Project Name

Final project for Data Mining course

Group number -

First name	Family name
Stav	Atias
Ido	Stern
Niv	Ben Avraham

ABSTRACT

Today, many stores invest a lot of money in advertising their products, in order to increase the clientele and thus increase their profits.

During the project, we used the "Customer Personality Analysis" database. After analyzing the data and classifying the customers according to their potential for accepting the campaign, we think that we have a good recommendation for the store manager.

In the process of choosing the classification model, we examined three different alternatives, and finally, the "Random Forest" algorithm was chosen.

We have created a model that predicts 80% of whether a customer will agree to the campaign.

We believe that with more long-term data we can build better customer segments and improve the model.

1 Introduction

Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors, and concerns of different types of customers. Customers show different personalities and behavior.
Therefore, while producing quality analysis of the data, will contribute to spending less money on campaigns, and to understanding which customers will likely spend more in the store. For example, instead of creating a general campaign, the store will do it for specific personalities.
The dataset for this project is provided by Dr. Omar Romero-Hernandez.

2 Material and Methods

Data

The purpose of our project is to predict whether consumers will agree to buy in the store through the campaign that the store will advertise.
Our target column is 'Aceepted_Cmp' which shows whether a customer has agreed to campaigns in the past.
The classification model under supervised learning we built, based on the 'Personality Customer Analysis' data, includes 2240 lines and 29 columns. The data columns describe the customer in terms of year of birth, income, number of children, marital status, education, seniority, and number of purchases in the store.

Data Cleaning

Issue	Manipulate	Reason
Income	Deleting all income > 150000 (7 rows).	Deleting outliers.
Missing cells	Delete 24 rows. (24/2240)	Ignore missing cells.

Data preparation:

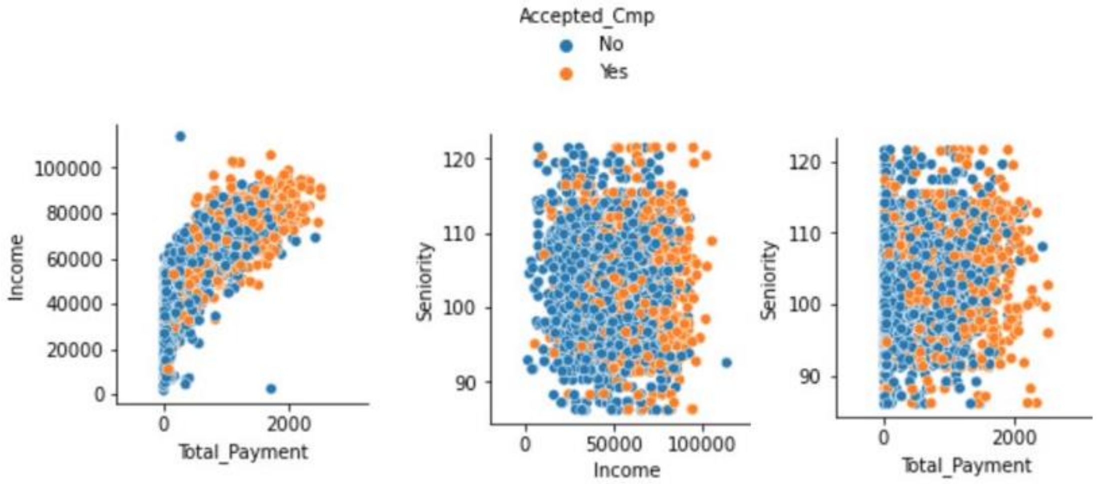
In order to better understand the data for our study, we have made certain columns categorical and combined columns into one column as follows:

a. Marital status - Is the client in a relationship or not.
b. Education - Is the client an academic or not.
c. Customer seniority - the time the customer registered for the store.

d. The amount of money the customer bought - summarizing all the products into a total payment.

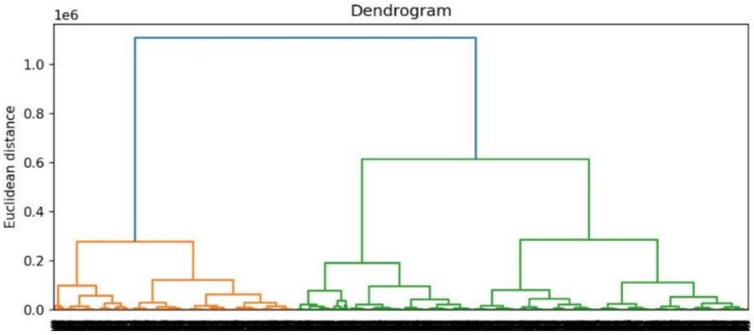
e. A number of transactions purchased - the sum of all transactions made in various services on the site.

