ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

BUSINESS
ANALYTICS
Master of Science

Athens University of Economics and Business

School of Business

Department of Management Science & Technology

Master of Science in Business Analytics

Program:	Full-time
Quarter:	2nd (Winter Quarter)
Course:	Statistics for Business Analytics II
Assignment №:	Project II
Students (Registration №):	Souflas Eleftherios-Efthymios (f2822217)

Table of Contents

Statistics for Business Analytics II

Project II

Student: Souflas Eleftherios-Efthymios

Introduction

Have you ever wondered what it takes to become the next President of the United States? Of course, you will answer a lot of money for the voting campaign and voters that believe in your abilities and your vision for making their lives better. However, apart from the wealth, the abilities of a politician and his skill to persuade people, there exist some key social and economic characteristics of the voting population that play a vital role in their decision of whom to elect. If the above information is known by the candidate president, he could concentrate his effort to persuade voters having the above characteristics.

This project's purpose is two-folded. Firstly, we will try to create some predictive models that, given some socioeconomic characteristics of a county, will classify the county into a "Trump County", in the sense that Trump is popular in that county for being elected as USA President amongst other candidates and will get more than 50% of the votes of that county, or not. Our model will be trained on data of the Presidential Elections of 2016, and more specifically in the USA Presidential Primaries, where the output was that Trump was elected as the presumptive Republican nominee for USA's Candidate 45th President. Secondly, we will try to create some clusters of the counties based on their demographic characteristics and then use the economic ones for the description of the forementioned clusters.

For this project, we used R programming language, which is heavily used for statistical computing and graphics and supported by the R Core Team and the R Foundation for Statistical Computing.

Dataset Cleaning

Firstly, let's explore our data and fix or remove possible incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within our dataset, or, simpler, let's "clean" our data. The dataset provided is not small. It is an Excel (.xlsx) file with three sheets. The first sheet, which is named as 'county facts', includes the socio-economic characteristics of each county of the United States, containing its FIPS Code, name of the county, the state it belongs to and 51 metrics of different socio-economic characteristics, which each of them is explained in the third ('dictionary') sheet. The second ('votes') sheet includes the counties with its respective

FIPS code and the state it belongs to along with the total votes and the fraction of the votes each candidate president got in each county grouped by the party (Democrats, Republicans) he/she belongs to.

Apart from the big “p” (columns) of the dataset, a big “n” (rows) also exists. In the ‘votes’ sheet, 24612 rows exist. From the above, only 3586 rows are needed, which are the outcome of the Trump votes aggregation per county. In the ‘county facts’ sheet there exist 3195 rows. From the above, if we exclude the rows containing aggregated data per state and country (USA), we will end up with 3143 rows depicting different counties.

We can notice the difference between the number of counties each sheet holds (3586 votes rows vs 3143 county rows). Apart from the forementioned difference, we discovered that some counties in the ‘county facts’ sheet have different FIPS code and/or name from the counties in the ‘votes’ sheet. Because we want to explain the dependency of the fraction of votes Trump got in the primaries to the socio-economic characteristics of a county, we must merge the two datasets. Firstly, we merged the two datasets on their common FIPS code, but not all counties had the same FIPS code on both datasets. Then, we merged the already created merged dataset with the ‘votes’ dataset by their common county name. However, not all counties were named the same in the two datasets. Below, are mentioned the challenges faced during the proper join of the two datasets and the actions taken.

Alaska House Districts (40 districts in total) don't match neither by FIPS Code nor by County Name in the 2 datasets, because a Voting House District incorporate fractions from different counties (29 in total in the ‘county facts’ sheet). Also, Congressional Districts (1,2,3, and 4) of Kansas, and Districts (1-47) of North Dakota from vote data, aggregate a plethora of Kansas (KS) and North Dakota (ND) counties respectively. For the above three States we did not de-aggregate vote data to each county as this would be approximate and not scientifically correct, because it would not depict the reality. Colorado State and the State of Maine (ME) have vote data only for the Democrats and not for the Republicans, thus not having data for Donald Trump. Minnesota (MN) State, District of Columbia (Washington, DC), and Kalawao County of Hawaii are not included in the vote data at all. The forementioned States and counties were not included in the dataset that the model explains the behaviour of its voters.

