Final project

Datamining

Sales Of Medical Devices

By Ahmad Alfahad

1. **Introduction**

In this project, my aim is to help a business owner understand the similarities and relationships within their customer data using the SalesOfMedicalDevices dataset. This dataset includes various attributes such as CustomerID, year, NoOfSalesCalls, NoOfTargetedEmails, NoOfSales, and customer satisfaction levels. The owner is interested in gaining insights from this data, and I will be using data mining techniques to assist them.

1.1Task Description:

The main objective of this report is to utilize data mining techniques to analyze the SalesOfMedicalDevices dataset and identify similarities and relationships between the customers' data. To achieve this, I will employ various techniques, such as regression analysis, including Linear Regression, Lasso Regression, and Logistic Regression, to obtain similar results. Additionally, I will use Principal Component Analysis (PCA) to reduce dimensions and discover relationships. Furthermore, Decision Trees will assist me in identifying the most important features that significantly impact customer satisfaction.

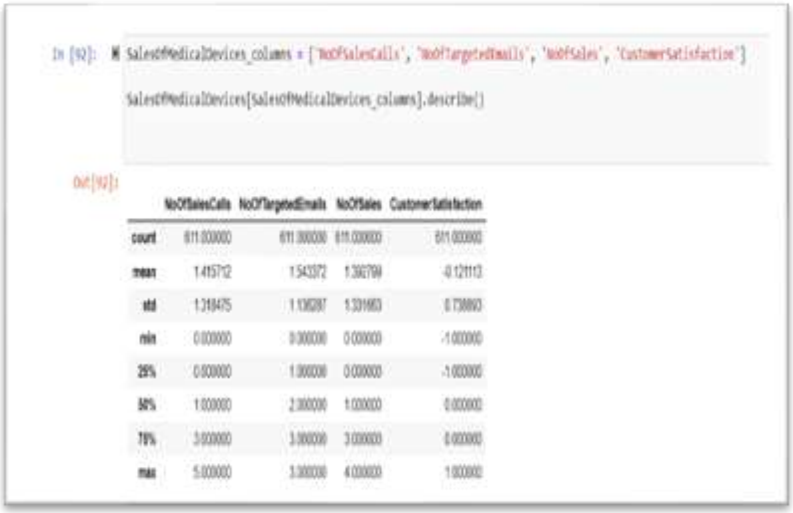
1.2 Data Description:

The dataset consists of 611 customers and 6 columns, namely CustomerID, year, NoOfSalesCalls, NoOfTargetedEmails, NoOfSales, and Customer Satisfaction, with the latter containing 3 values (-1, 0, 1). The dataset is at the customer level granularity.

2. Data Preparation:

2.1 Data Exploration

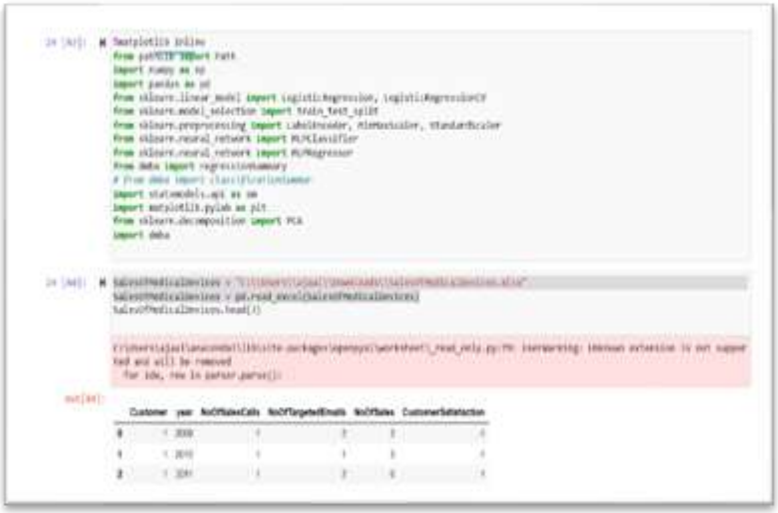
In the first step, I imported all the necessary libraries and read the 'SalesOfMedicalDevices' dataset. I then displayed the first 3 rows of the dataset to get a quick overview. To gain more insights about the data, I used the 'describe()' function which provided a more detailed summary of the dataset.



```
In [92]: SalesOfMedicalDevices_columns = ['NoOfSalesCalls', 'NoOfTargetedEmails', 'NoOfSales', 'CustomerSatisfaction']

         SalesOfMedicalDevices[SalesOfMedicalDevices_columns].describe()
```

Out[92]:

	NoOfSalesCalls	NoOfTargetedEmails	NoOfSales	CustomerSatisfaction
count	611.000000	611.000000	611.000000	611.000000
mean	1.415712	1.543372	1.390799	-0.121113
std	1.218475	1.136287	1.331663	0.738860
min	0.000000	0.000000	0.000000	-1.000000
25%	0.000000	1.000000	0.000000	-1.000000
50%	1.000000	2.000000	1.000000	0.000000
75%	3.000000	3.000000	3.000000	0.000000
max	5.000000	3.000000	4.000000	1.000000



```
In [43]: %matplotlib inline
         from pathlib import Path
         import numpy as np
         import pandas as pd
         from sklearn.linear_model import LogisticRegression, LogisticRegressionCV
         from sklearn.model_selection import train_test_split
         from sklearn.preprocessing import LabelEncoder, MinMaxScaler, StandardScaler
         from sklearn.neural_network import MLPClassifier
         from sklearn.neural_network import MLPRegressor
         from dmba import regressionSummary
         # from dmba import classificationSummary
         import statsmodels.api as sm
         import matplotlib.pylab as plt
         from sklearn.decomposition import PCA
         import dmba
```

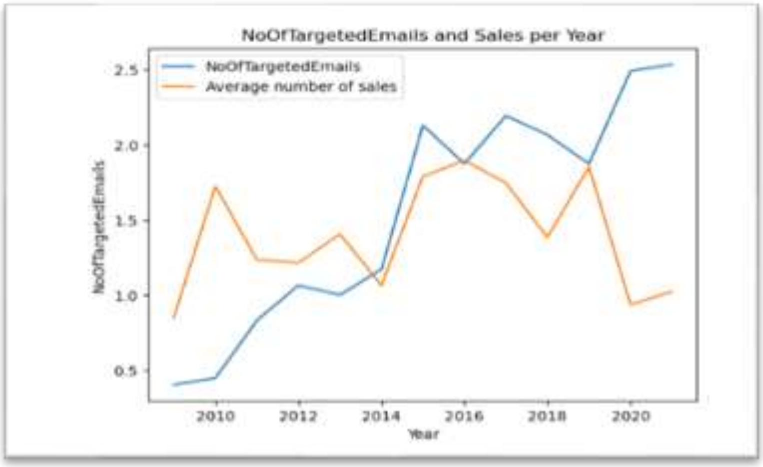
```
In [44]: SalesOfMedicalDevices = "[illegible]SalesOfMedicalDevices.xlsx"
         SalesOfMedicalDevices = pd.read_excel(SalesOfMedicalDevices)
         SalesOfMedicalDevices.head(3)

C:\Users\[illegible]\anaconda3\lib\site-packages\openpyxl\worksheet\_read_only.py:79: UserWarning: Unknown extension is not suppor
ted and will be removed
  for idx, row in parser.parse():
```

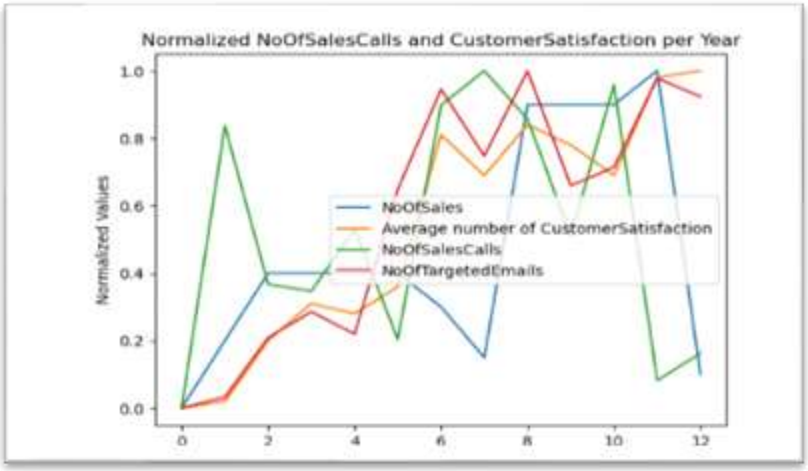
Out[44]:

	Customer	year	NoOfSalesCalls	NoOfTargetedEmails	NoOfSales	CustomerSatisfaction
0	1	2009	1	2	2	1
1	1	2010	1	1	1	1
2	1	2011	1	2	0	1

