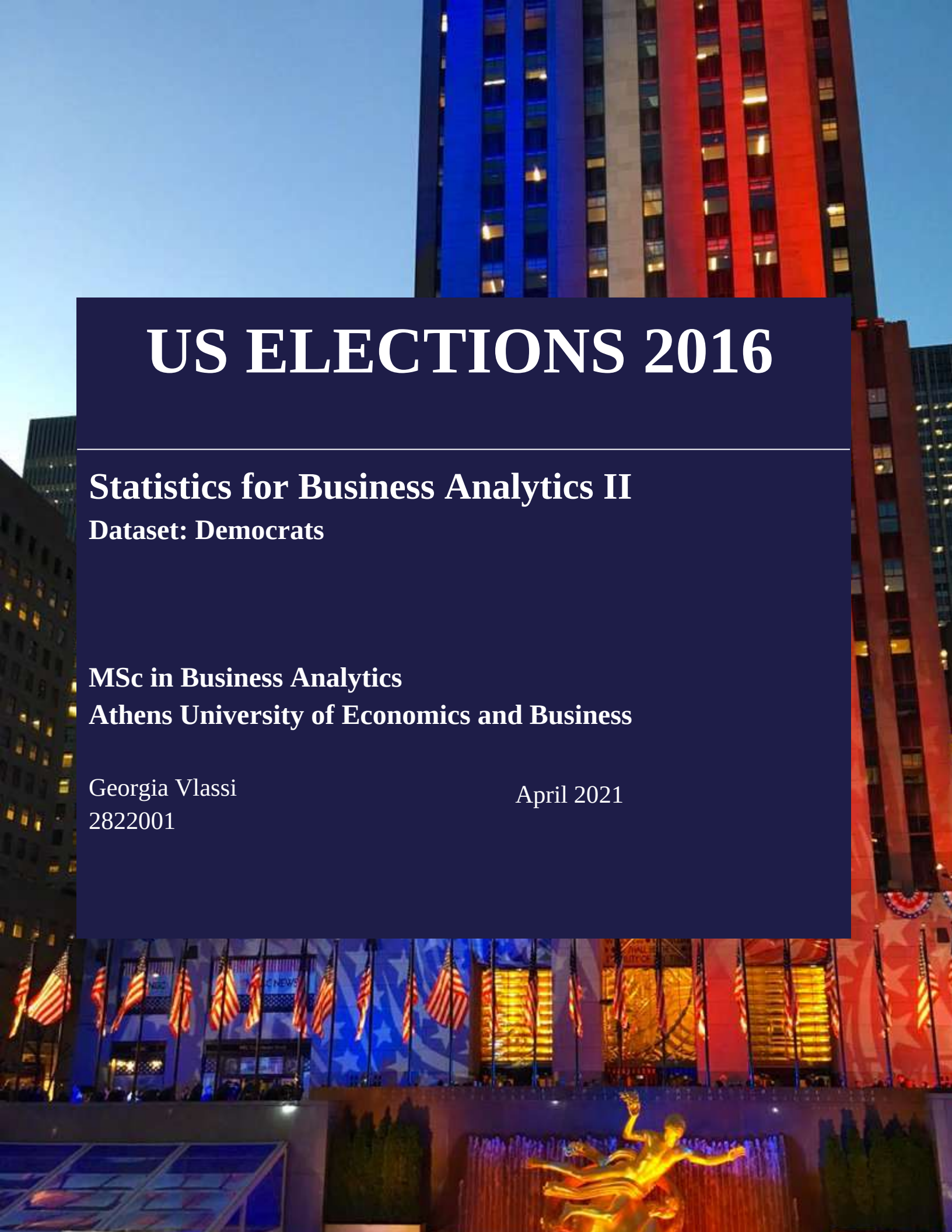

US ELECTIONS 2016

Statistics for Business Analytics II

Dataset: Democrats

MSc in Business Analytics
Athens University of Economics and Business

Georgia Vlassi
2822001

April 2021

Table of Contents

Abstract

The data of this report refer to votes of the Democratic party, by US County and the presidential candidates, Hillary Clinton and Bernie Sanders. Having already analyzed the demographic and economic data of the voters, we will move on predictions voting Hillary Clinton or Bernie Sanders. The main scope of this report is to create a predictive model to classify whether Clinton or Sanders will win the county. In order to achieve the classification, distinct methods, like the Logistic Regression, the Naïve Bayes and the Decision Tree, will be used. The predictions of each classifier will be further assessed. After the classification, the clustering of the economic related variables will be implemented to describe the data.

