## **Uzma Younas**

## **04071813047**

## **IML#1**

## 

## 

## **Customer Categorization**

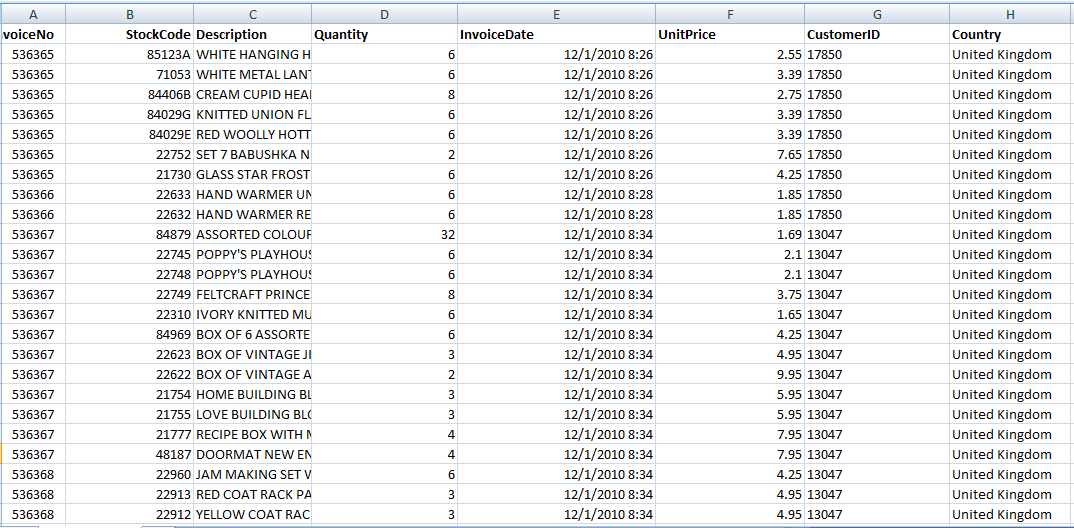
### 

### **Introduction**

The customer categorization model aims to personalize suggestions to customers of an e-commerce site based on spending patterns. It enables the company to sell more products with less marketing expenditure and recommend products or services to customers under certain price ranges. The goal is to develop a model to categorize the customers in a particular segment based on their buying patterns

### **Dataset**

Sample of the taken dataset.



### 

### 

### 

### **Attributes of the dataset**

These are the attributes in this dataset with a description of each:

1. **InvoiceNo:** A unique Invoice ID for each transaction.
2. **StockCode:** Product/item code.
3. **Description:** Description of the Product.
4. **Quantity:** Quantities for each product per transaction.
5. **InvoiceDate:** date and time for an invoice or transaction.
6. **UnitPrice:** Price of a Product per unit.
7. **CustomerID:** ID for Customer identification.
8. **Country:** The country name of the buyers.

### 

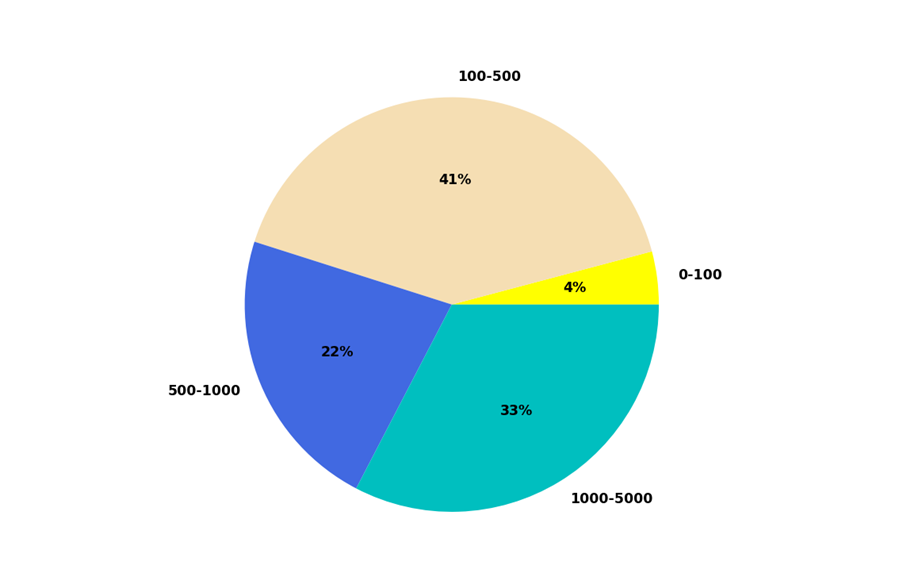
### **Data Pre-processing:**

* Created a new column **Amount** from the product of **Quantity** and **UnitPrice** for a single invoice.
* Calculated **Net Spendings** on a particular transaction by adding up all the **Amount** associated with a particular Customer Invoice.
* Divided the customer into four categories based on the net spending. The pie chart below illustrates 4 types of target customers with the spending ranges of each mentioned.