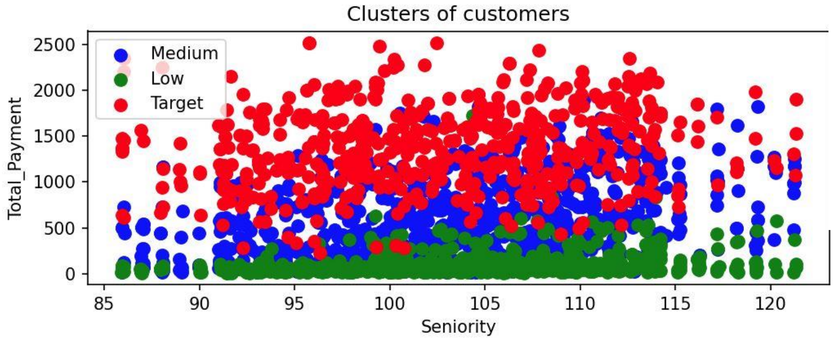
f. ״Accepted_Cmp״ - Whether a customer Accepted a campaign or not.

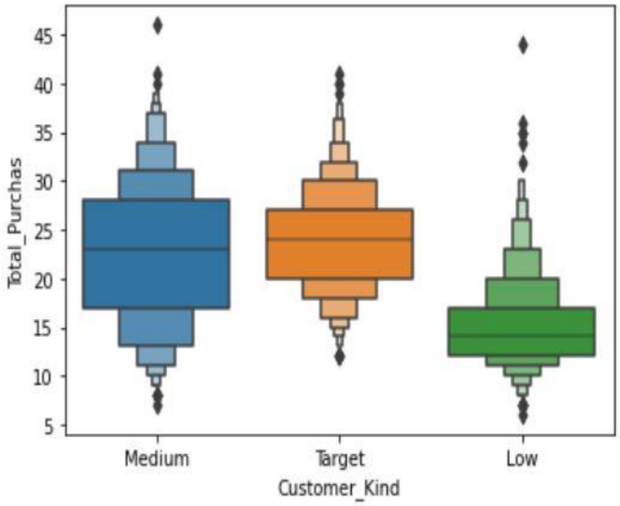


In order to find the potential customers, we used the method of hierarchical clustering. The clustering made on the following columns:

income, seniority, and the total amount of money bought by the customer. We found that the optimal separation is to three kinds of customers. We converted the columns to one that calls ‘Customer’Kind’.







Preparations for algorithms

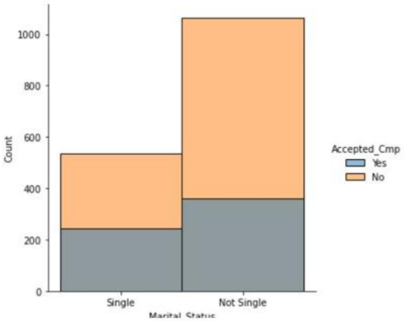
1. We converted all the columns into a categorical data type.
2. For finding the columns that match our target column ("Accepted_Cmp"), we performed correlation tests (Chi-Square) and found that the most suitable columns for predicting are:
   a. Customer Kind: P-value = 0.003.
   b. Marital Status: P-value = 0.01.
   c. Kids : P-value = 1.35×10^{-9}
3. The data split for 70% training and 30% test.

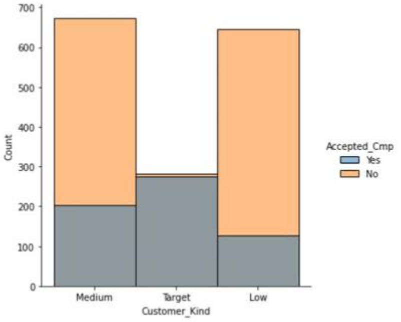
3 Numerical Analysis and Results

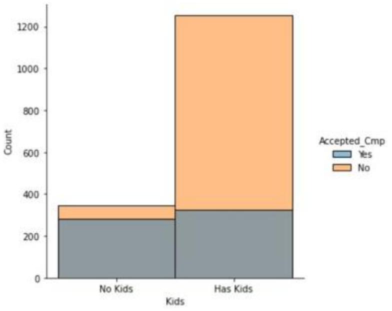
Analysis of the results

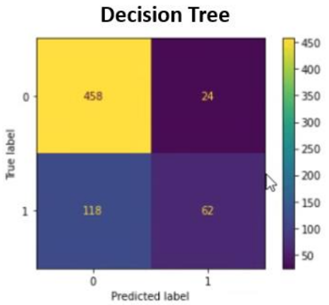
While predicting the results, we performed three algorithms:

1. Decision Tree -The columns were inserted into a 'one hot' matrix, and we chose that the depth of the tree would be 3. The "Marital Status" column has the highest 'Gini' grade.
2. Naïve Bayes – The coulmns data type that inserted were categorial.
3. Random Forest.