2.2.Data Visualization

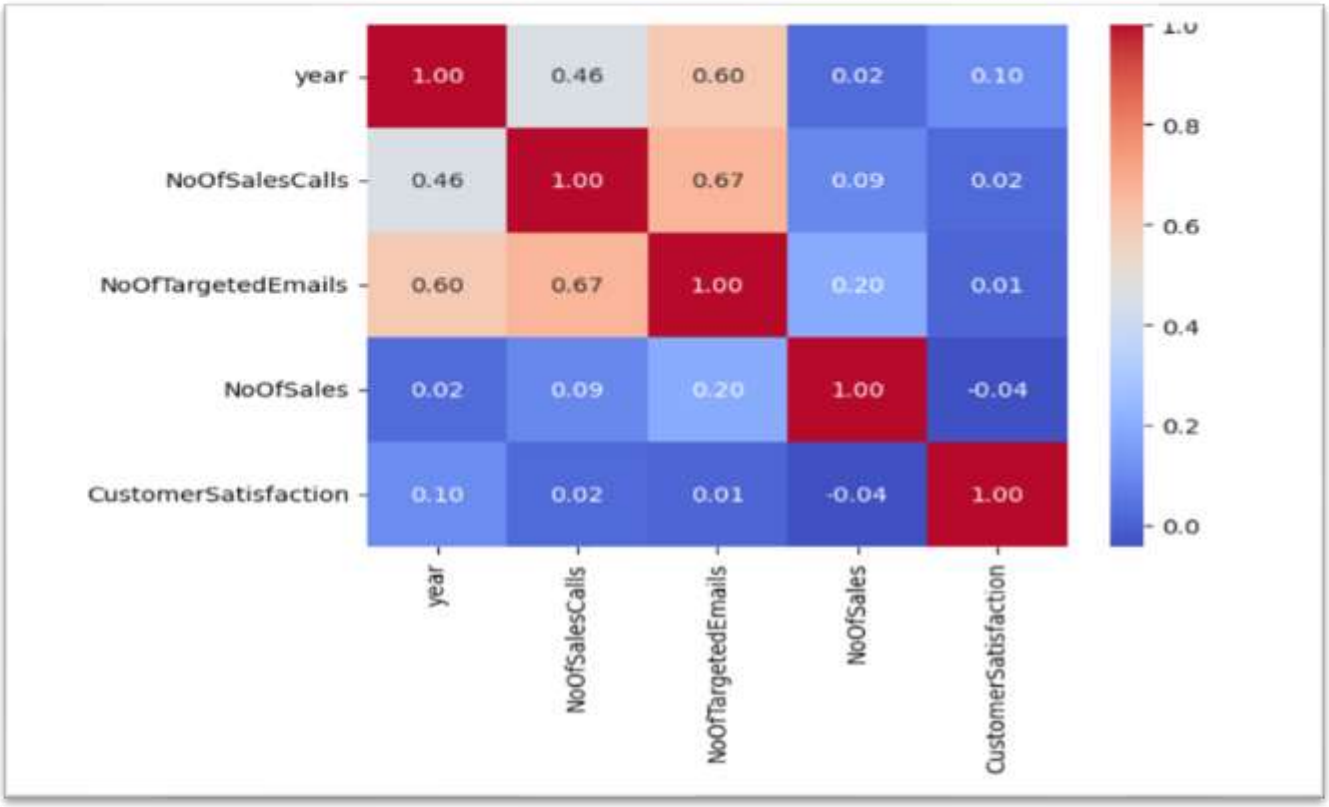
In this section, I normalized the dataset to ensure that all the variables are on the same scale. Then, I visualized the dataset to gain a better understanding of the data by using MinMaxScaler() and plotted a line chart.



