A blue and white sign with white letters

Description automatically generated

# School of IT & Business Technologies

# Graduate Diploma in Data Analytics (Level 7)

# Cover Sheet and Student Declaration

This sheet must be signed by the student and attached to the submitted assessment.

|  |  |  |  |
| --- | --- | --- | --- |
| **Course Title:** | **Machine Learning and AI** | **Course code:** | GDDA708 |
| **Student Name:** | Aju Peter | **Student ID:** | 764706847 |
| **Assessment No & Type:** | Summative Assessment 1[Project] | **Cohort:** |  |
| **Due Date:** | 09-02-2024 | **Date** **Submitted:** | 09-02-2024 |
| **Tutor’s Name:** | Harsh Tiwari |  |  |
| **Assessment** **Weighting** | 40% |  |  |
| **Total Marks** | 100 |  |  |

**Student Declaration:**

I declare that:

* I have read the New Zealand School of Education Ltd policies and regulations on assessments and understand what plagiarism is.
* I am aware of the penalties for cheating and plagiarism as laid down by the New Zealand School of Education Ltd.
* This is an original assessment and is entirely my own work.
* Where I have quoted or made use of the ideas of other writers, I have acknowledged the source.
* This assessment has been prepared exclusively for this course and has not been or will not be submitted as assessed work in any other course.
* It has been explained to me that this assessment may be used by NZSE Ltd, for internal and/or external moderation.

**Student signature:** Aju Peter

**Date:** 09-02-2024

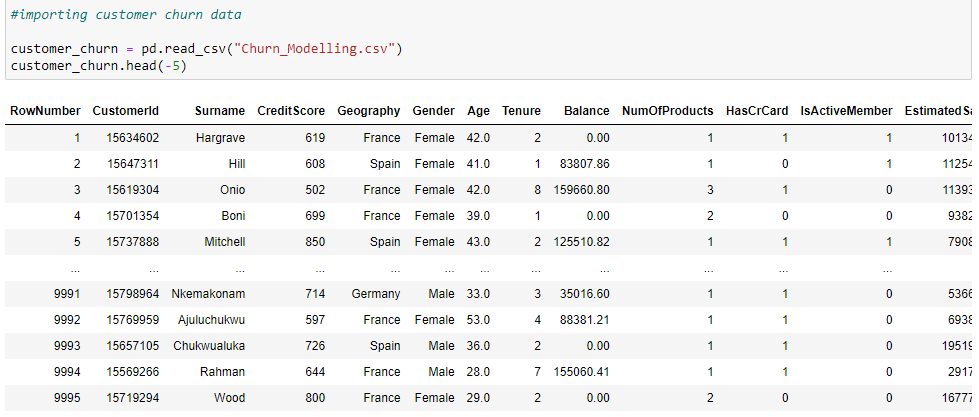
|  |  |  |
| --- | --- | --- |
| **Tutor only to complete** |  | |
| **Assessment results:** | **Part A** **Part B Part C**  (max. 25 marks) (max. 25 marks)(max. 50 marks) | |
| **Total Marks: /100** | **Grade:** |

## PART A – Business Decision Making using Supervised Machine Learning Methods (Classification)

**Scenario: Analyse and predict Bank customer churn**

### Task 1 Data Preparation

1. To perform this entire analysis, I have chosen Jupyter Notebook. Firstly, I successfully imported the required libraries and then imported the Bank customer churn data.



1. Applied two techniques to handle missing values.
2. Identifying and Handling Missing Values:

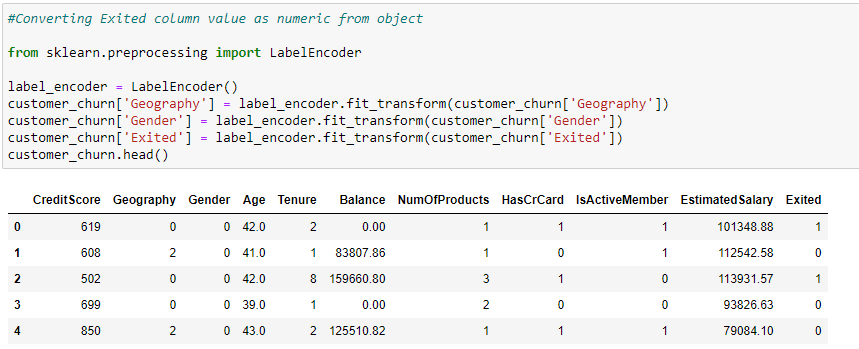
I used the isnull().sum() method to find any null values present in the dataset. I found some missing values in the 'Age' column and used the ‘fillna’ function to fill the data.