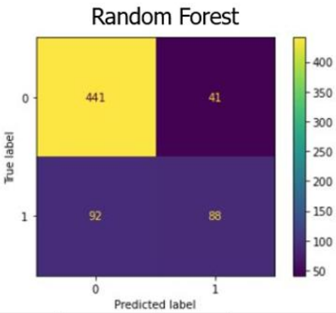




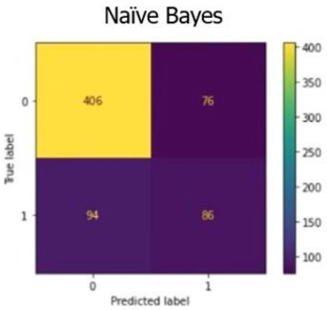




Accuracy	Precision	Recall	F_1 *Score*
0.8	0.72	0.34	**0.46**



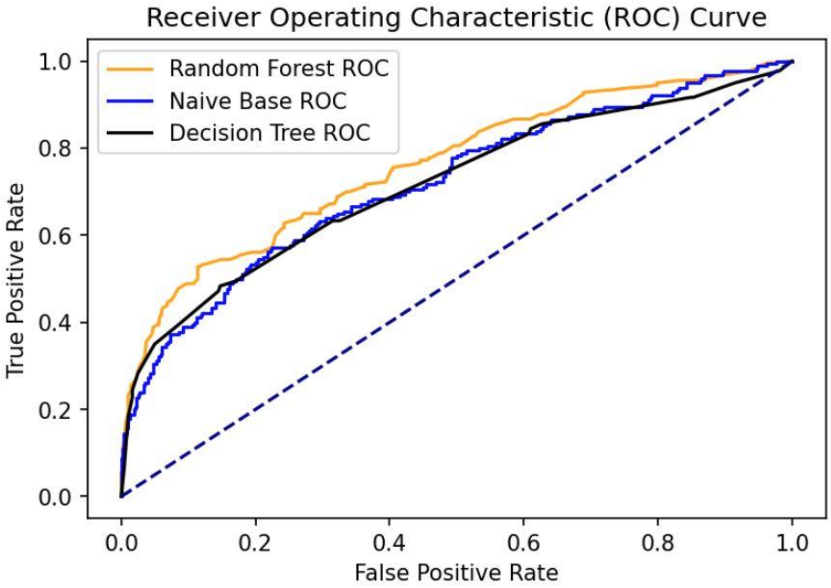
Accuracy	Precision	Recall	F_1 *Score*
0.8	0.68	0.49	**0.57**



Accuracy	Precision	Recall	F_1 *Score*
0.75	0.53	0.48	**0.50**

Discussion

The algorithm that has the best performance is Random Forest.



4 Conclusions

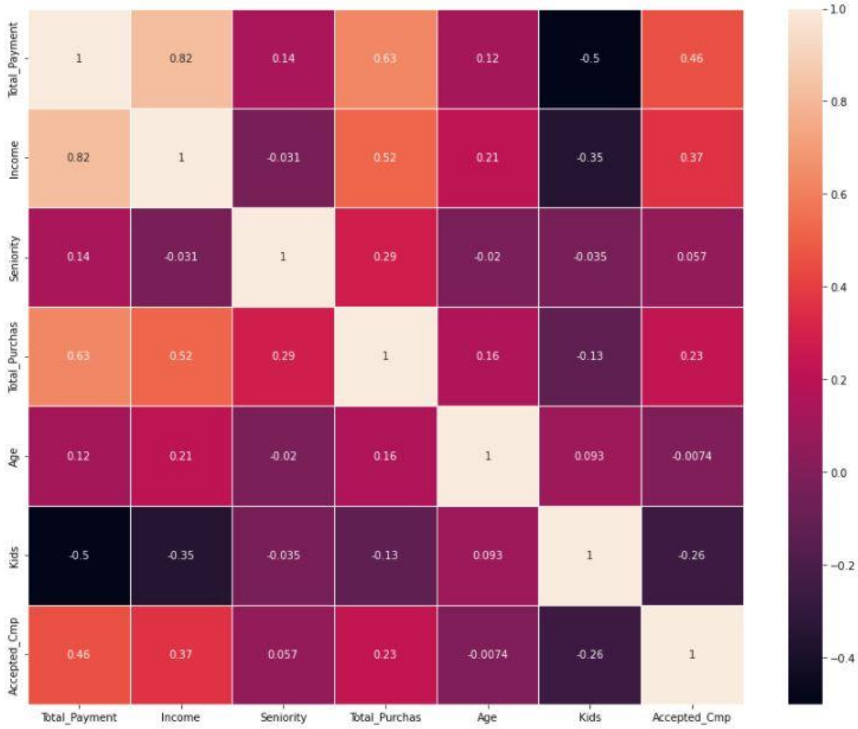
In this project, our statement was helping the store to adjust the advertising campaigns for customers kinds, in order to get the optimal result from it.
For example, if the store wants to set a new meat product campaign, the model will analyze the customers that have a high percentage of buying the new product.
Our model gives the store manager a prediction according to 3 characteristics:
1. The customer's potential to buy - is characterized by his income, seniority, and his total payment. If the customer will classify as a 'target' customer, probably will accept the campaign.
2. The marital status of the customer.
3. Customer without kids has bigger percentage to accept the campaign.
Our recommendation for further work is to collect long-term data regarding the customer characteristics and his treatments.
Customer segmentation is the key for the prediction of whether the client will accept the campaign.