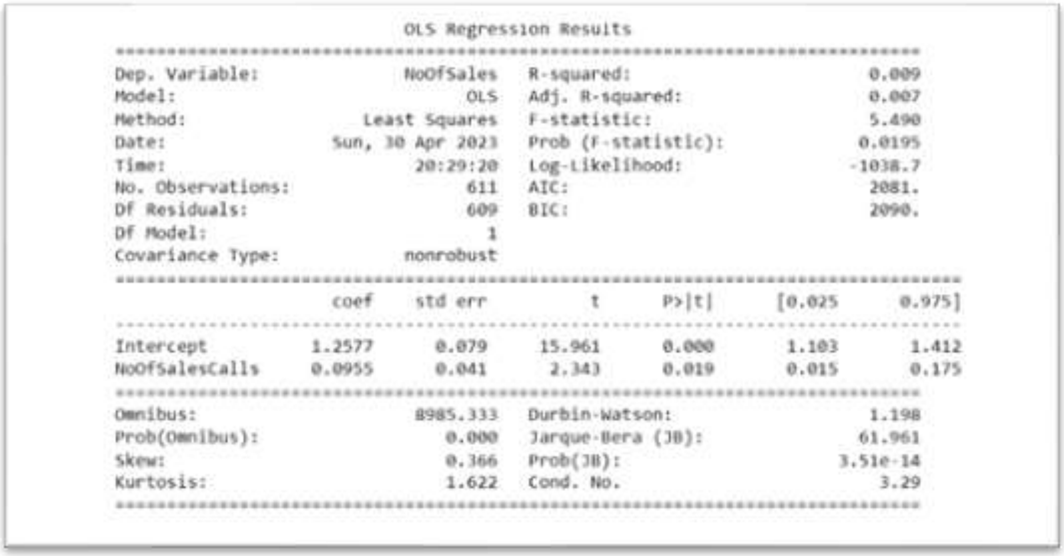


3. Analysis and find the relation:

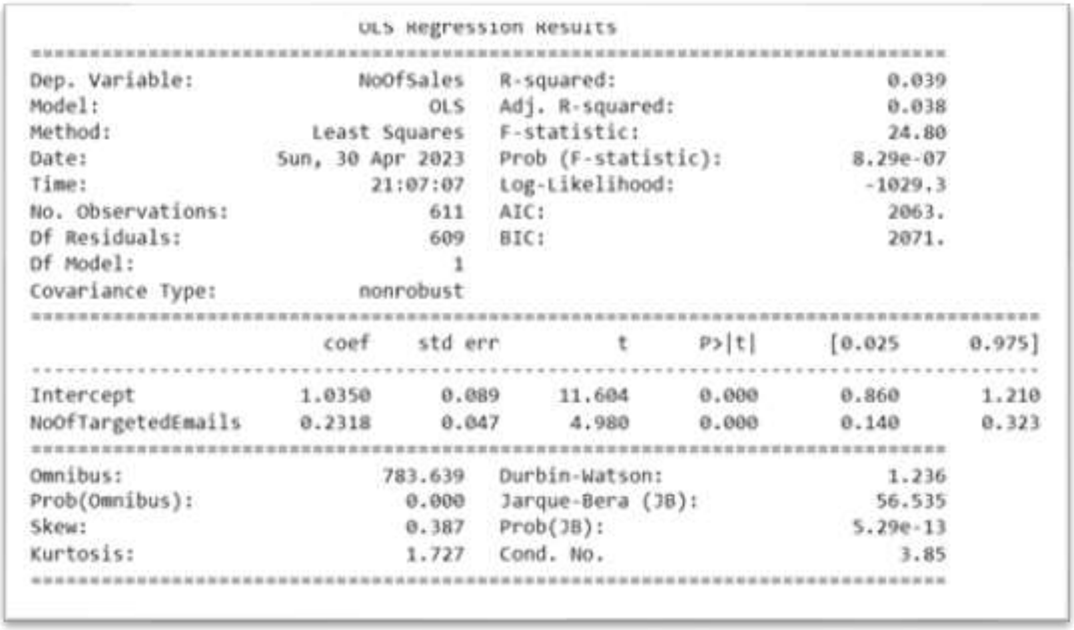
- I decided to use a Heatmap as a tool to visualize the correlation between the variables in the dataset. I used the MinMaxScaler to normalize the data to be on the same scale. The Heatmap graph shows the correlation between all the variables in the dataset. The graph is color-coded with red indicating a positive correlation, blue indicating a negative correlation, and white indicating no correlation. The intensity of the color in each square represents the strength of the correlation coefficient.Upon analyzing the Heatmap, I found that 'NoOfSalesCalls' and 'NoOfTargetedEmails' have a positive correlation, indicating that higher number of emails result in higher number of sales calls. On the other hand, 'NoOfSales' and 'CustomerSatisfaction' have a negative correlation. It can be concluded that the correlation between the number of emails and sales is better than the correlation between the number of sales calls and sales.



- Another technique I used was regression analysis, which helped me find the relationship between NoOfSalesCalls and NoOfSales using the 'statsmodels' library. By finding the coefficients and other statistics for the model, I was able to further support the heatmap technique. The result showed a significant positive relationship between the number of sales calls and the number of sales, with a coef of 0.0955. This means that for each additional sales call made to a customer, the number of sales to that customer is expected to increase by an average of 0.0955. I also performed the same analysis on NoOfTargetedEmails and NoOfSales, which gave me a higher number than the number of sales calls. The coef for this was 0.2318, indicating that for each additional targeted email sent to a customer, the number of sales to that customer is expected to increase by an average of 0.2318. Here are the results:

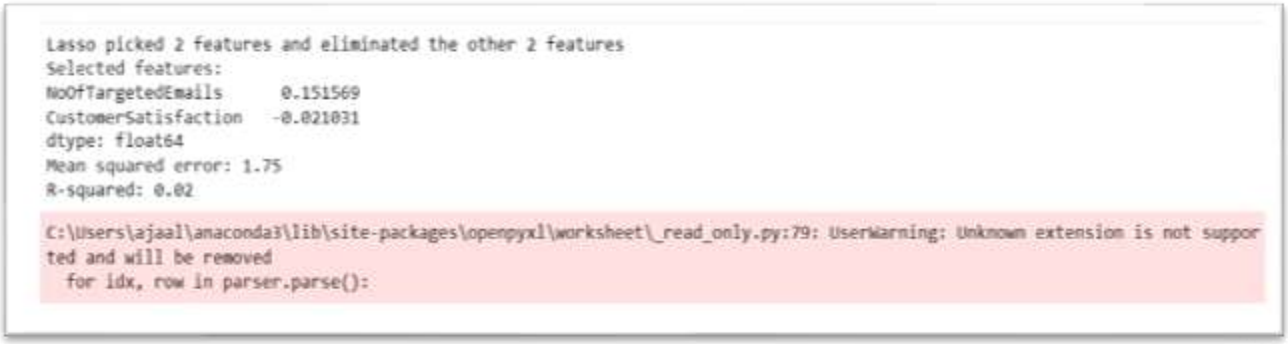


```
                            OLS Regression Results
==============================================================================
Dep. Variable:              NoOfSales   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     5.490
Date:                Sun, 30 Apr 2023   Prob (F-statistic):             0.0195
Time:                        20:29:20   Log-Likelihood:                -1038.7
No. Observations:                 611   AIC:                             2081.
Df Residuals:                     609   BIC:                             2090.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          1.2577      0.079     15.961      0.000       1.103       1.412
NoOfSalesCalls     0.0955      0.041      2.343      0.019       0.015       0.175
==============================================================================
Omnibus:                     8985.333   Durbin-Watson:                   1.198
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               61.961
Skew:                           0.366   Prob(JB):                     3.51e-14
Kurtosis:                       1.622   Cond. No.                         3.29
==============================================================================
```