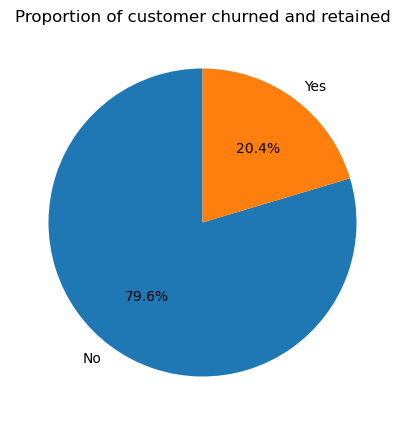
1. Dropped Irrelevant Columns:

I used the drop() method to remove irrelevant columns Row Number, Customer ID and Surname. This method allowed me to drop columns that are not needed for my analysis, thus simplifying the dataset.

By performing these data cleaning methods, I ensured that the dataset is free from missing values, incorrect entries, and irrelevant columns. This preparation is essential for further analysis tasks in my Jupyter Notebook.

1. I have used one effective method to convert categorical data to numerical data for better analysis.



1. Applied three specific data visualisation techniques to analyse data distribution and patterns.
2. Pie Chart: The pie chart provides a clear overview of customer retained and churned.

80% of customers choose to stay with the bank and only 20% exited.

1. Pair Plot: from the below pair plot we can identify patterns, trends, and potential correlations between variables, allowing to quickly spot interesting relationships or dependencies in the data.

A screenshot of a graph

Description automatically generated

1. Bar diagram: To review the customer retention and churn relation with categorical variables.

A group of blue and orange bars

Description automatically generated

These visualization techniques provide insights into the distribution of data and patterns within the dataset, aiding in understanding customer behaviour and potential factors influencing churn.

Observations:

* Retention rate is high in customer from France.
* High churn rate shows from female customers.
* Customers who have credit cards shows high retention.
* Customers who all are inactive shows higher churn rate.

### Task 2 Feature Engineering

a. Her I have applied 2 most effective feature selection method.

1. Correlation method: Correlation matrix shows weak positive correlation and weak negative correlation.

Features Correlations with Customer Churn:

Exited 1.000000

Age 0.285278

Balance 0.118533

Geography 0.035758

EstimatedSalary 0.012097

HasCrCard -0.007138

Tenure -0.014001

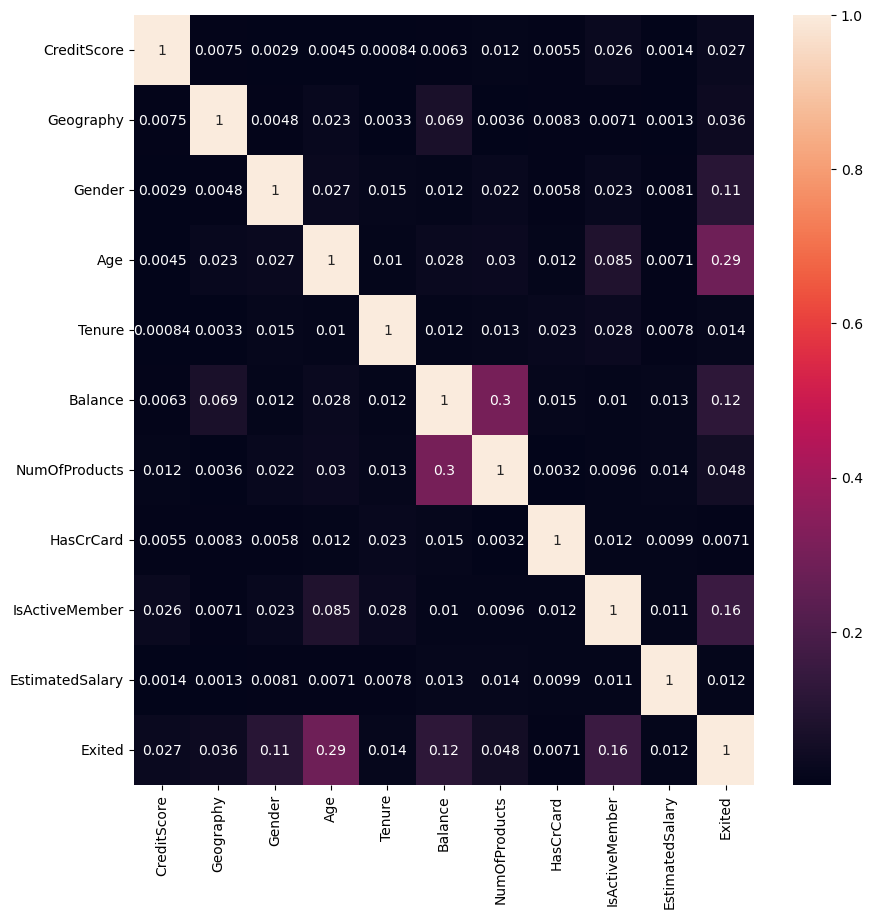
CreditScore -0.027094

NumOfProducts -0.047820

Gender -0.106512

IsActiveMember -0.156128

Name: Exited, dtype: float64



A graph with different colored bars

Description automatically generated

2. Random Forest Classifier: I have used this method to find 5 most correlated features for further analysis.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

b. I have successfully implemented the Standard Scaler method to standardize all feature values within the range of -1 to 1. This approach can be applied to identify and select relevant features for the prediction model on the selected datasetA screenshot of a computer

