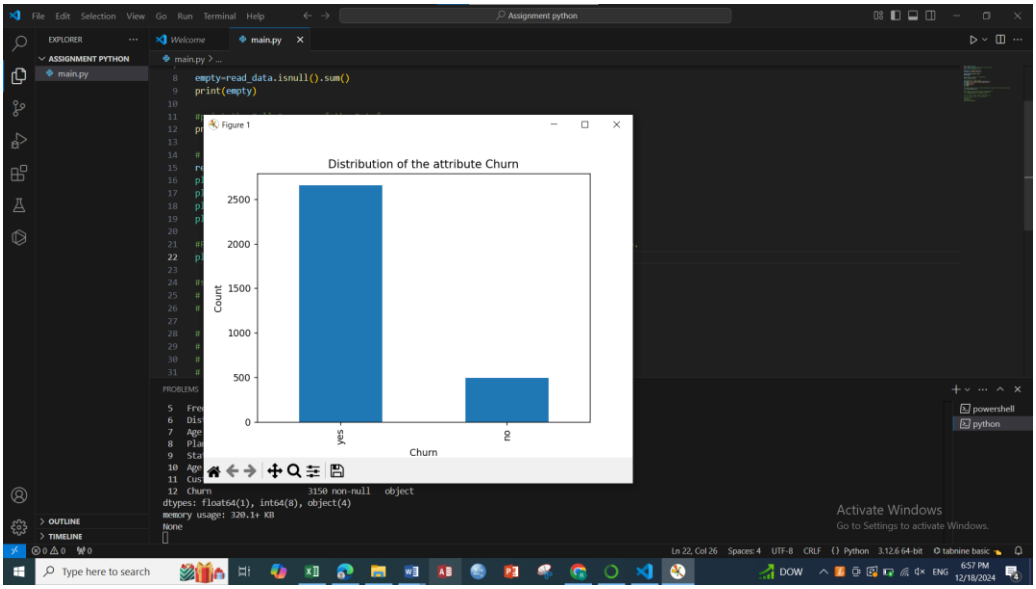


Figure 3

I represented the bar chart to display the distribution because “Churn” is a categorical data. I displayed the value of Churn on the x-axis and the count of it on the y-axis.

A bar chart with two bars: a bar for Churn = “yes” and a bar for Churn = “no”

This shows the relative proportion of churned vs. non-churned customers.

From figure 3, the Churn rate is significantly larger than the non-Churn rate, which means that a lot of the company customers have left or unsubscribed the company service or product so is must reconsider their service.

Note: in this task, I have included rows that have Freq. of Use = 0, Freq. of SMS = 0, Status = Not-active, and Churn = No, so provide useful insights into subscriber declines or increased inactive users.