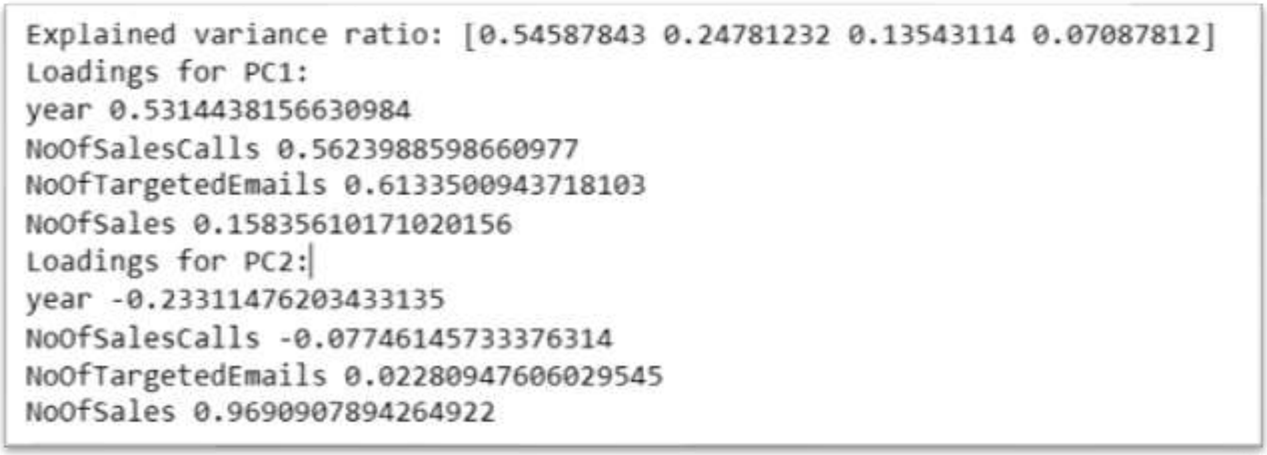
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              NoOfSales   R-squared:                       0.039
Model:                            OLS   Adj. R-squared:                  0.038
Method:                 Least Squares   F-statistic:                     24.80
Date:                Sun, 30 Apr 2023   Prob (F-statistic):           8.29e-07
Time:                        21:07:07   Log-Likelihood:                -1029.3
No. Observations:                 611   AIC:                             2063.
Df Residuals:                     609   BIC:                             2071.
Df Model:                           1
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept              1.0350      0.089     11.604      0.000       0.860       1.210
NoOfTargetedEmails     0.2318      0.047      4.980      0.000       0.140       0.323
==============================================================================
Omnibus:                      783.639   Durbin-Watson:                   1.236
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               56.535
Skew:                           0.387   Prob(JB):                     5.29e-13
Kurtosis:                       1.727   Cond. No.                         3.85
==============================================================================
```

- I used the Lasso regression algorithm to help identify the most important features for predicting the target variable and to perform feature selection by reducing the impact of less important features. Lasso regression selected two features (NoOfTargetedEmails and CustomerSatisfaction) and eliminated the other two features (Year and NoOfSalesCalls). This means that, according to Lasso regression, these two features are not important in predicting the target variable (NoOfSales). Additionally, Lasso regression showed a positive coefficient (0.151569) for NoOfTargetedEmails, indicating that as the number of targeted emails sent to a customer increases, the number of sales to that customer tends to increase as well. There is also a negative coefficient (-0.021031) for CustomerSatisfaction, indicating that as the customer satisfaction index decreases, the number of sales to that customer tends to decrease as well. Therefore, Lasso regression supported the previous results.



```
Lasso picked 2 features and eliminated the other 2 features
Selected features:
NoOfTargetedEmails      0.151569
CustomerSatisfaction   -0.021031
dtype: float64
Mean squared error: 1.75
R-squared: 0.02