1. Introduction – Data Analysis

The dataset which will be used for our analysis contains information regarding the demographic and economic data of the voter. Some indicative data are the origin and the race of the voters, their wealth, their work or their language. Our analysis is limited only in the democratic voters, and more specifically to Hillary Clinton's and Bernie Sander's voters.

Before starting to analyze the data, a data handling and transformation is needed. A new response variable is created, which indicates whether Hillary Clinton won over a county or Bernie Sanders. The new variable named **Winner** is defined as:

$$\text{Winner} = \begin{cases} 1: \text{vote over Hillary Clinton} \\ 0: \text{vote over Bernie Sanders} \end{cases}$$

The data that are irrelevant with the prediction are omitted. More specifically, columns regarding the counties names and abbreviations are not taking into consideration. Moreover, our analysis will be implemented in numeric variables only, so the null values at these columns will be deleted.

In order to better understand the relationship among the most important socioeconomic variables and the response variable Winner, the correlation plot is designed to give more emphasis to the relationships among attributes. If the ellipse leans towards the right, it is positive correlation and if it leans to the left, it is negative correlation. As shown to the correlation plot below(Figure 1), the highly correlated variables to the Winner, are those indicating races, like Black or African, Asian and Native Hawaiian.

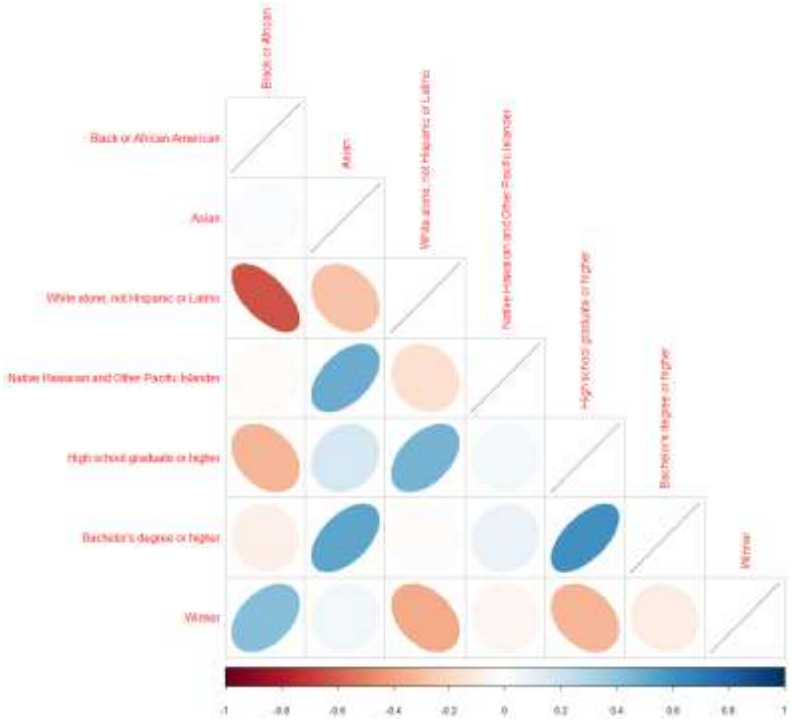

Figure 1 Correlation Plot

2. Classification

For classification three distinct methods are used and tested to classify the data between the two groups. As we have a dataset with 52 variables, for each classifier different selection of variables will be used.

2.1 Data Preparation

As our scope is to test different classifiers, we should prepare our dataset and split it into train and test. In order to achieve that k-fold cross-validation is used, where randomly split the data set into k-subsets. After some tries, we conclude to use k = 6 folds, where we will have 2.308 rows for testing and 461 rows for prediction. This separation of the dataset will be used for all the below methods.