3. For each age group, draw a histogram detailing the amount of churn in each sub-group.

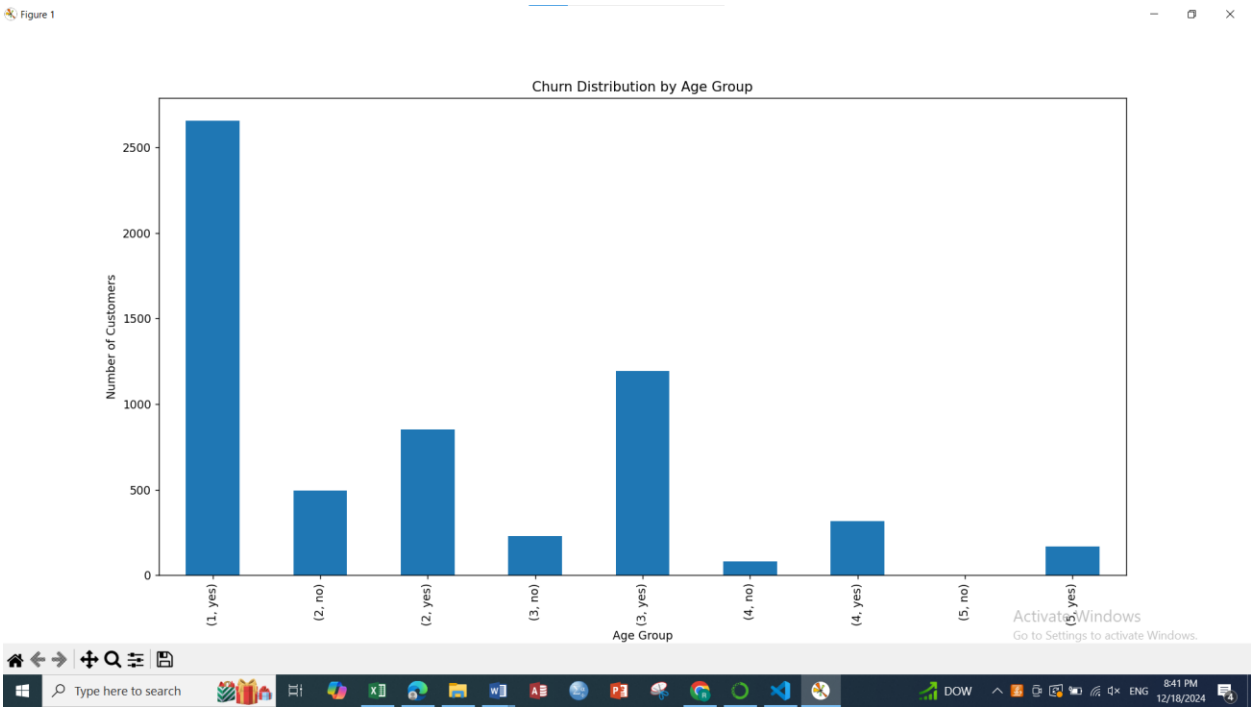


Figure 4

I turned the figure4 into this bellow to mke it more clear.