C:\Users\ajaal\anaconda3\lib\site-packages\openpyxl\worksheet\_read_only.py:79: UserWarning: Unknown extension is not suppor
ted and will be removed
  for idx, row in parser.parse():
```

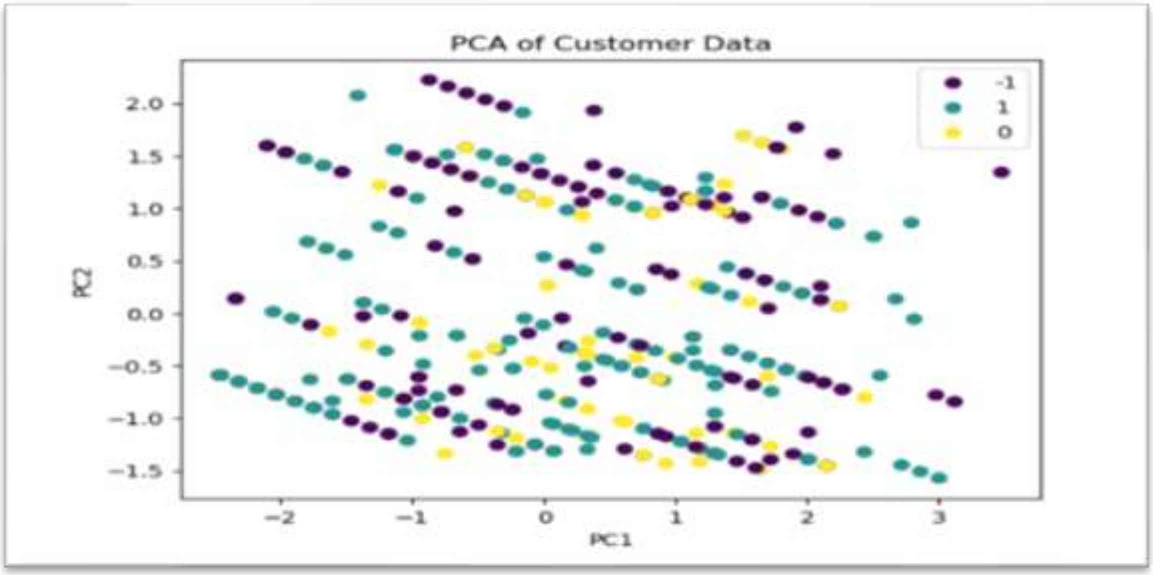
- Also, I used PCA. It will be a useful technique in the dataset and also because there are multiple variables (NoOfSalesCalls, NoOfTargetedEmails , NoOfSales, and CustomerSatisfaction) that may be correlated with each other. Using the PCA helped me to reduce the dimensionality of the data and identify the principal components that capture the most variation in the data and explain the relationships between the variables more simply.



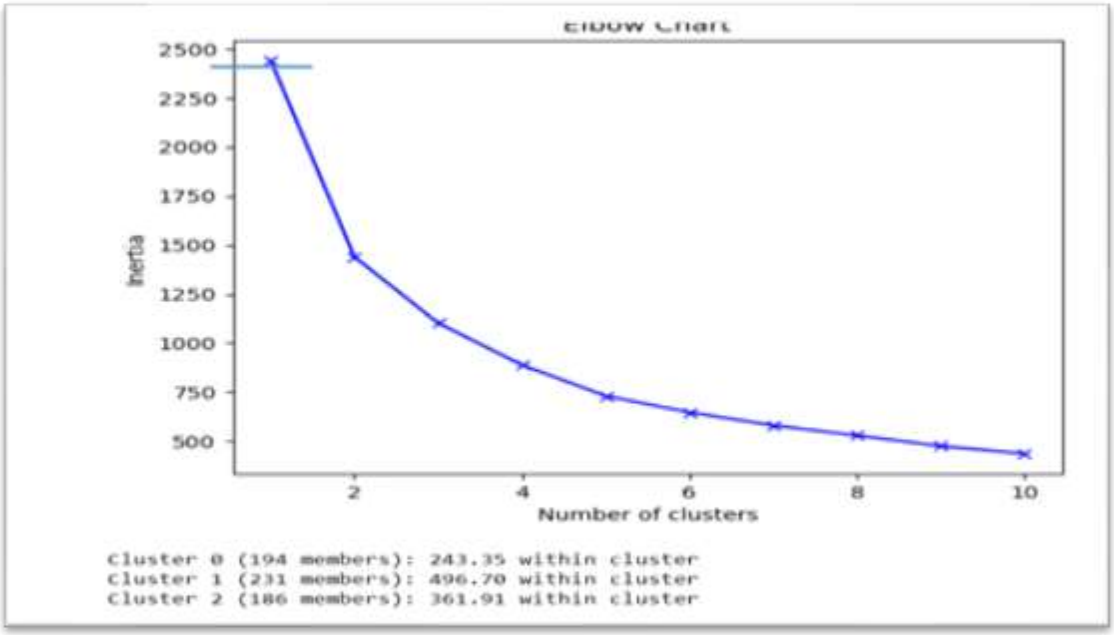
```
Explained variance ratio: [0.54587843 0.24781232 0.13543114 0.07087812]
Loadings for PC1:
year 0.5314438156630984
NoOfSalesCalls 0.5623988598660977
NoOfTargetedEmails 0.6133500943718103
NoOfSales 0.15835610171020156
Loadings for PC2:
year -0.23311476203433135
NoOfSalesCalls -0.07746145733376314
NoOfTargetedEmails 0.02280947606029545
NoOfSales 0.9690907894264922
```

To explain the result: In the first principal component (PC1), it shows that the variables of year, NoOfSalesCalls, NoOfTargetedEmails, and NoOfSales are all positively correlated. This means that when these variables increase together, they tend to be in the same direction. Also, PC1 explains 55% of the total variance, so I can say that it has most of the data. In PC2 the situation is a little bit different. The variable "NoOfSales" has the highest loading on PC2. The variable NoOfTargetedEmails" has a small positive loading on PC2, which means that it has a weak positive correlation with the number of sales in this dataset.