By searching the internet, we ended up to the following useful information regarding false insertions of counties’ name which we updated in order to merge properly the two datasets. The most common mistake was to name the county after its biggest and most famous city. We updated the following rows in the ‘vote’ sheet:

1. Middletown to Middlesex, CT.

2. Cook County to Cook Suburbs, IL.
3. Pittsfield to Berkshire, MA.
4. Taunton to Bristol, MA.
5. Edgartown to Dukes, MA.
6. Amherst to Hampshire, MA.
7. Cambridge to Middlesex, MA.
8. Boston to Suffolk, MA.
9. Warwick to Kent, RI.
10. South Kingstown to Washington, RI.
11. St. Johnsbury to Caledonia, VT.
12. Morristown to Lamoille, VT.
13. Derby to Orleans, VT.
14. Bedford to Bedford City, VA.

We then merged the two datasets. Because we want to create a model using as response whether Trump got more than 50% of the votes at each county for the Republicans and as explanatory variables the socio-economic characteristics of the counties, we decided to create two classes. Thus, we transformed Trump's fraction of votes to two groups. The first group (named '0') includes all counties that Trump got equal to or less than 50% of the votes for the Republicans and the second group (named '1') contain the rest i.e., the counties that Trump got more than 50% of the votes. We then dropped all other variables (columns), except of the columns depicting the state that the county belongs to, the socioeconomic characteristics of each county and the category ('0' or '1') it belongs depending on the votes Trump got on the Primary Elections.

Also, for Wyoming (WY) State, there exist 23 counties in the 'county facts' sheet, whereas in the 'votes' sheet, there exist 11 combinations of two counties and Laramie County separately (12 rows in total), all having different FIPS Code and County name in order to make the join with the 'county facts' sheet. However, the forementioned State seems to be a Democrat state as in Albany-Natrona, Campbell-Johnson, Converse-Niobrara, Crook-Weston, Goshen-Platte, Hot Springs-Washakie, Sheridan-Big Horn, Uinta-Lincoln, and Laramie Counties Trump got 0 votes, while on Sweetwater-Carbon and Fremont-Park Trump got less than 50% of the votes for the Republican Party. Only on Teton-Sublette, Trump got more than 50% of the votes for the Republicans. Thus, after the merge of the two datasets and the drop of the unnecessary columns, we updated the group to all belonging to the '0' group, except the Teton and Sublette counties, which belong to the '1' group.

We, then, dropped all rows that were not placed into a group i.e., those mentioned earlier, which were in total 356. Thus, from the 3143-rows dataset, we ended up with a 2787-rows dataset. Finally, we added a column with the division that each state belongs to, because some algorithms cannot handle categorical variables with many levels, like the levels of a USA State that a county can belong to, whereas it can work with fewer-levels categorical variable, like the Division it belongs to, which can take 9 possible values:
New England (NE); Middle Atlantic (MA); East North Central (ENC); West North Central (WNC); South Atlantic (SA); East South Central (ESC); West South Central (WSC); Mountain (M); Pacific (P).
We, then, searched for any anomaly in the captured data of the dataset and found that no missing values and no weird data (e.g., negative income or population, percent of population negative or above 100 etc.) existed.

Classification (Predictive) Models

In order to create a predictive model to classify whether Trump "will" get more than 50% of the votes, we firstly checked if our response variable were grouping the observations in two unequal groups. This was not the case, as our dataset was grouped in the two groups, '0' and '1', in a 64-36 percent manner respectively. Then, we created a bootstrap sample (sample with replacement) from the full dataset as the train set and the observations not included in the train set as test dataset. We, then, decided to use 3 distinct methods, assess how good are the predictions made by those models and select the best out of them. The classification methods used are: Decision Tree, Random Forest and Support Vector Machine.
The classification tree's aim is to classify each observation to a given number of categories. It belongs to the group of Hard Classification, as no probability model is used to handle uncertainty (overlapping). The Decision Trees are powerful and popular tools for classification and prediction, as they generate easy to understand rules, are non-parametric and flexible, but can become over-complex and have a tendency to overfit the data. In most real applications, we prune the tree by cutting nodes that improve only slightly the performance to end up with a more parsimonious tree. Due to the restriction of the fact that factor predictors used in R's 'tree' library must have at most 32 levels, we excluded the variable holding the USA State that the county belongs to from the training model. We then ran a 20-fold cross-validation experiment to find the best size i.e., number of terminal nodes, that minimizes the misclassification error rate of the tree. This, was found to be 5 (size of tree) and will be used for the pruning of the tree. Below, we can observe the total number of misclassifications and

the respective misclassification error rate per size of the tree as a result of the cross-validation ran. We can also observe the firstly fitted unpruned classification tree and the respective one after the pruning of the former with the use of the hyper-parameters implemented by the cross-validation.