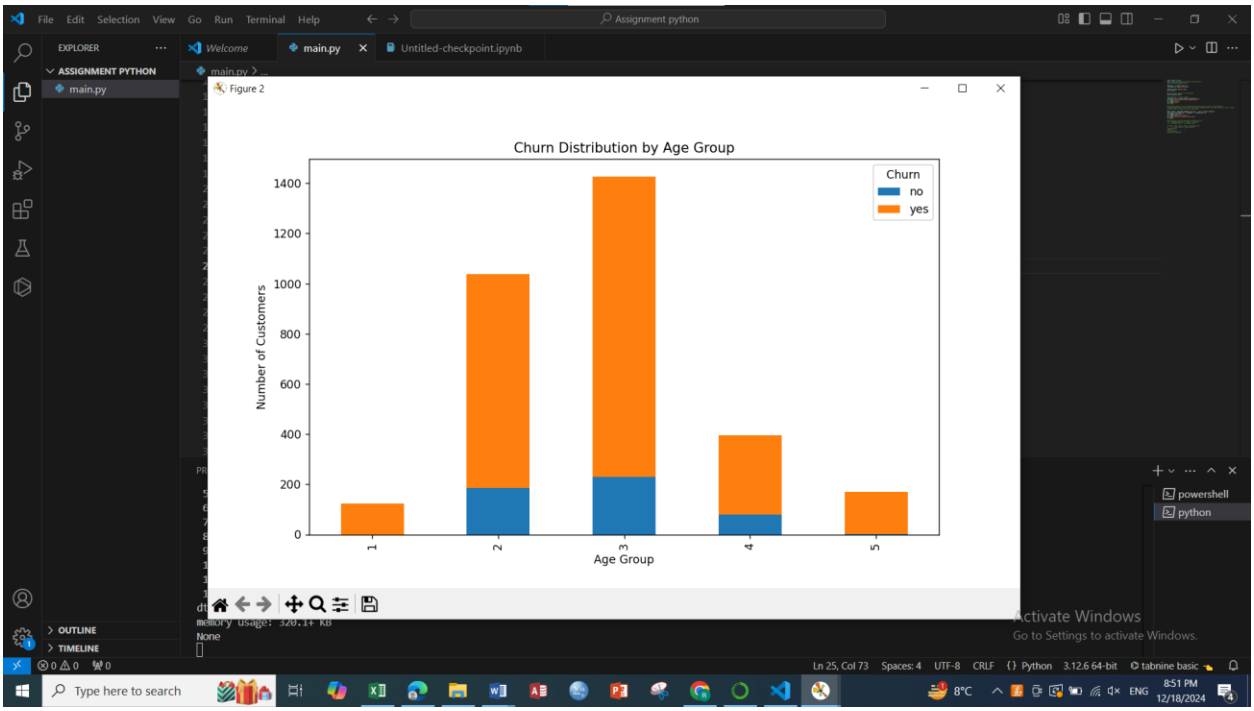


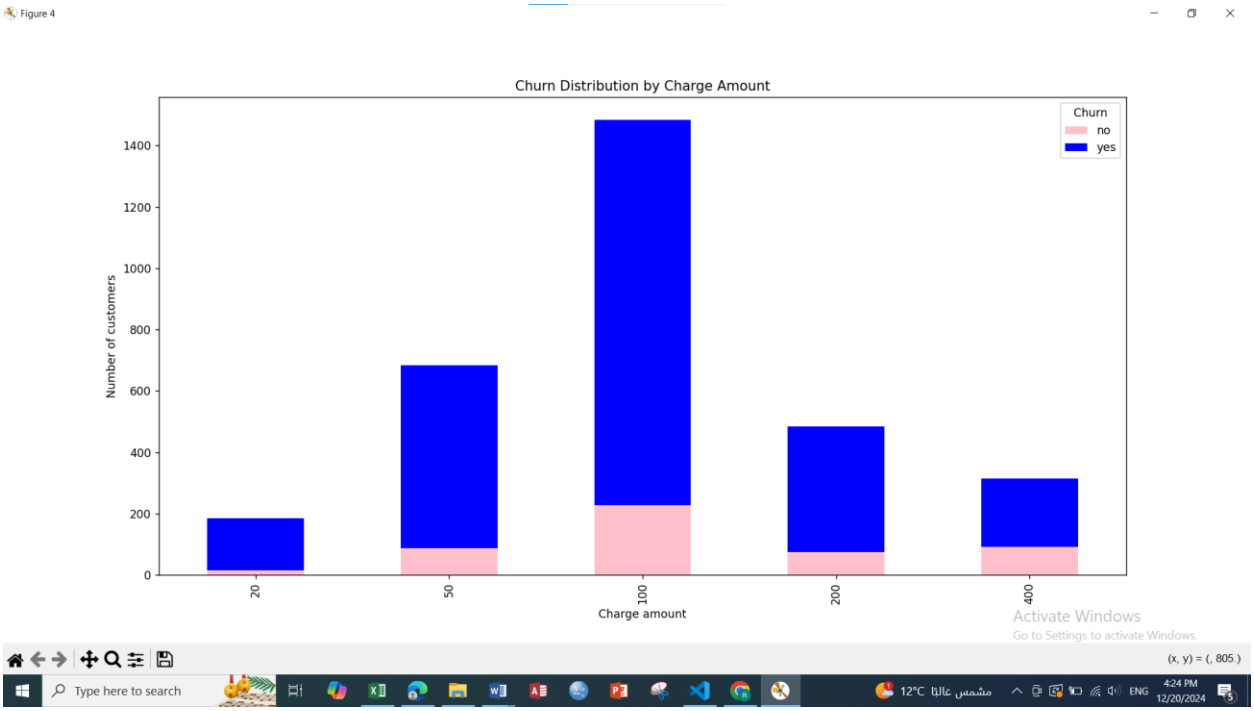
Figure 5

From Figure 5, the age group 3 (age=30) has the highest churn rate and number of customers. Also, the data is skewed to the right meaning that the mean is larger than the median this means that the customers with the older age becomes less.

```
minimum age for age group
Age Group
1    15
2    25
3    30
4    45
5    55
Name: Age, dtype: int64
maximum age for age group
Age Group
1    15
2    25
3    30
4    45
5    55
```

Note: in this task, I have included rows that have Freq. of Use = 0, Freq. of SMS = 0, Status = Not-active, and Churn = No, so provide useful insights into subscriber declines or increased inactive users.

4. For each charge amount, draw a histogram detailing the amount of churn in each sub-group.



For the customers who paid 100 for the service, the number of churned customers is larger so the company should reconsider the price of the service versus the quality. Also, this histogram is skewed to the right, and the churned customers rate is larger than the not churned customers meaning that even with more payment the customers are not satisfied with the service.