Also, here is the result of the Plot for customer satisfaction based on the PCA.The plot shows the data in a two-dimensional space where the x-axis represents (PC1) and the y-axis represents (PC2). Each point in the plot represents a row in the original dataset, and the color of each point explains the customer satisfaction rating for that row. Every group of points with similar customer satisfaction is clustered together in the plot and they have certain combinations of variables.

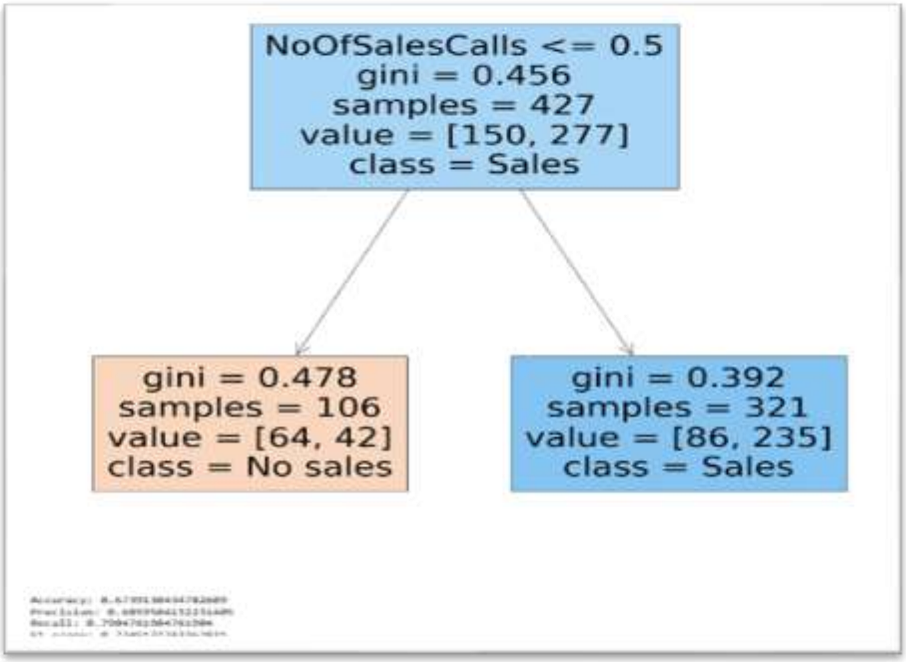


- I decided to use also K-means to cluster the result into 3 groups based on the elbow chart also I got this result of the total squared distance of each data point from the center of its cluster.



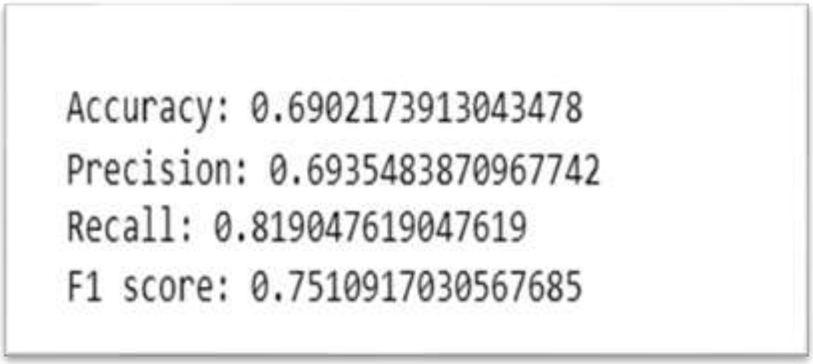
```
Cluster 0 (194 members): 243.35 within cluster
Cluster 1 (231 members): 496.70 within cluster
Cluster 2 (186 members): 361.91 within cluster
```

- Also, I decide to use a decision tree in my research and the accuracy of the model is 0.67, which means that it correctly predicts the class of 67% of the instances in the testing data. The precision of the model is 0.69, which means that when it predicts the positive class. It is correct 69% of the time. The recall of the model is 0.79, which means that it correctly identifies 79% of the positive instances in the testing data. The tree shows that the leaf node with the label "Class No Sales" represents customers who are unlikely to make a purchase based on the given features and the leaf node with the label "Class Sales" represents customers who are likely to make a purchase based on the given features.



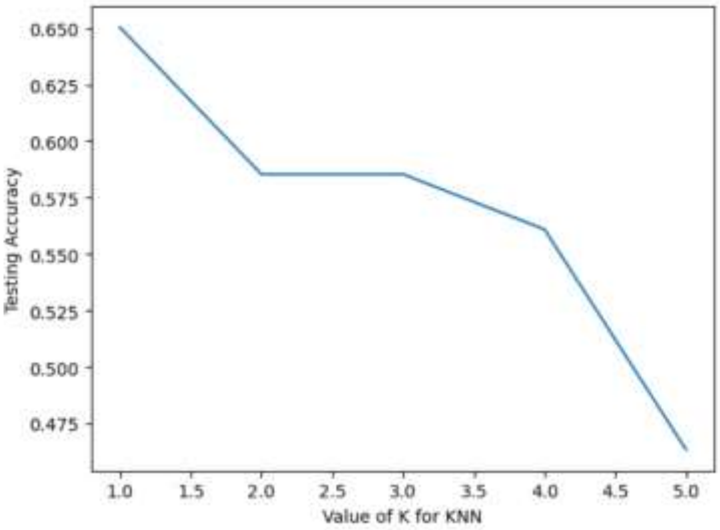
```
Accuracy: 0.6739130434782609
Precision: [illegible]
Recall: 0.7047619047619048
F1 score: [illegible]
```

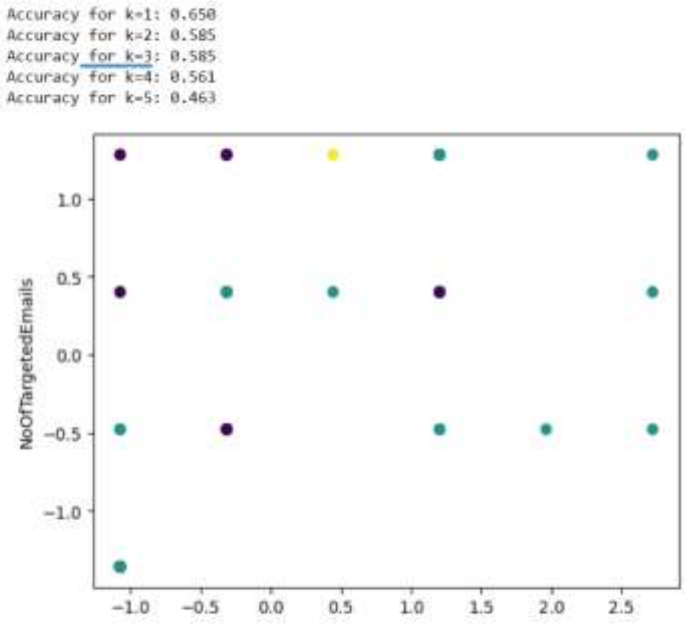
- After that, I decided to increase the accuracy of the model by decreasing the max_depth to (5) and min_samples_split to (3) and I got this result:



```
Accuracy: 0.6902173913043478
Precision: 0.6935483870967742
Recall: 0.819047619047619
F1 score: 0.7510917030567685
```

- I decided to use KNN to classify the database based on CustomerSatisfaction and gain insights into the data. To obtain the best accuracy, I varied the number of K from 1 to 5. I visualized the results using a scatter plot of KNN and a line plot of the accuracy.:





```
Accuracy for k=1: 0.650
Accuracy for k=2: 0.585
Accuracy for k=3: 0.585
Accuracy for k=4: 0.561
Accuracy for k=5: 0.463
```

It’s clear that the result shows that in k=1, the model achieved an accuracy of 0.650, indicating that 65% of the test set was correctly classified. As the value of k increased to 2 and 3, the accuracy dropped slightly to 0.585, suggesting that including more neighbors did not significantly improve the model's performance. However, when the number of neighbors was increased to 4, the accuracy decreased further to 0.561, suggesting that including too many neighbors may not be helpful for this particular problem. Finally, when k was set to 5, the model's accuracy dropped significantly to 0.463, indicating that the model was not performing well when more neighbors were considered.

4. The conclusions:

After doing all of these technics on the dataset I got this result. Increasing the frequency of sales calls and targeted emails has a direct impact on the number of sales made to a customer. In other words, customers are more likely to make a purchase when they receive a higher number of sales calls and targeted emails. Also, using the PCA technique helped me to find that the results of PCA indicate that the first principal component is strongly influenced by the number of sales calls, targeted emails, and the year of operation, which means a significant correlation among these variables.

The result that I got after using the decision tree achieved an accuracy of 69%, which is not bad and gives me a graph to explain the result.

Also, I used Lasso regression which helped me to identify NoOfTargetedEmails and CustomerSatisfaction as the most important features for predicting sales. This gives me the idea that sending targeted emails and ensuring customer satisfaction may be particularly important for driving sales.

Also, I used K-means clustering to identify any group of clustering and I got three groups of customers with different patterns of sales calls, targeted emails, and the number of sales. This could be useful for understanding each group and each group has unique characteristics.

As all the above points have demonstrated that sales calls, targeted emails, and customer satisfaction are very important in my analysis and the dataset for driving sales in the medical device.