### 

Initial EDA on dataset

It is obvious from the graph we have a few of transactions which creates problems in scaling. So we will consoder only values under 5000 range as maximum transactions are under this range.



Limit the data in under 5000 range

### 

### **Feature Vectors and Target Class:**

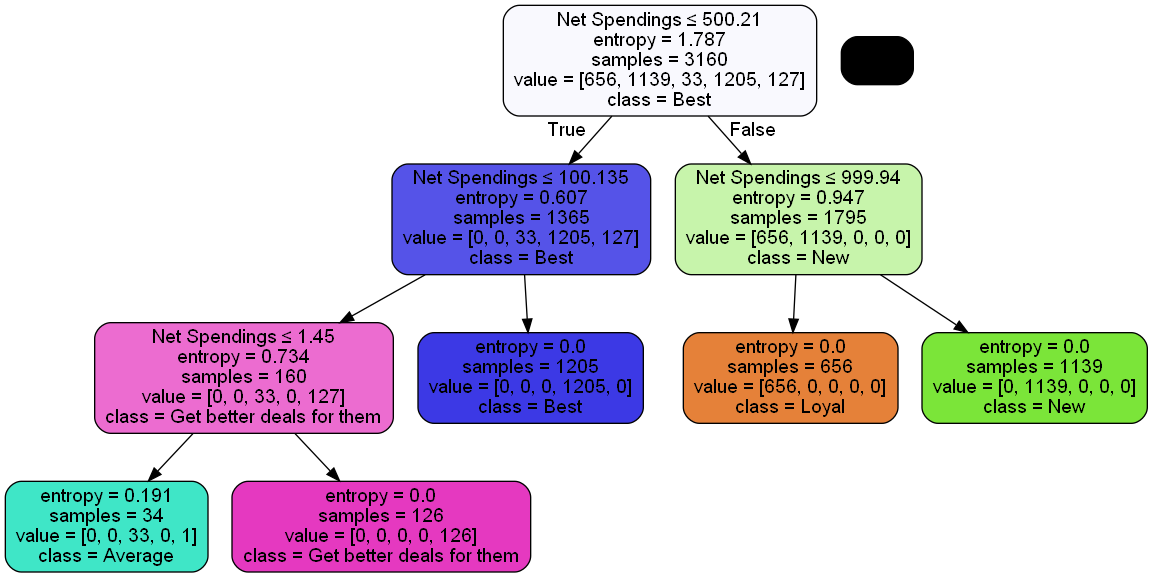
**Feature** CustomerID, InvoiceNo, Quantity, UnitPrice, Amount, Net Spendings

**Target** CustomerType

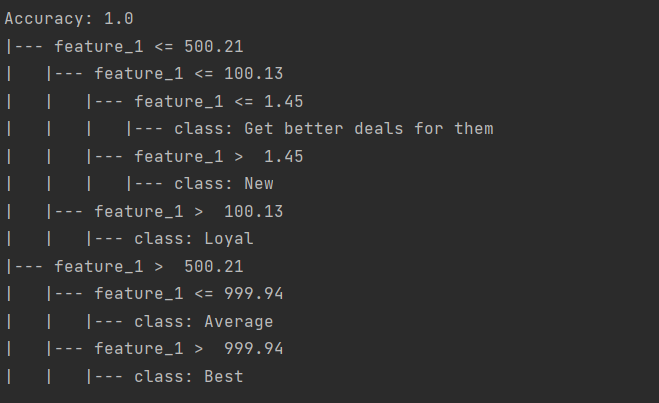
### 

### **Algorithms**

### **Decision Trees:**



Decision Tree for the model

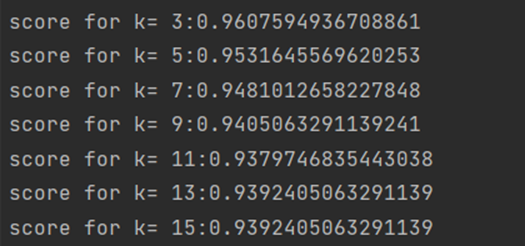


### **Logistic Regression:**



### **K-Nearest Neighbor:**

Determined accuracy for different values of K. K was chosen as an odd number to reduce biasness.