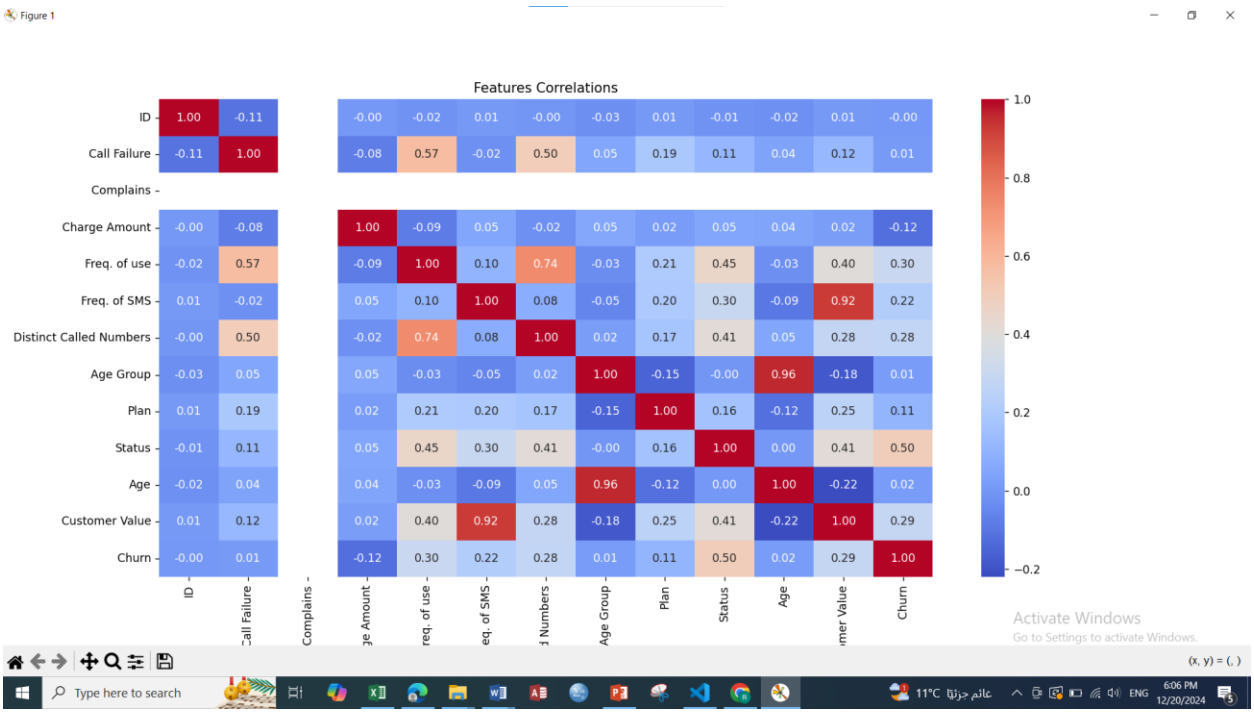
5. Show the details of the charge amount of customers.

```
Charge Amount Details
count    3150.000000
mean      129.882540
std       102.790931
min        20.000000
25%        50.000000
50%       100.000000
75%       200.000000
max       400.000000
Name: Charge Amount, dtype: float64
```

Interquartile= 75th percentile – 25th percentile =150

Since the maximum value is greater than the 75th percentile + Interquartile, the maximum value is an outlier. That means that the number of people who subscribed to the most expensive plan is a small number of customers.

6. Visualise the correlation between all features and explain them in your own words.



Freq. of SMS has a high correlation with customer value =0.92.

And the features with very low correlation with Churn are (ID, Call Failure, Complains, Plan, Age, Age Group) which have a values of (0, 0.01, -, 0.11, 0.02, 0.01 respitivally).

So the features that has well a correlation relationship with Churn are freq. of use, freq. of sms, distinct called numbers, status, customer value.

7. Split the dataset into training (70%) and test (30%).

```
training set for linear regression
(2205, 8)
testing set for linear regression
(945, 8)
```

Regression Tasks

1. Apply linear regression to learn the attribute “Customer Value” using all independent attributes (call this model LRM1).

```
Linear Regression "First model" measurements
Mean Squared Error (LRM1):  0.0753438483808761
Mean Absolute Error (LRM1):  0.1684998320008554
R-squared (LRM1):  0.4561419686998296
```

From the MAE value, we can see that the model’s predictions are so close to the true values, Which is good. Also, I noticed that MSE> MAE and that means that there are outliers that affected the MSE.