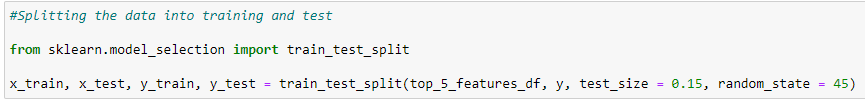
Description automatically generated

### Task 3 Model building and prediction

1. I have selected three supervised machine learning model(s) using the chosen dataset.

* Logistic Regression
* SVC (Support Vector Classifier)
* Decision Tree Classifier

To evaluate the three supervised machine learning and their performance, I divided the dataset into training (85%) and testing (15%) sets using the train-test-split method. This approach ensures adequate data for training while providing a separate set for evaluating model performance. This process enables a comparative analysis of the models' predictive capabilities and assists in selecting the most suitable one for the dataset.



1. Logistic Regression

Accuracy of logistic regression classifier on test set: 0.77

A screen shot of a computer

Description automatically generated

1. SVC (Support Vector Classifier)

Accuracy of SVC (RBF) classifier on test set: 0.84

A white screen with black text

Description automatically generated

1. Decision Tree Classifier

Accuracy of Decision Tree Classifier: 0.812

A screen shot of a computer

Description automatically generated

1. Summary:

Based on the analysis of the predictive system and the performance of three different supervised machine learning models Logistic Regression, Support Vector Classifier with RBF kernel, and Decision Tree Classifier. The accuracy scores on the test are follows.

* Accuracy of logistic regression classifier on test set: 0.77
* Accuracy of SVC (RBF) classifier on test set: 0.84
* Accuracy of Decision Tree Classifier: 0.812

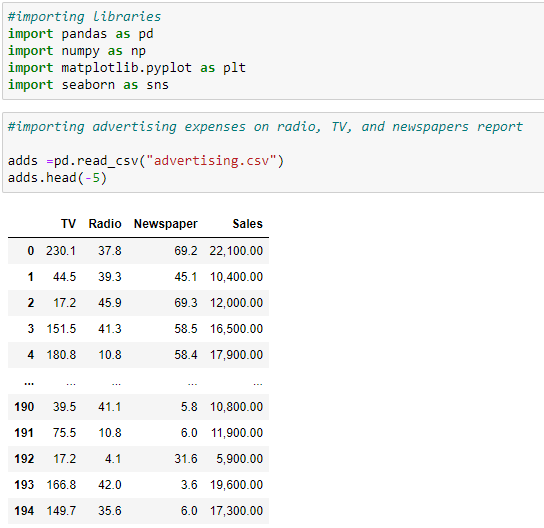
Among these models, the Support Vector Classifier with RBF kernel demonstrated the highest accuracy score of 0.84, making it the preferred choice for building predictive models.

## 

## PART B – Business Decision Making using Supervised Machine Learning Methods (Regression)

### Scenario: Advertising expenses on radio, TV, and newspapers

### Task 1 Data Preparation

1. To perform this entire analysis, I have chosen Jupyter Notebook. Firstly, I successfully imported the required libraries and then imported the Advertising expenses data. I thoroughly checked the data and used data cleaning methods.

1. I used the isnull().sum() method to find any null values present in the dataset.

By performing this data cleaning methods, I ensured that the dataset is free from missing values, incorrect entries, and irrelevant columns. This preparation is essential for further analysis and modeling tasks in my Jupyter Notebook.