2.2 Logistic Regression Model

The first classifier that will be used is the Logistic Regression, which is a powerful statistical way of modeling a binomial outcome with more explanatory variables. For the purposes of this analysis, two methods were tested for variable selection, the AIC and the BIC and we conclude to continue with AIC, as it is indicated for estimating predictions. We keep only the statistically significant variables after implementing the AIC, as shown with their estimates below:

Intercept	-72.81

Population 2014	-0.00007143
Population (%)	0.1405
Population 2010	0.00006284
Persons under 5 years(%)	-0.3522
Persons under 18 years (%)	0.2385
Persons 65 years and over (%)	0.2366
White(%)	0.7757
Black or African American(%)	0.9939
American Indian and Alaska Native(%)	0.7591
Asian(%)	0.9520
Native Hawaiian and Other Pacific Islander(%)	1.174
White not Hispanic or Latino(%)	-0.03343
High school graduate or higher(%)	-0.09050
Bachelor's degree or higher(%)	-0.06976
Veterans	0.0001032
Homeownership rate	-0.05136
Housing units in multi-unit structures(%)	-0.02929
Median value of owner-occupied housing units	-0.000006642
Per capita money income	0.0001948
Private nonfarm establishments	-0.001271
Nonemployer establishments	0.0003884
Merchant wholesaler sales	0.0000001121
Retail sales	0.0000005839
Population per square mile	0.0006062

Table 1 Coefficients of Logistic Regression

From the Table 1, we can understand that the coefficients related to races have greater impact on voting over Hillary Clinton. Most specifically the Native Hawaiian, the Asian and the Black or African population.

At Table 2 we can see the contingencies of the test dataset. Implementing the Logistic Regression, we conclude that its predictions have accuracy equal to 80.04%.

	0	**1**
0	144	55
1	37	225

Table 2 Logistic Regression contingency table

2.3 Decision Tree Model

The second classifier that will be used is the Decision Tree. A decision tree classifier can be considered as a flowchart diagram, which uses a structure of branching decisions, with the terminal nodes representing classification outputs/decisions. Starting with the initial dataset, we measure the entropy to find a way to split the set until all the data belongs to the same class. The scope of this method is to fit the data into a tree as shown below.

Analyzing the below tree, we observe that the root of the tree is the Black or African American population. Similar to Logistic Regression this coefficient seems to be very important factor to decide the winner between Hillary Clinton and Bernie Sanders. In the first leaf, we can see that the population having graduated high school has also great impact for the decisions, followed by Hispanic or Latino population.