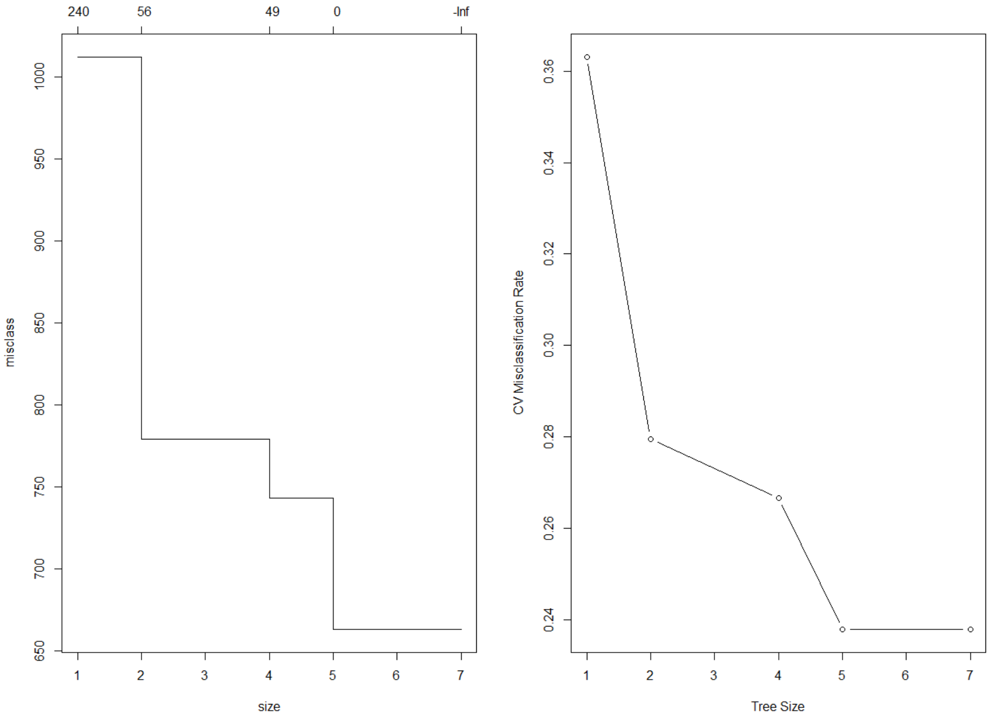

Figure 1 - Total number of misclassifications and respective misclassification error rate per size of the tree as a result of the cross-validation

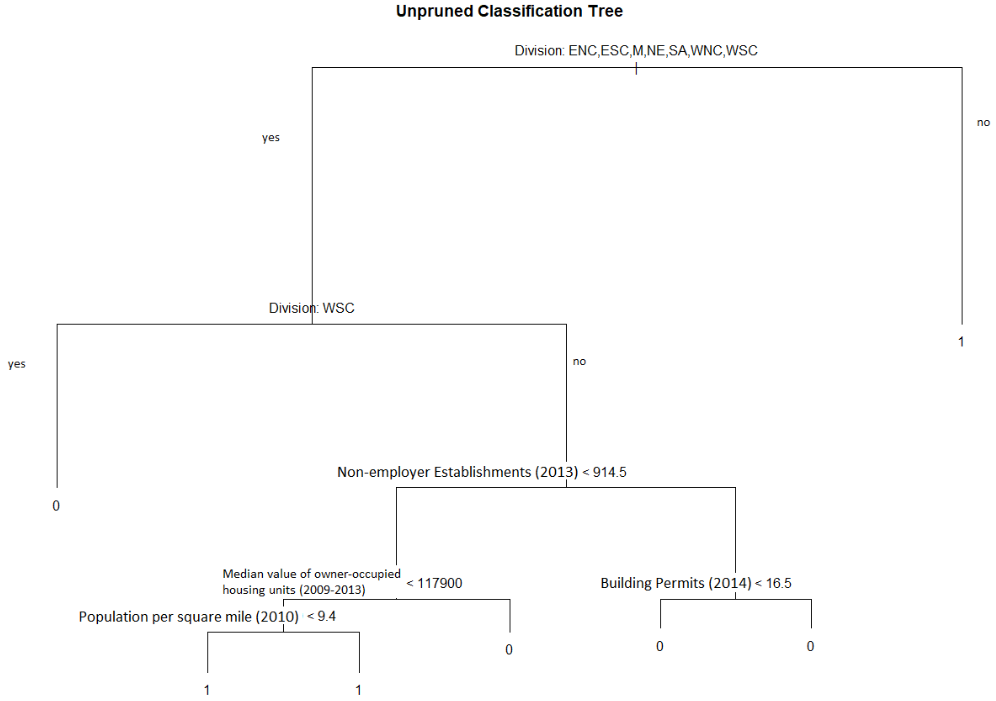
Unpruned Classification Tree
Division: ENC,ESC,M,NE,SA,WNC,WSC
yes
no
1
Division: WSC
yes
no
0
Non-employer Establishments (2013) < 914.5
Median value of owner-occupied housing units (2009-2013) < 117900
Building Permits (2014) < 16.5
Population per square mile (2010) < 9.4
0
0
0
1
1