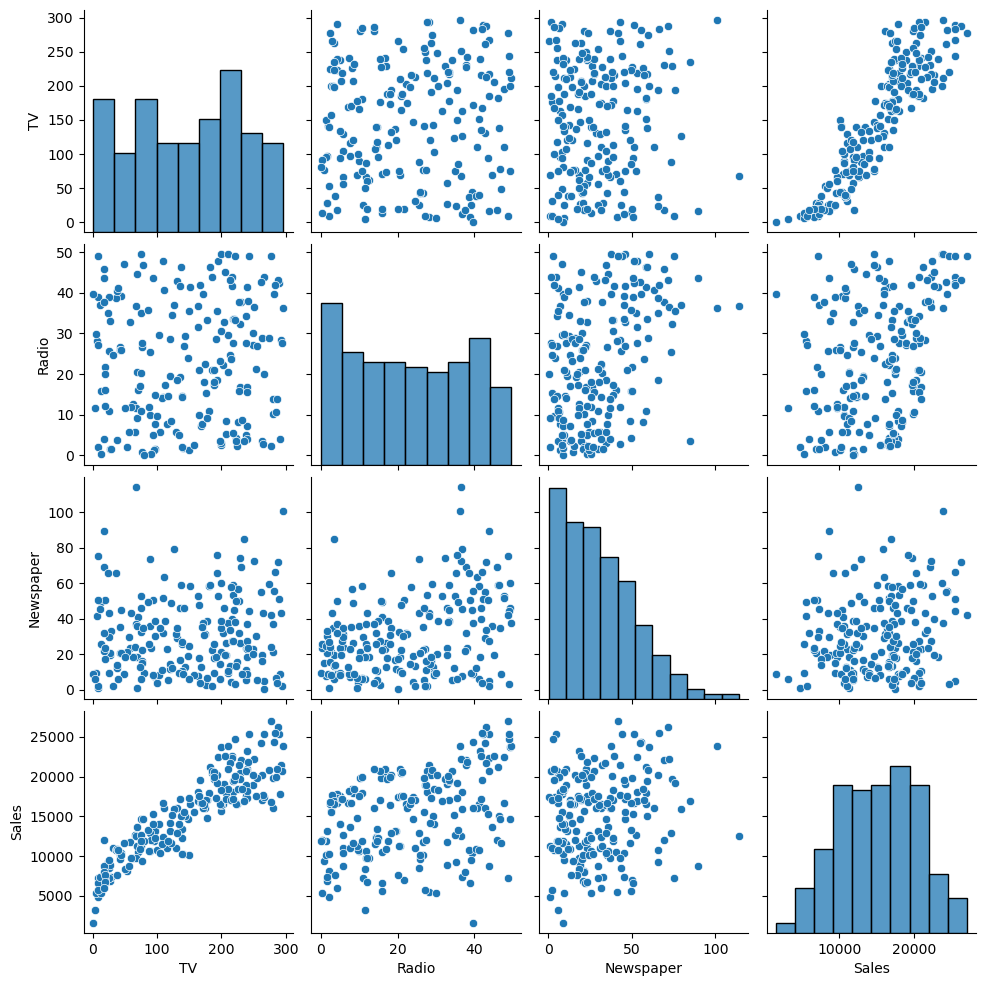
1. categorical to numerical variables.

I found that numerical values are represented as objects. To convert them to integers, I used the strip function to remove commas from the dataset.



1. Applied two specific data visualisation techniques to analyse data distribution and patterns.

1. Pair Plot: from the below pair plot we can identify patterns, trends, and potential correlations between variables, allowing to quickly spot interesting relationships or dependencies in the data.



1. Heatmap: From the heatmap below, we can clearly identify that TV advertising expenses are highly correlated with sales, and Radio advertising expenses also show a positive correlation. Newspaper advertising expenses exhibit a very low positive correlation with sales.A chart of a television show

   Description automatically generated with medium confidence

3. Scatter Plot: From the below scatter plot we can identify patterns, trends, and potential correlations between variables.

A diagram of a scatter plot

Description automatically generated

### Task 2 Feature Engineering

1. I have successfully applied Linear Regression Feature selection method to choose relevant features for the prediction model on the selected dataset.

A screenshot of a computer program

Description automatically generated

Selected Features: Index(['TV', 'Radio'], dtype='object')

1. I have used Min Max Scaler method on the subset of features from the dataset.

**A screenshot of a computer

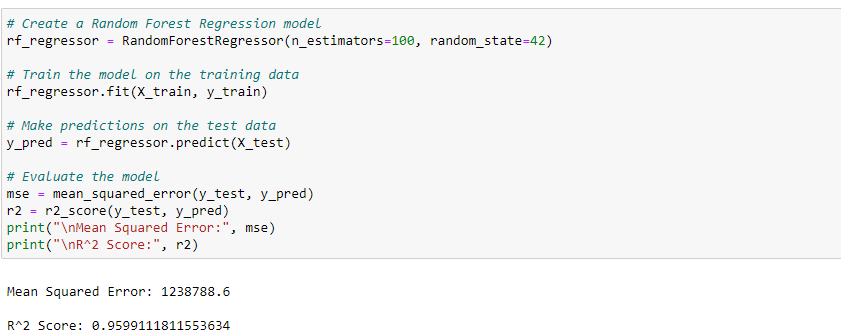
Description automatically generated**

### Task 3 Model building and prediction

1. I have used 3 supervised machine learning models using the chosen dataset.
2. Random Forest Regression:

Mean Squared Error: 1238788.6

R^2 Score: 0.9599111811553634



1. Linear Regression:

Mean Squared Error: 2846616.122131541

R^2 Score: 0.9078797802624651

A screenshot of a computer program

Description automatically generated

1. SVR- Support Vector Regression:

Mean Squared Error: 34262991.01562693

R^2 Score: -0.10879518902650487

A screenshot of a computer program

Description automatically generated

b. Summary:

For SVR, the MSE is high (34262991.0156) and the R^2 score is negative (-0.1088), indicating poor model performance and a weak fit to the data.