The value of the R^2 measure indicates a moderate fit for this model, since that a R-squared value closer to 1 suggests a better fit.

This result means that the model needs more improvement, particularly through better feature selection, and regularization.

2. Apply linear regression using the set of the 2 most important features (from your point of view); and explain why did you use these 2 attributes (call this model LRM2).

The most important features, in my opinion, are “Freq. of use” and “Charge Amount” since they are highly affecting the target class “Customer Value”.

The feature “Freq. of use” is important because the large number of calls means that the customer may face many problems and may also be satisfied with the service, which is why he has many calls.

The feature “Charge Amount” is important because its can also provide insight into the likelihood of churn, and it means that they are subscribing on an expensive plans which is gonna increase their customer value.

```
Linear Regression "Second model" measurements
Mean Squared Error=  244785.69726492662
Mean Absolute Error=  363.3442491561602
R squared=  0.14767235814854351
```

Mean square error is huge meaning that the model's predictions is so far away from the actual y. Also the R squared measure is so small meaning that the model's accuracy is very bad!

3.Apply linear regression using the set of the most important features (based on the correlation coefficient matrix) and explain why did you use these attributes (call this model LRM3).

Note : For the splitting data task, I have discarded some unuseful attributes that mislead the statistics from the coming tasks.

First, I discarded the "ID" because it is just an identifier that is not a useful feature for prediction and dispersion of the final statistic. Also, I have discarded the feature "Churn" since it is suitable for the classification task not for the regression. Lastly, I have discarded the feature "Age" since I have the feature "Age Group" which is considered sufficient to find the relationship between the output and the feature.

Also, regarding on the correlation matrix I decided the features ('Freq. of Use' , 'Freq. of SMS', 'Distinct Called Numbers', 'Status') for this model because they have a strong correlation value with the customer value

```
Linear Regression "Third model" measurements
Mean Absolute Error=  63.88364912672473
Mean Squared Error=  11561.881099645907
accuracy=  0.9597422931031678
R squared=  0.9597422931031678
  Model            MSE       R^2
0  LRM1       0.168500  0.456142
1  LRM2  254714.859700  0.113100
2  LRM3      63.883649  0.959742
```

```
Mean Squared Error=  11561.881099645907
```

Since the mean square error is a huge value, then the model's predictions are so far to the true values!

Since the MSE is sensitive to the outliers then maybe the model makes good predictions but it is huge because of the outliers. Because of that, we must calculate the Mean absolute error (MAE) since it is not sensitive to outliers.

```
Mean Absolute Error=  63.88364912672473
Mean Squared Error=  11561.881099645907
```

From the MAE value, we can see that the model's predictions are so close to the true values! Which is good.

```
accuracy=  0.9597422931031678
```

Accuracy is equal to the proportion of the correct prediction, accuracy is not that used and appropriate in regression tasks. R^2 measure is more suitable for linear regression.

```
R squared=  0.9597422931031678
```

R^2 =0.9597422931031678 usually indicates that the model is very accurate. We can improve the accuracy of the model by adding more useful features to the test set.

Overall, a small MAE typically indicates very good predictions and low R^2 measure suggests that the model is very accurate.

4. Compare the performance of these models using adequate performance metrics

	Model	MSE	MAE	R^2
0	LRM1	0.075344	0.168500	0.456142
1	LRM2	254714.859700	370.144164	0.113100
2	LRM3	11561.881100	63.883649	0.959742

The best and most effective feature is LRM3 because it has a low MAE and the highest R-squared measure. It avoids overfitting and underfitting by not using all features and by not using a few features.

These tasks highlight the importance of feature selection, which affects the model accuracy.

Classification Tasks

1. Run k-Nearest Neighbours classifier to predict churn of customers (the “Churn” feature) using the test set.

```
K-Nearest Neighbours classifier measurements
Accuracy Score=  0.9227513227513228
Confusion Matrix:
 [[113  44]
 [ 29 759]]
classifiction report =
              precision    recall  f1-score   support

           0       0.80      0.72      0.76       157
           1       0.95      0.96      0.95       788

    accuracy                           0.92       945
   macro avg       0.87      0.84      0.85       945
weighted avg       0.92      0.92      0.92       945

ROC-AUC Score: 0.9165750590061107
```

Note: Before training kNN model, we must scale the data before because it is a distance algorithm based since it calculates the distance between data points from training set and the inputs from the test set to classify them). So if the features have different scales, features with larger numerical ranges will control the distance calculation, leading to biased predictions.