Figure 2 - Firstly fitted unpruned classification tree

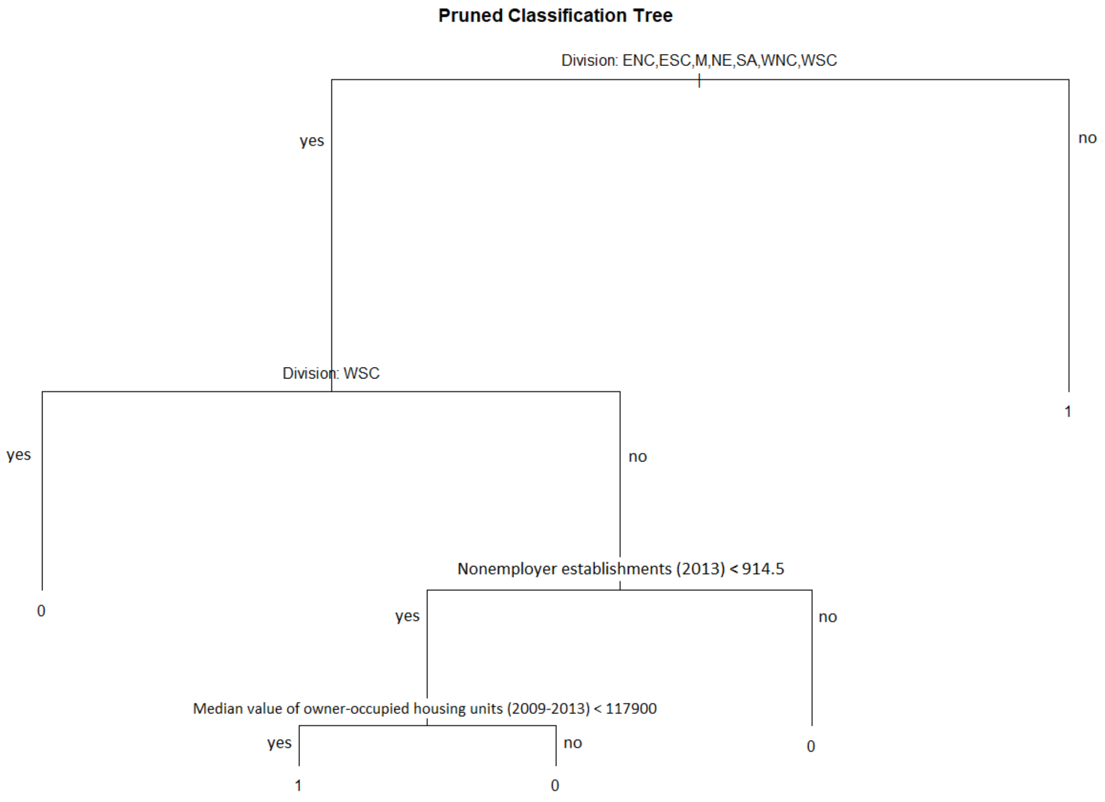
Pruned Classification Tree
Division: ENC,ESC,M,NE,SA,WNC,WSC
yes
no
Division: WSC
1
yes
no
0
Nonemployer establishments (2013) < 914.5
yes
no
Median value of owner-occupied housing units (2009-2013) < 117900
yes
no
0
1
0