Linear Regression performs better with a lower MSE (2846616.1221) and a higher R^2 score (0.9079), suggesting a relatively good fit to the data.

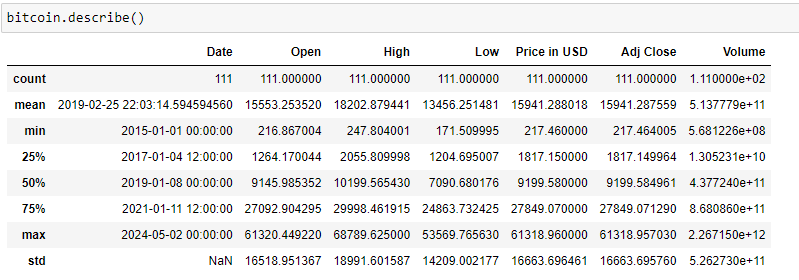
Random Forest Regression has a better performance with the lowest MSE (1238788.6) and the highest R^2 score (0.9599), indicating superior predictive accuracy and a strong relationship between the variables.

## PART C – Time Series Trend Analysis and Forecasting

**Scenario: Discover historical prices of Bitcoin USD (BTC-USD)**

### Task 1: Data Exploration

To perform this entire analysis, I have chosen Jupyter Notebook. Firstly, I successfully imported the required libraries and then imported the Bitcoin Price data. I thoroughly checked the data and used data cleaning methods.



Based on the descriptive analysis, we can see that the dataset spans from 2015 to 2024 and contains historical data. The average price of Bitcoin over this period is approximately USD 16,000, with the maximum price reaching USD 61,000. This substantial difference in prices over the years indicates significant changes in the Bitcoin market. Especially, the minimum price recorded is surprisingly low at USD 217.0.

### Task 2: Trend Analysis:

Trend analysis shows a substantial difference in prices over the years indicates significant changes in the Bitcoin market. From starting at USD 217 and reaching up to USD 45,000 now.

A graph of a line graph

Description automatically generated with medium confidence

A graph of a price

Description automatically generated with medium confidence

### Task 3: Seasonality Assessment

Conducted a detailed analysis to identify and describe any seasonal patterns in the data.

How the seasons affect the price of Bitcoin, and their impact on Bitcoin price.A graph of different types of data

Description automatically generated with medium confidence

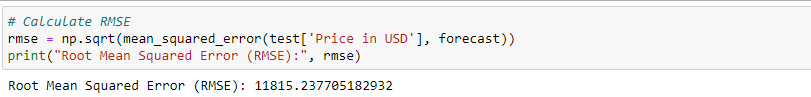
|  |  |  |
| --- | --- | --- |
| **Task 4: Anomaly Detection**  I could see significant anomalies in the Bitcoin historical price movements. And detected significant deviation from the expected behaviour in the price data. To identify the animalities we need more deeper analysis on this subject. |  |  |
| **Task 5: Predictions and Recommendation** |  |  |

I have used the SARIMA model for Bitcoin price prediction and calculated the Root Mean Square Error value, and the value is in a higher side which indicates greater prediction variability. In this case, the RMSE value of 11815.24 suggests that the time series model's predictions may have moderate to significant variability compared to the actual Bitcoin price data.

A graph showing the growth of bitcoin price

Description automatically generated

A graph showing the price of bitcoin

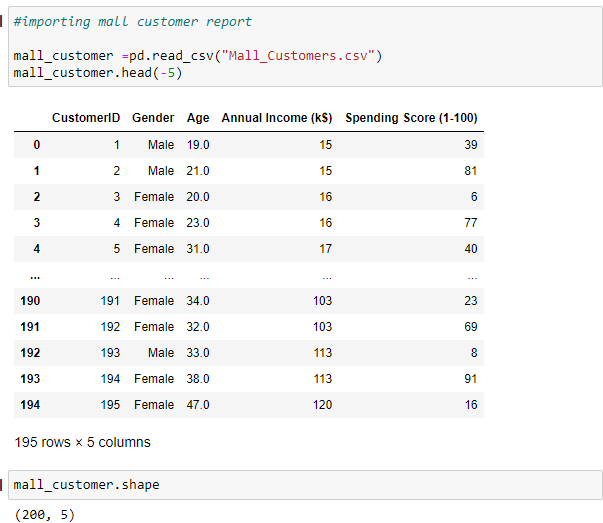
Description automatically generated

## PART D – Business Decision Making using Un-Supervised Machine Learning

**Methods (Clustering)**

#### Task 1 Data Preparation

1. To perform this entire analysis, I have chosen Jupyter Notebook. Firstly, I successfully imported the required libraries and then imported the mall customer segmentation data. I thoroughly checked the data and used three major data cleaning methods.