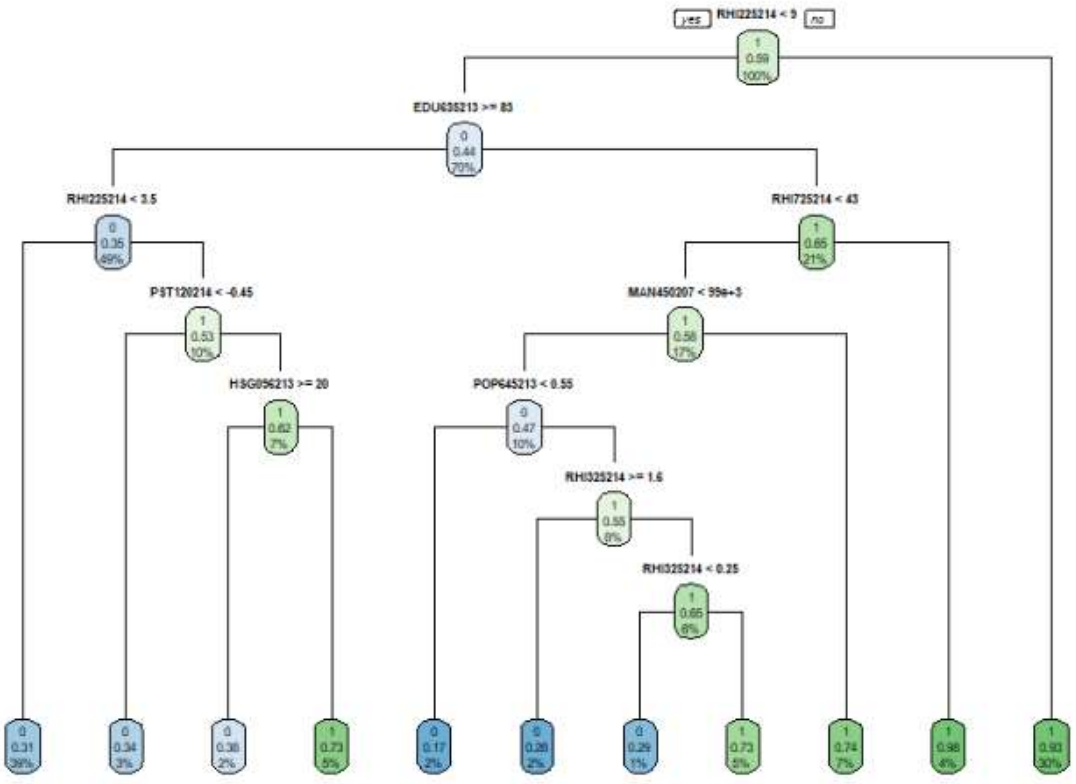

Figure 2 Decision Tree

	0	1
0	147	79
1	34	201

Table 3 Decision Tree contingency table

At Table 3 we can see the contingencies of the test dataset. Implementing the Decision Tree, we conclude that its predictions have accuracy equal to 75.49%.

2.4 Naïve Bayes Model

The third technique for classification is a supervised learning technique based on Bayes' Theorem with an assumption of independence among predictors. For this method we do select specific variables, but we have to use the initial model including all the numeric variables. By using this classifier, we can assume that the presence of a particular feature in a class is unrelated to the presence of any other feature. This assumption could lead us to a not a very accurate model.

	0	1
0	167	139
1	14	141

Table 4 Naive Bayes contingency table

At Table 4 we can see the contingencies of the test dataset. Implementing the Naïve Bayes, we conclude that its predictions have accuracy equal to 66.81%.

2.5 Model Comparison

Below we can see the comparison between the three classifiers.

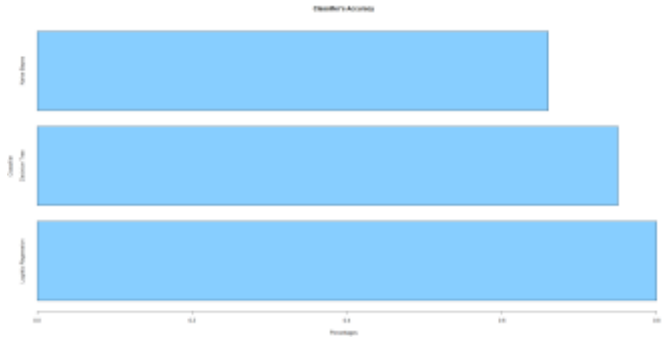

Figure 3 Classifier's accuracy

Comparing the results, we can conclude that the most accurate is the Logistic Regression, having the highest accuracy. The Decision tree is close to Logistic Regression having almost the same coefficients for the decisions. Finally, there is the Naïve Bayes method, having a poor performance regarding the others.

3. Clustering

3.1 Hierarchical Clustering

The hierarchical clustering is an approach, which builds a hierarchy from the bottom-up, and does not require to specify the number of clusters beforehand. Regarding the algorithm, it puts each data point in its own cluster. The closest two clusters are identified and combined into one cluster. In hierarchical clustering distances are used to find data closely related with each other.

In order to use the Silhouette method, we scaled our data and take the Euclidean distance between the observations. At clustering we use only the economic variables. As show in Figure 4, it seems that two clusters are better for clustering.