5 References

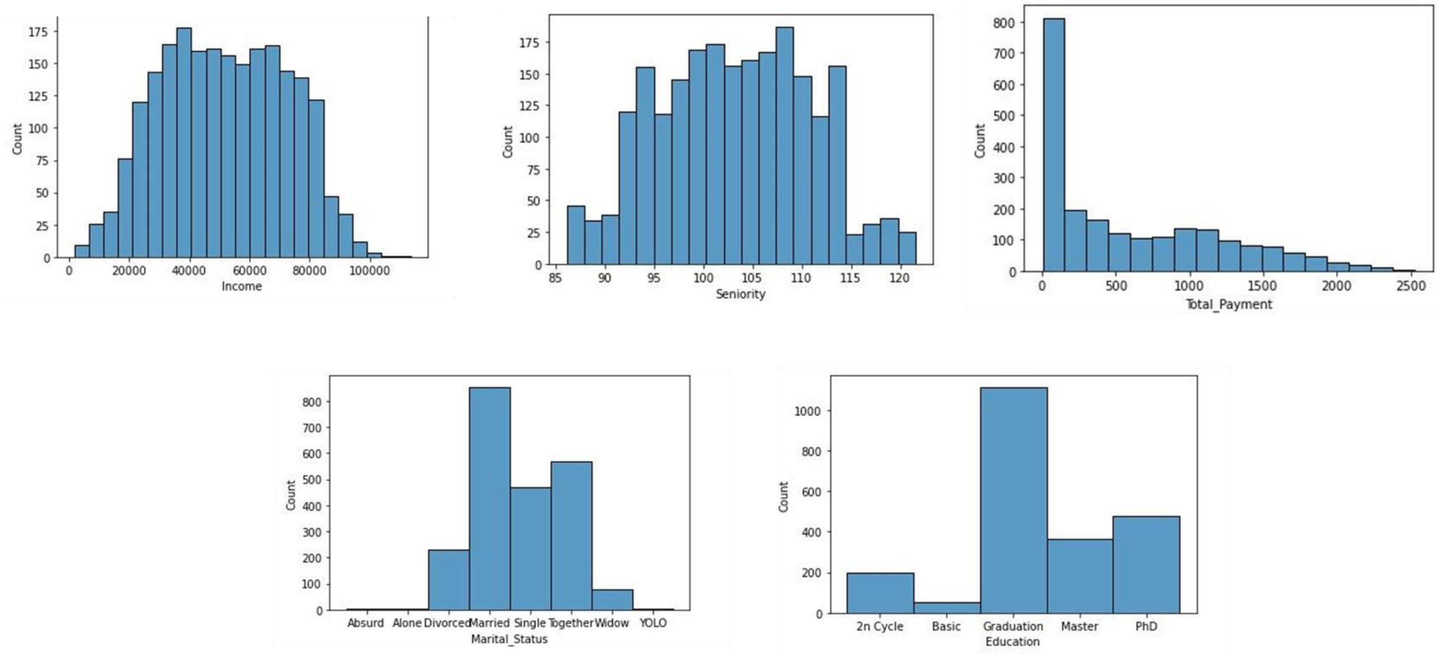
[1] https://www.kaggle.com/imakash3011/customer-personality-analysis

6 Appendix

Correlation between numeric columns:



Histograms :


Count
Income
Count
Seniority
Count
Total_Payment
Count
Absurd
Alone
Divorced
Married
Single
Together
Widow
YOLO
Marital_Status
Count
2n Cycle
Basic
Graduation
Master
PhD
Education