1. Identifying and Handling Missing Values:

I used the isnull().sum() method to find any null values present in the dataset. I found some missing values in the 'Age' column and used the mean function to fill the data.

1. Identifying and Handling Incorrect Entries:

I utilized the unique() function to identify any wrong entries in the dataset. I found one incorrect entry in the 'Gender' column. I noticed that there is a numeric value in the 'Gender' column, and I replaced the numeric value with an object using the replace() function.

1. Dropping Irrelevant Columns:

I used the drop() method to remove irrelevant columns from the dataset. This method allowed me to drop columns that are not needed for my analysis, thus simplifying the dataset.

1. Convert object to numeric for successful analysis.

I used Label Encoder function to convert object/ string to numeric in the dataset. This method allows me to convert Gender columns as numeric. This method is essential for further analysis.

A table with numbers and text

Description automatically generated

By performing these data cleaning methods, I ensured that the dataset is free from missing values, incorrect entries, and irrelevant columns. This preparation is essential for further analysis and modelling tasks in my Jupyter Notebook.

1. Applied three data exploration techniques.
2. Descriptive analysis

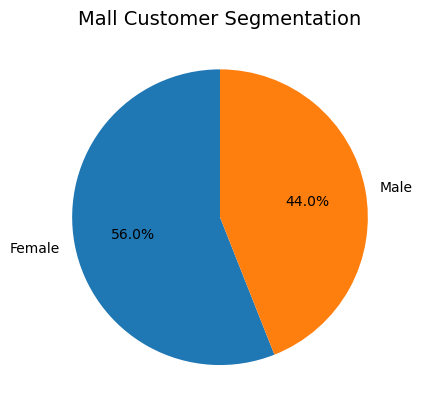
Below is the overall descriptive analysis of the mall customer data. We have a total of 200 records in this dataset, which includes information on Gender, Age, Annual Income, and Spending Score.

* We can observe that the average annual income of customers is $60,000.
* Average spending score of customers is 50
* The majority of customers fall within the age range of 20 to 50.
* Average age of customers is 39

Further analysis can provide deeper insights into customer behaviour and preferences.

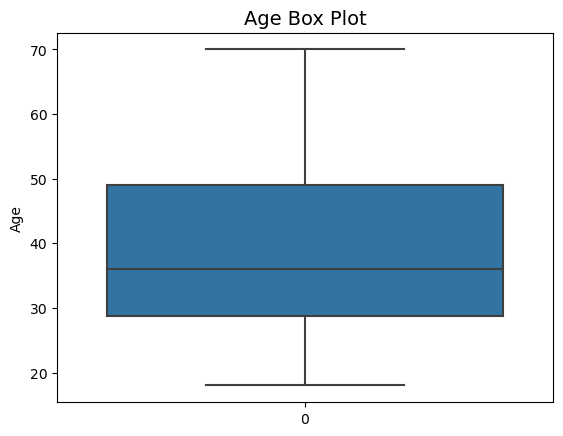
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
| count | 200 | 200 | 200 | 200 |
| mean | 0.44 | 38.86935 | 60.56 | 50.2 |
| std | 0.49763 | 13.96633 | 26.264721 | 25.823522 |
| min | 0 | 18 | 15 | 1 |
| 25% | 0 | 28.75 | 41.5 | 34.75 |
| 50% | 0 | 36 | 61.5 | 50 |
| 75% | 1 | 49 | 78 | 73 |
| max | 1 | 70 | 137 | 99 |

1. A graph of age distribution

   Description automatically generatedData Visualisation

Histogram

The histogram chart provides an overview of the age distribution in the mall customer data. It reveals that the majority of customers are youngsters (20-50). This insight highlights the significant representation of youngsters in the mall, suggesting there is potential opportunities can be targeted from this segment.

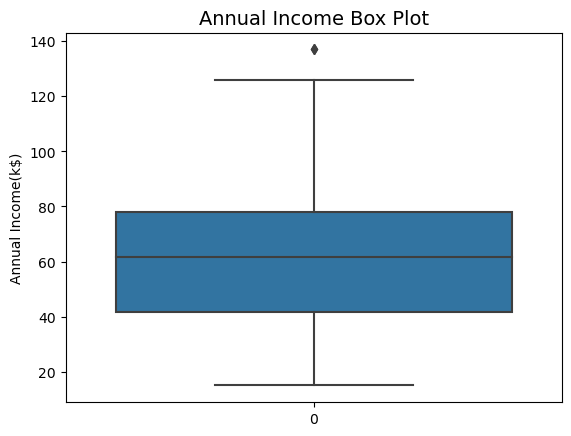
A graph of blue bars

Description automatically generated

Pie Chart

The pie chart provides an overview of the gender segmentation in the mall customer data. It reveals that the majority of customers are female, comprising 56.0% of the total customer base. This insight highlights the significant representation of female customers in the mall, suggesting there is potential opportunities can be targeted from this segment.