Experimentally, after trying an odd value of k between 3 and 10, k=3 has the highest accuracy score.

From the output the confusion matrix is as this:

	Predicted 0	Predicted 1
Actual 0	113	44
Actual 1	29	759

TP:759 , TN=113 , FP=44 , FN=29

From the output of the cassification report

	0 (churn=no)	1 (churn =1)	Weighted avg
precision	0.8	0.95	0.92
recall	0.72	0.96	0.92
F1-score	0.76	0.95	0.92

Precision → the proportion of positive examples that are correctly classified

80% of customers predicted as not churned were actually correct.

95% of customer predicted as churned were actually correct.

Recall → the proportion of positive examples that were correctly classified (from the dataset)

72% of actual non-churned customers predicted as not churned correctly.

96% of actual churned customers were predicted as churned correctly.

The model works well on predicting churn, with high precision (95%) and high recall (96%), and minimizes the FN as possible. While the performance of predicting non-churned is lower than predicting churn (recall=0.72).

2. Run Naive Bayes classifier to predict churn of customers (the “Churn” feature) using the test set

```
Naive Bayes classifie measurements
Accuracy Score=  0.7597883597883598
Confusion Matrix:
 [[141  16]
 [211 577]]
classifiction report =
              precision    recall  f1-score   support

           0       0.40      0.90      0.55       157
           1       0.97      0.73      0.84       788

    accuracy                           0.76       945
   macro avg       0.69      0.82      0.69       945
weighted avg       0.88      0.76      0.79       945

ROC-AUC Score: 0.9095549484302758
```

	Predicted 0	Predicted 1
Actual 0	141	16
Actual 1	211	577

TP=577, TN=141, FP=16, FN=211

	0 (churn=no)	1 (churn =1)	Weighted avg
precision	0.4	0.97	0.88
recall	0.90	0.73	0.76
F1-score	0.55	0.84	0.79

Precision → the proportion of positive examples that are correctly classified

40% of actual non-churned customers predicted as not churned correctly.

97% of actual churned customers were predicted as churned correctly.

Recall → the proportion of positive examples that were correctly classified (from the dataset)

90% of actual non-churned customers predicted as not churned correctly.

73% of actual churned customers were predicted as churned correctly.

The Naive Bayes model reached high precision of 97% of predicting churned customers, meaning that when the model predicts a customer will churn, it is correct 97% of the time. However, the recall was 73%, meaning that the model missed about 27% of the actual churn cases what inidactes a lower portion of positive examples captures in the model.

3. Run Decision Tree classifier to predict churn of customers (the"Churn" feature) using the test set.

```
Decision Tree classifier Results
Accuracy Score on train set 0.9383219954648526
Accuracy Score on test set=  0.9153439153439153
Confusion Matrix:
 [[ 86  71]
 [  9 779]]
classifiction report =
              precision    recall  f1-score   support

           0       0.91      0.55      0.68       157
           1       0.92      0.99      0.95       788

    accuracy                           0.92       945
   macro avg       0.91      0.77      0.82       945
weighted avg       0.91      0.92      0.91       945

ROC-AUC Score: 0.9116201623072199
```

	Predicted 0	Predicted 1
Actual 0	86	71
Actual 1	9	779

TP=779, TN=86, FP=71, FN=9

	0 (churn=no)	1 (churn =1)	Weighted avg
precision	0.91	0.92	0.91
recall	0.55	0.99	0.92
F1-score	0.68	0.95	0.91

Precision → the proportion of positive examples that are correctly classified

91% of actual non-churned customers predicted as not churned correctly.

92% of actual churned customers were predicted as churned correctly.

Recall → the proportion of positive examples that were correctly

classified (from the dataset)

55% of actual non-churned and predicted as not churned correctly

99% of actual churned persons were predicted as churned correctly

The Decision Tree model reached high precision of 92% of predicting churned customers, meaning that when the model predicts a customer will churn, it is correct 92% of the time. However, the recall was 99%, meaning that the model missed about 1% only of the actual churn cases which indicates a higher portion of positive examples captured correct in the model.