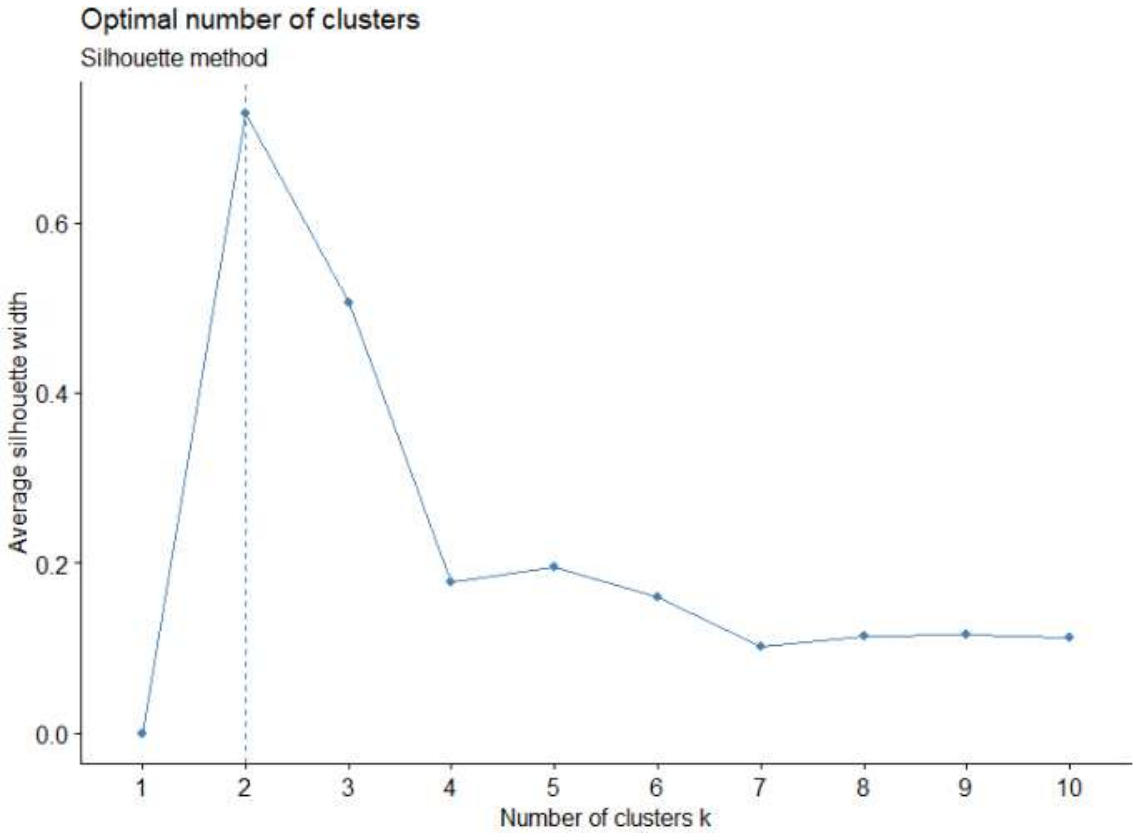

Figure 4 Hierarchical Clustering Silhouette method

In Figure 5, we observe that our data are clustered in two clusters by using the "WARD" method, which tries to minimize the total within sum of squares (WSS) of the clustering.