A blue and orange colored lines

Description automatically generated with medium confidenceA graph of a number of bars

Description automatically generatedA blue and orange diagram

Description automatically generated

1. Correlation Analysis.

A chart of a number of people

Description automatically generated with medium confidenceI have applied various correlation methods, including a heatmap and correlation matrix.

The heatmap did not reveal significant correlations among the variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
| Gender | 1 | 0.062446 | 0.05641 | -0.058109 |
| Age | 0.062446 | 1 | -0.014338 | -0.328109 |
| Annual Income (k$) | 0.05641 | -0.01434 | 1 | 0.009903 |
| Spending Score (1-100) | -0.05811 | -0.32811 | 0.009903 | 1 |

The

correlation matrix identified several weak correlations:

* A weak positive correlation (0.062) exists between Gender and Age.
* A weak negative correlation (-0.058) is observed between Gender and Spending Score.
* Similarly, a weak negative correlation (-0.328) is noted between Age and Spending Score.
* Moreover, correlations between Annual Income and other variables are also weak, with coefficients close to 0.

A group of graphs showing different sizes of distribution

Description automatically generated with medium confidence

Pair Plot Analysis: Pair plots were utilized to explore the relationships between variables.

* Pair plot indicated that individuals aged 20-50 tend to have higher annual incomes and spending scores.
* Spending score isn’t affecting high or low annual income.

A diagram of a scatter plot

Description automatically generatedA diagram of a scatter plot

Description automatically generatedA diagram of red dots

Description automatically generated

Scatter Plot Analysis: Further insights were obtained through scatter plots, which identified correlations between pairs of variables. the Age vs. Annual Income scatter plot revealed that individuals aged 30-50 typically have higher annual incomes. Additionally, it suggested that younger individuals tend to earn more and spend more.

In summary of correlation methods revealed weak associations between variables, pair plots and scatter plots provided deeper insights into the relationships and distributions within the dataset. These visualizations highlighted trends such as higher earnings and spending patterns among individuals aged 30-50, contributing to a better understanding of consumer behaviour within the dataset.

### Task 2 Unsupervised Algorithm Implementation

1. I have used below 3 unsupervised machine learning algorithms.
2. K-Means
3. Hierarchical
4. DBSCAN
5. K-Means Clustering

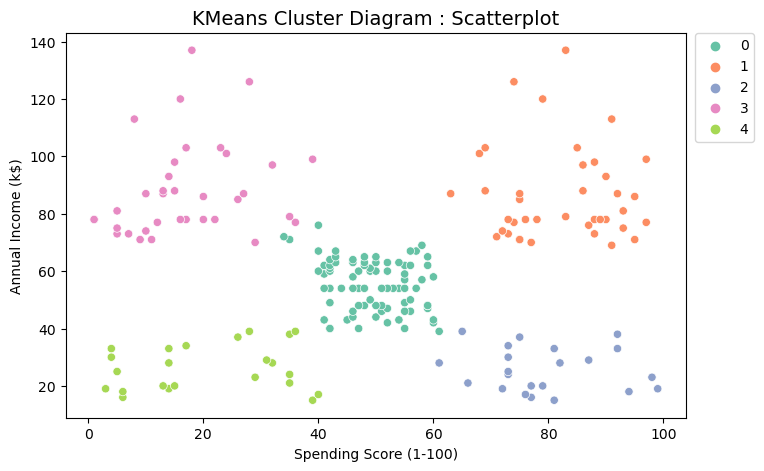
A white rectangular object with black text

Description automatically generatedThe Elbow method was used to determine the optimal number of clusters. From the plot, it's evident that the most significant slope occurs at k = 5. Therefore, the dataset was clustered into 5 distinct groups. Additionally, the knee locator confirmed the presence of 5 clusters. The Silhouette Coefficient method also suggested 5 as the ideal number of clusters.

A line graph with numbers and symbols

Description automatically generatedA graph with blue dots and white text

Description automatically generatedA K-Means model was then fitted with 5 clusters, and the resulting clusters were visualized using a scatter plot.



Observations from K-Means Clusters:

Customers with low annual income tend to have a low spending score.

Customers with high annual income exhibit varying spending behaviours, with some having a low spending score.