The Decision Tree accuracy on the training set equals to 0.938, while the accuracy on the test set equals to 0.915, meaning that the model has learned the patterns in the training set and does good work on the test set without memorizing specific details of the training set so this the model is not overfitting.

4. Compare the performance of Logistic regression, Naive Bayes, and kNN classifiers in an appropriate results section. Compare the classification performance of the generated classification models and make sure to use the appropriate performance metrics. You should include at least the ROC/AUC score and the Confusion Matrix. Report the results in an appropriate table and explain in your own words why one model outperforms the other.

```
Logistic Regression measurements
Accuracy Score=  0.9068783068783068
Confusion Matrix:
 [[ 76  81]
 [  7 781]]
classifiction report =
              precision    recall  f1-score   support

           0       0.92      0.48      0.63       157
           1       0.91      0.99      0.95       788

    accuracy                           0.91       945
   macro avg       0.91      0.74      0.79       945
weighted avg       0.91      0.91      0.89       945

ROC-AUC Score: 0.9303970383782212
```

	Logistic Regression	Naive Bayes	kNN
Accuracy	0.9068783068783068	0.7597883597883598	0.9227513227513228
ROC-AUC	0.9303970383782212	0.9095549484302758	0.9165750590061107
Precision (churn)	0.91	0.97	0.95
Recall (churn)	0.99	.73	0.9
F1-score (churn)	0.95	0.84	0.95
Confusion matrix	TP=781, TN=76 FP=81 , FN=7	TP=577 TN=141 FP=16 FN=211	TP:759 , TN=113 , FP=44 , FN=29

A) Logistic Regression:

Accuracy: 0.907 means that the model correctly predicted 92% of the test set.

ROC-AUC: 0.930 means that the Logistic Regression model has very good performance in recognition between churn and non-churn customers.

Precision and Recall: The precision and recall for predicting churn are very high (precision=0.91 and recall=0.99), meaning the model is both accurate in predicting churned and non-churned customers.

Confusion Matrix: The confusion matrix shows 81 false positives (actual churned customers incorrectly classified as non-churned), 7 false negatives (churn customers missed by the model that is predicted as non-churned), 781 True positives (actual churned customers correctly classified as churned), and 76 true negatives (actual non churned customers correctly classified as non churned).

B) KNN:

Accuracy: 0.923 means that the model correctly predicted 92% of the test set.

ROC-AUC: 0.917 means that the k-Nearest Neighbor model has very good performance in predicting between churn and non-churn customers.

Precision and Recall: The precision is very high as it reached 95%, while the recall equals 0.96, meaning that the proportion of positive examples that were correctly classified is 96% of the data set.

Confusion Matrix: The confusion matrix shows 16 false positives (actual churned customers incorrectly classified as non-churned), 211 false negatives (churn customers missed by the model that is predicted as non-churned), 577 True positives (actual churned customers correctly classified as churned), and 113 true negatives (actual non churned customers correctly classified as non churned).

C) Naive Bayes:

Accuracy: 0.759 means that the model correctly predicted 75% of the test set.

ROC-AUC: 0.910 means that the Naïve Bayes model has very good performance in prediction between churn and non-churn customers.

Precision and Recall: The precision is very high as it reached 91%, while the recall equals 0.73, meaning that the proportion of positive examples that were correctly classified is 73% of the data set.

Confusion Matrix: The confusion matrix shows 16 false positives (actual churned customers incorrectly classified as non-churned), 211 false negatives (churn customers missed by the model that is predicted as non-churned), 577 True positives (actual churned customers correctly classified as churned), and 141 true negatives (actual non churned customers correctly classified as non churned).

Logistic Regression is larger than both Naive Bayes and kNN in terms of accuracy and ROC-AUC score, showing it is a very good performance model for predicting churn with a good balance of precision and recall. Therefore, if you are interested in high recall and accuracy KNN and logistic regression would be the best choice. And if you are interested in high precision Naïve Bayes would be a good choice.

Finally, Logistic Regression provided the best overall performance, also it balanced both precision and recall effectively with high accuracy and very good performance in recognition.