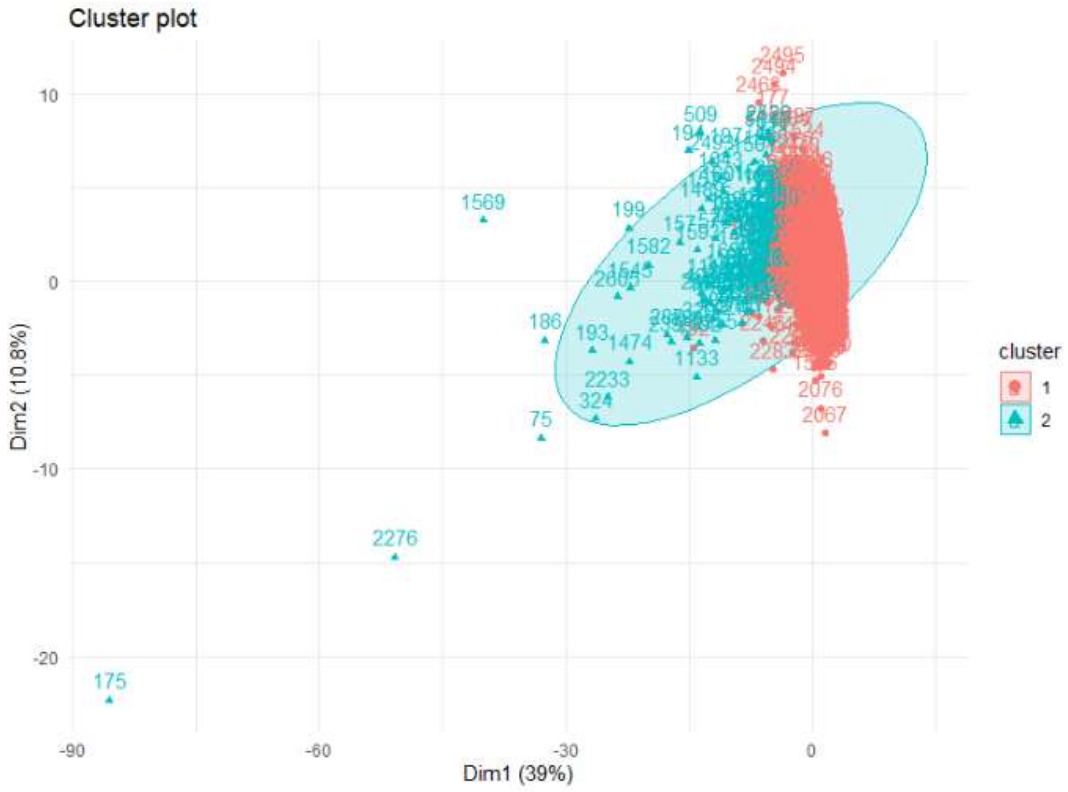

Figure 5 Hierarchical Clustering

3.2 K-Means

Another algorithm for clustering is the k – means. We use kmeansruns() function, in which we give a range of Ks, like 10, and it returns the best K according to two different criteria. These criteria are the Calinski Harabasz Index (ch) that measures how separately the clusters are from each other's, and the Average Silhouette Width (asw) that measures how tightly grouped all the points in the cluster are. As shown in Figure 6, in both methods the K equal to 2, is the better number of clusters. In Figure 7, we observe that there are similarities in the plot with the one occurred from hierarchical clustering, especially in the percentages that the two clusters are separated, 10.8% and 39% respectively, in the two methods.