Figure 3 - Classification Tree after pruning the firstly fitted classification tree with the use of the hyper-parameters implemented by the cross-validation

The forementioned pruned tree is interpreted as following:

If the county belongs to the Middle Atlantic (MA) or Pacific (P) division then classify it to the group '1'. If the county belongs to the West South Central (WSC) division then classify it to the group '0'. If the county belongs to the East North Central (ENC), East South Central (ESC), Mountain (M), New England (NE), South Atlantic (SA), or West North Central (WNC), then if the non-employer establishments recorded in it in 2013 are 915 or more, then classify the county to group '0'. Else, if the median value of owner-occupied housing units recorded during the 2009-2013 period is less than $117,900, classify the county to group '1', otherwise to group '0'. The forementioned variables (Division, non-employer establishments and median value of owner-occupied housing units) are the most important variables.

We, then, used Random Forest. Random Forests construct an ensemble predictor by averaging over a collection of binary trees, thus being also a method belonging to the group of hard classification. So, instead of building one decision tree, several of them are being built to create a forest, whose predictions are taken and combined creates a generic prediction. We trained a Random Forest of 200 individual trees, with 2787 (as the size of the dataset) observations taken, using sampling with replacement from the full dataset (bootstrap). As another hyper-parameter of the algorithm, in each tree's node, 10 variables were randomly sampled as candidates, in order to make the decisions, whose importance will be assessed. No pruning was applied, resulting the trees to grow with different structure, aiming at capturing different effects in the data. As the forementioned restriction of the levels of categorical variables did not apply here, we included the variable holding the USA State that the county belongs to. The measured Variable Importance by the Random Forest trained for the top 10 used (important) variables is:

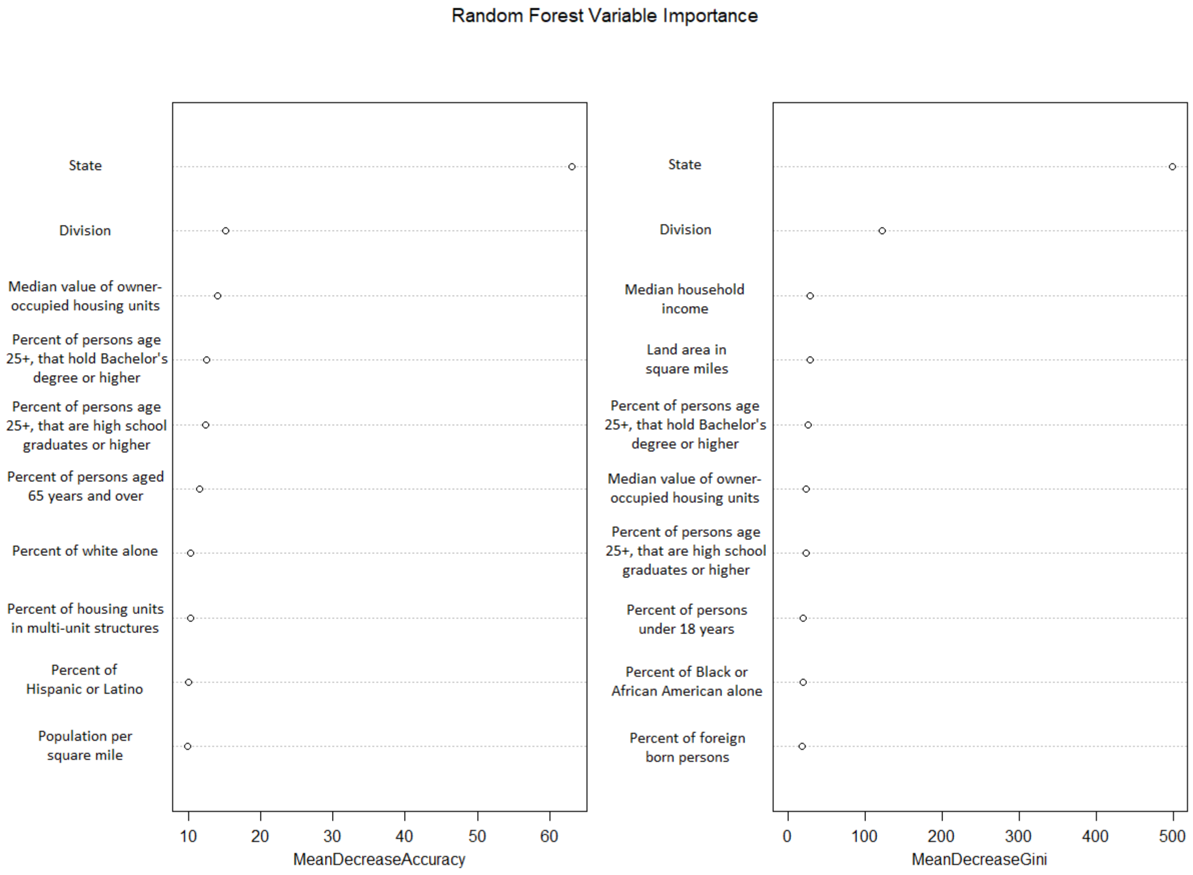

Figure 4 - Random Forest Variable Importance

The higher the value of Mean Decrease Accuracy or Mean Decrease Gini scores, the higher is the importance of the variable in the model. The Mean Decrease Accuracy measures how much the model accuracy decreases if we drop that variable, whereas the Mean Decrease Gini measures the variable importance based on the Gini index used for the calculation of splits in trees. The plot above (Figure 4) indicates that firstly the State and then the Division that the County belongs to are the most helpful variables in the model.

Finally, we used Support Vector Machine (SVM). SVMs belong to the group of hard classification methods. They use learning algorithms that analyze data and recognize patterns, simultaneously minimizing the empirical classification error and maximizing the geometric margin. The simple idea used by SVMs is that they can separate the two classes by a hyperplane. We trained a support vector machine to carry out the classification, using a linear kernel, after we scaled the variables to have zero mean and unit variance. Then, we plotted the Receiver Operating Characteristic (ROC) curves of the three models. ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. We can use it to compare the three different classification methods trained.