Accuracy values for different values of K

### 

Model Accuracy on testing data

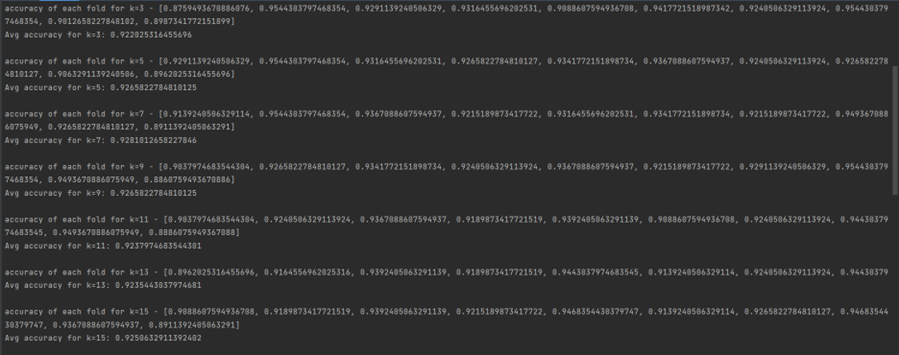
**SVM:**



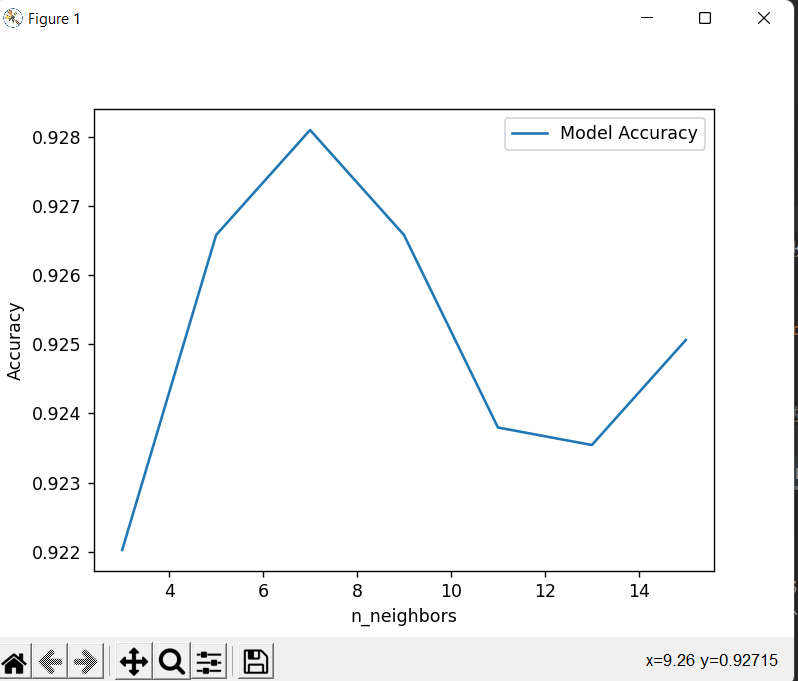
Model Accuracy on testing data

**K-Fold Validation:**

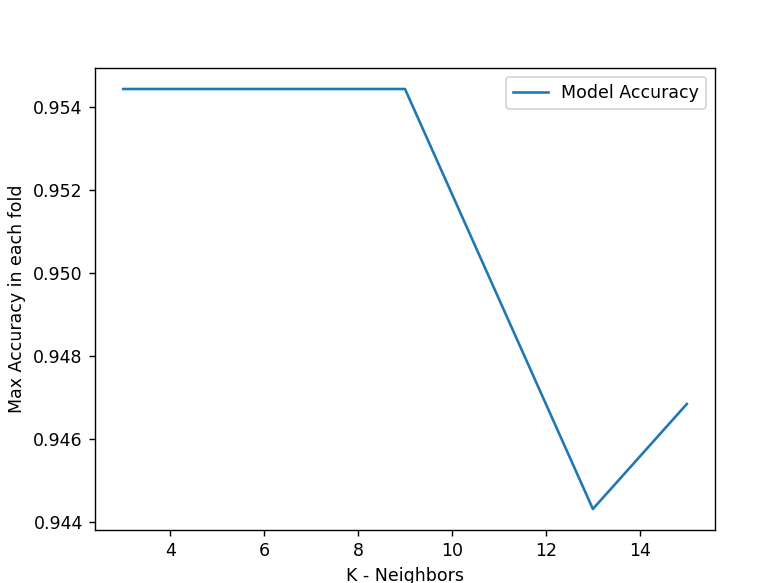
Applied **K-Fold validation** for different values of **knn** to calculate accuracy and average accuracy for each fold with a fold size of 10.



Accuracy using K-fold method for KNN



Avg accuracy for each fold



Max accuracy achieved in each fold

### **Results and Assessment**

The model yield best results on KNN, and K-Fold validation with KNN. It performed poor on SVM.