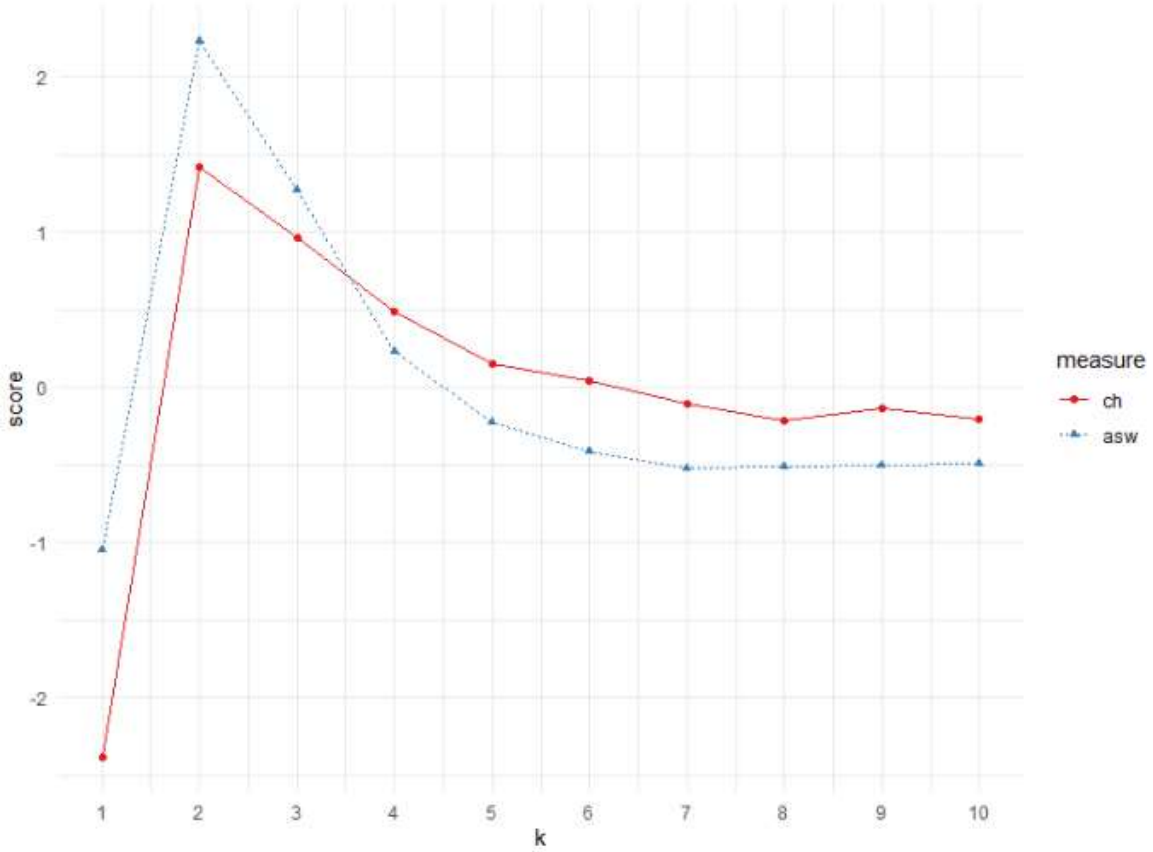

Figure 6 K - Means Clustering with ch and aws methods

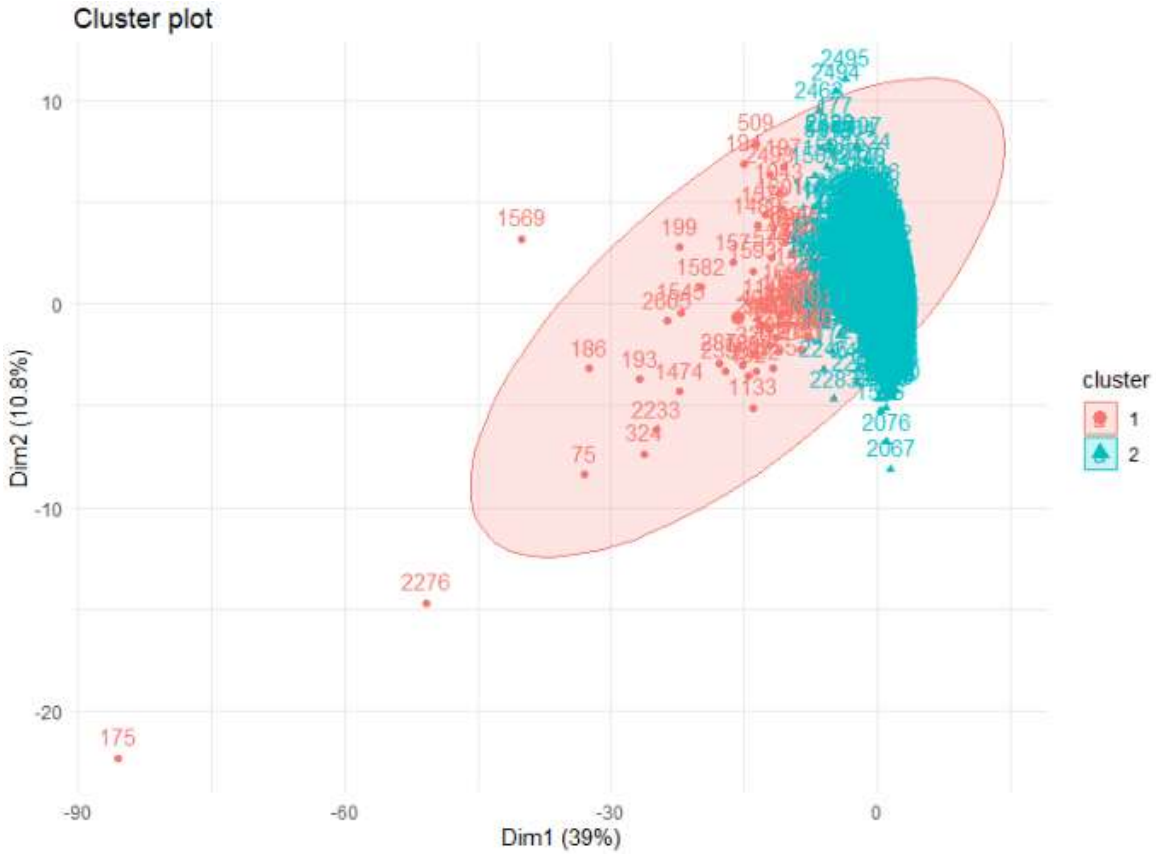

Figure 7 K - Means Clustering

4. Conclusions - Discussion

Completing our analysis, such conclusions can be made. Regarding the dataset of the US Elections of 2016, we analyzed the attributes of the voters over the Democratic party. Three distinct classifiers were implemented, showing that the Logistic Regression method achieved to predict with 80% accuracy the voters over the candidate Hillary Clinton. A very good classifier is also the Decision Tree, having 75% accuracy at prediction. The attributes that highly affect the prediction are the human races, with the Native Hawaiians, the Blacks and the Asians, having a great impact on prediction of Hillary Clinton's win. Finally, we tried to cluster the economic attributes and we concluded that two clusters of both implementing hierarchical clustering and k-means are the best solution regarding our dataset. A further analysis of the demographic attributes of the clusters can be implemented, in order to have an overview of the general attribute of the voter.