# Data Mining Project

Palestine Technical University – Kadoorie   
College of Engineering and Technology  
Department of Computer Systems Engineering

Project title:

**CUSTOMER SEGMENTATION CLASSIFICATION**

By:Ruaa Qashoo – 201911196

Instructor: Dr. Anas Melhem

E

###### 1 January 2023

INTRODUCTION

# Problem Definitaion:

An automobile company has plans to enter new markets with their existing products (P1, P2, P3, P4, and P5). After intensive market research, they’ve deduced that the behavior of the new market is similar to their existing market.

In their existing market, the sales team has classified all customers into 4 segments (A, B, C, D). Then, they performed segmented outreach and communication for a different segment of customers. This strategy has worked exceptionally well for them. They plan to use the same strategy for the new markets and have identified 2627 new potential customers.

The goal of this project is to study and predict the right group of new customers for an automotive company, so the company can adopt the specific proven marketing strategy to each of them and be more successful in the business.

I decided to use Weka to analyze the problem and then solve it.

dataset description:

There is 8068 instances (customers) for training, and 2627 instance (new potential customers) for testing.

This dataset has 11 Variables (10 attributes & 1 class/target), divided as follows:

Link of The Dataset: [Customer Segmentation Classification | Kaggle](https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation)

|  |  |  |
| --- | --- | --- |
| **DEFINITION** | **TYPE** | **VARIABLES** |
| Unique ID | numeric | ID |
| Gender of the customer | binary | GENDER |
| Marital status of the customer | binary | EVER\_MARRID |
| Age of the customer | numeric | AGE |
| Is the customer a graduate? | binary | GRADUATED |
| Profession of the customer | nominal | PROFESSION |
| Work Experience in years | numeric | WORK\_EXPERIANCE |
| Spending score of the customer | Ordinal | SPENDING |
| Number of family members for the customer (including the customer) | numeric | FAMILY SIZE |
| Anonymized Category for the customer | nominal | VAR\_1 |
| (target) Customer Segment of the customer | nominal | SEGMANT |

**Problems with this data & how I solve it** << Clean/Prepare the data >>

# Missing Values:  
 after uploading the data set:

click open file chose our training dataset this information about each attribute will appear:

* NOTE: as we can see there is 4 colors in each chart  
    
   represent data from class A  
    
   represent data from class B  
    
   represent data from class C  
    
   represent data from class D
* There is no missing value in Segmentation attribute (class)

* There is 144 missing value in Ever\_Married attribute
* There is no missing value in Gender attribute
* There is no missing value in Age attribute
* There is 78 missing value in Graduated attribute
* There is 124 missing value in Profession attribute
* There is 829 missing value in Work\_Experience attribute
* There is no missing value in Spending attribute
* There is 335 missing value in Family\_Size attribute
* There is 76 missing value in Var\_1 attribute

## **#Missing Value processing**

As above, we see that data contains some missing values. there are total of 8068 entries, but some columns have less than 8068 entries, which means they have missing values:

Columns " Family Size "and " Work Experience " have huge number of null values. Both are numeric attribute, so can be replaced with their mean values. **But** because each of these features represents an integer value (1,2, 3…………) and the mean of each respectively is (2.85, 2.64) which is not an integer.

# Also replacing missing with the mean value may affect outliers.

**So,** I decided to replace the missing values in these columns with the median value of each one:

* Median of family\_size = 3
* Median of work\_Experience = 1

**Replacing them in weka:**

open the training dataset click Edit write click on attribute which needed to edit then choose set messing value to and enter the median value (that we calculated above):

**We can see some of messing value**

After replacing them we click " **OK "** to save the changes we made

Columns " Ever\_Married " , " Graduated " , " Profession " , " Var\_1 " have null values.

All of them either nominal or binary, so can be replaced with their **Mode** values.  **Note:** in weka there is 3 replacing filter I chose " ReplaceMissingValues " because this filter allow replacing the missing with Mode value.

Now as we can see we replace all missing values (**No missing values now**)

**Outlier/ extreme values:**

Now we need to find if there are any outlier/ extreme values in our data set:

A common rule for identifying suspected outliers is to single out values falling at least 1.5×IQR above the third quartile or below the first quartile.

**# Finding them in weka:  
  
 \*Note: First you must process the missing values as above**

Filter unsupervised attribute InterquartileRange (IQR)

# After we apply this filter 2 new column added to dataset (Outlier , ExtremeValue):

In our dataset there is no Outlier/Extreme

**Note:**   
If there were outliers, we could solve them by using this filter which remove all instances that contain the outliers:

Filter 🡪unsupervised🡪 instance🡪 **RemoveWithValue.**

After completion of the process, the variables (Extreme Values, outliers) deleted.

**Note:**   
we can Remove column ID as it´s not important to the model

**Note:**   
Because we need to make clustering for data, All Numeric attributes should be transformed to a similar scale to be effective:

Filter 🡪unsupervised🡪 instance🡪 **Normalize.**

Normalizes all numeric values in the given dataset.

The resulting values are in [0,1] for the data used to compute the normalization. intervals.

We finish data preparation (fill in missing values ​​and get rid of the extreme values and outliers (if there is) ).

**Now our dataset ready for applying data mining model**

**Data Mining Modeling**

**# Classification:**

### **# Classification using Decision Tree:**

Classify 🡪 choose 🡪 tree 🡪j48

We can see that the **correctly** classified instance is: 68.6292%   
We can see that the **incorrectly** classified instance is: 31.3708%

### 

### **# Classification using Naïve Bayes:**

Classify 🡪 choose 🡪 bayes🡪**NaiveBayes**

We can see that the **correctly** classified instance is: 50.6693%   
We can see that the **incorrectly** classified instance is: 49.3307%

### **We can see that classification by using the first model (j84 Decision Tree)** **more accurate than classification by using the second model (naïve Bayes). So, we chose the j48 Decision Tree for Classification**

### **# Clustring:**

### **Simple K-Means**

**Decision**

In this project we went through all the process from defining the business objective, knowing that dataset, clean and preparing the data exploring features and distributions, data modelling and presenting different algorithms to select the best to predict the Customer Segmentation, what will help the business adopt the best marketing strategies to each of them and bring more market share and revenue to the company.

The chosen model was j84 model since it´s the most accurate, although it doesn’t present a high accuracy. We could reach a more accurate model having more data about customers, it´s something to explore and go deeper in the organization with the business team and the data engineer in order to explore if more relevant features are available.

The End