There is a cluster of customers with high annual income and high spending score.

Another cluster consists of customers with low annual income but high spending score.

There is also a group of customers with average annual income and average spending score.

1. Hierarchical Clustering:

The method begins with normalizing the data using MinMaxScaler. Subsequently, a dendrogram is constructed to determine the optimal number of clusters. The dendrogram diagram suggests the presence of 5 major clusters within the data.

Next, a new dataframe is created with an additional column indicating the cluster assignments. This appended dataframe is then used to visualize the hierarchical clusters through a scatter plot.

A diagram of multiple colored dots

Description automatically generated

Observations from Hierarchical Clustering:

Customers with low annual income tend to have a low spending score.

Customers with high annual income exhibit varying spending behaviours, with some displaying a low spending score.

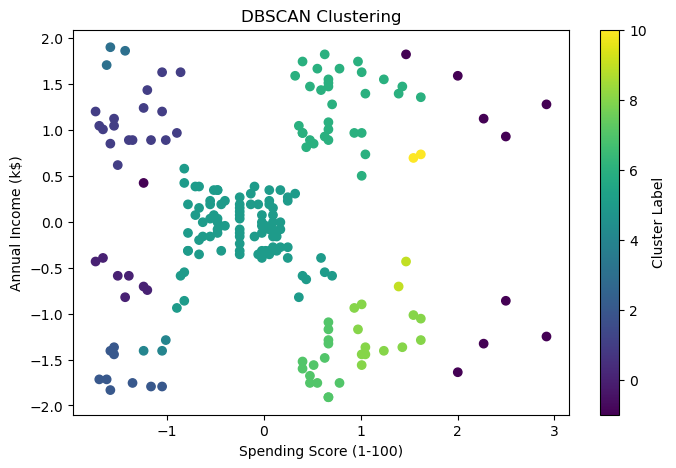
A distinct cluster consists of customers with high annual income and high spending scores.

Another cluster comprises customers with low annual income but high spending scores.

Additionally, there is a group of customers with average annual income and average spending scores.

DBSCAN Clustering

DBSCAN clustering was performed on the scaled data using StandardScaler. The resulting clusters from DBSCAN, Hierarchical clustering, and K-Means clustering algorithms were found to be consistent.



The DBSCAN algorithm identified the following 5 clusters:

Clients with low annual income and high spending score.

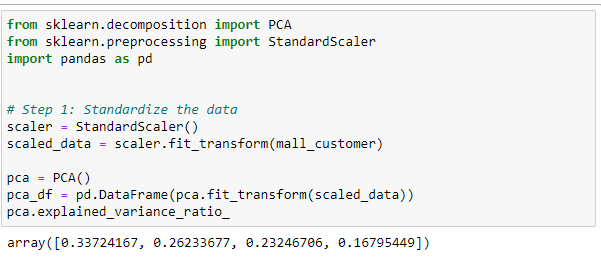
Clients with medium annual income and medium spending score.

Clients with high annual income and low spending score.

Clients with high annual income and high spending score.

Clients with low annual income and low spending score.

1. Applied Principal Component Analysis (PCA) to reduce the dimensionality of a given dataset.



A graph with blue lines

Description automatically generated

I have used PCA analysis for dimensional reduction below is the cumulative variance ratio and dimensionally reduced data.

1. 2 examples of visualisation types

There are two examples of visualization types that can effectively represent data.

Here I have used,

1. Cluster scatter plot: This visualization allows a clear separation between clusters and helps identify any patterns or trends within each cluster. Additionally, it provides insights into the distribution of data points across different clusters.

I have used the clustered scatter plot to visualize the clustering algorithms such as K-Means, DBSCAN, and Hierarchical clustering. Each point represents a customer, and clusters are differentiated by colour. This helps identify customer segments based on their characteristics such as annual income and spending score.

1. Elbow Plot and Silhouette Plot: It helps in determining the optimal number of clusters for a given dataset.

The plot typically shows a decrease in inertia as the number of clusters increases. However, the rate of decrease tends to diminish after a certain point. These visualization aids in fine-tuning clustering parameters and selecting the optimal number of clusters.

#### Task 3 Conclusion

Identified Customer Segments: The segmentation analysis revealed distinct customer segments based on their annual income and spending behaviour. These segments include:

* + Customers with low annual income and high spending score.
  + Customers with medium annual income and medium spending score.
  + Customers with high annual income and low spending score.
  + Customers with high annual income and high spending score.
  + Customers with low annual income and low spending score.

The segmentation results were consistent across multiple clustering algorithms, including K-Means, Hierarchical clustering, and DBSCAN. Each algorithm identified similar customer segments, indicating robustness and reliability in the segmentation process. Hence, these clustering helps the mall in understanding the various customer groups and ways to improve the customer's spending score that helps the business of the mall.