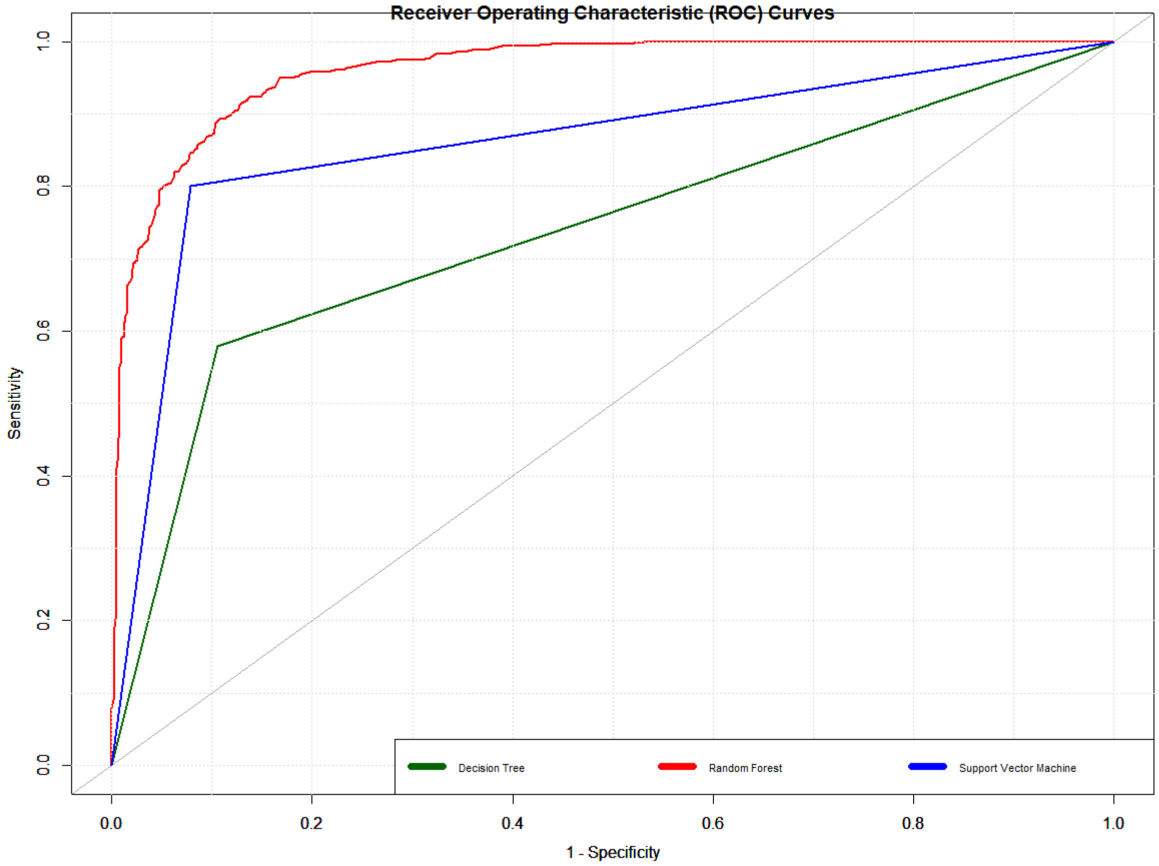

Figure 5 - ROC Curves of the three models trained

It appears that the Random Forest has the best diagnostic ability from the classification models trained. However, because we used a bootstrap dataset as training and the rest observations as test dataset only once, we would like to observe how our models behave on a lot different bootstrap training models in order to have an overall average assessment of our models. So, we executed the forementioned procedure on 100 different bootstrap datasets, recording each time the In-Sample Accuracy, Out-Of-Sample Accuracy, Adjusted Rand Index and Area Under the Curve (AUC), calculating at the end their average scores. The In-Sample Accuracy measures how well the model classifies the observations that it was trained. It does not produce much to the assessment of the predictive ability of the model, but compared to Out-Of-Sample Accuracy, we can identify if our model overfits the trained data (e.g., large value of the former with low value of the latter). The Out-Of-Sample Accuracy measures the proportion of correct classifications on new 'unseen' data. The Adjusted Rand Index measures the degree of overlapping between two data classes. Area Under the Curve is the summarization of ROC curve in one number and the higher the AUC score is, the most the model's curve deviates from the diagonal of the ROC curve plot (random classifier), thus implying better classification. The average scores from the procedure executed for our models are:

Metrics / Models	Accuracy (In-Sample)	Accuracy (Out-Of-Sample)	ARI	AUC
Tree	0.79	0.76	0.27	0.71
Forest	1.00	0.89	0.59	0.87
SVM	0.90	0.88	0.57	0.86

Table 1 - Average Metrics of the three models for the 100 bootstrap samplings

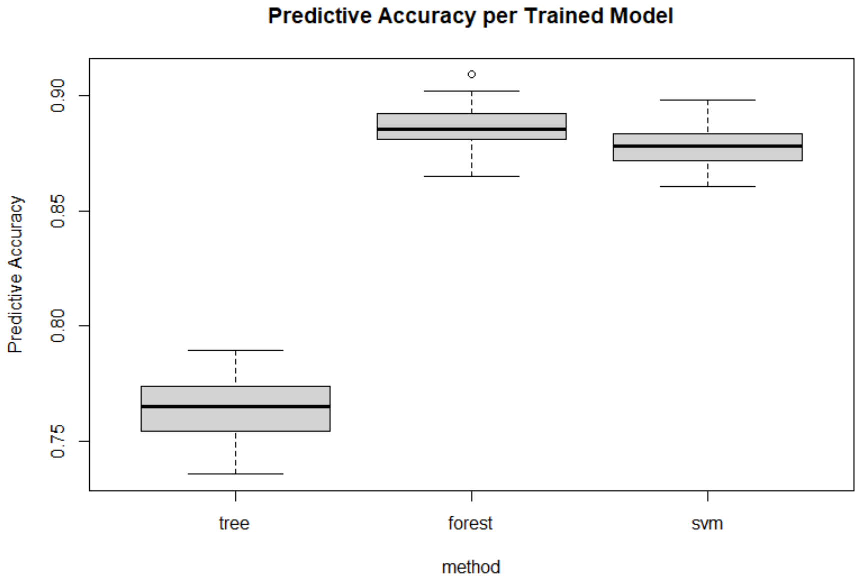

Figure 6 - Predictive Accuracy of the three models for each of the 100 bootstrap samplings

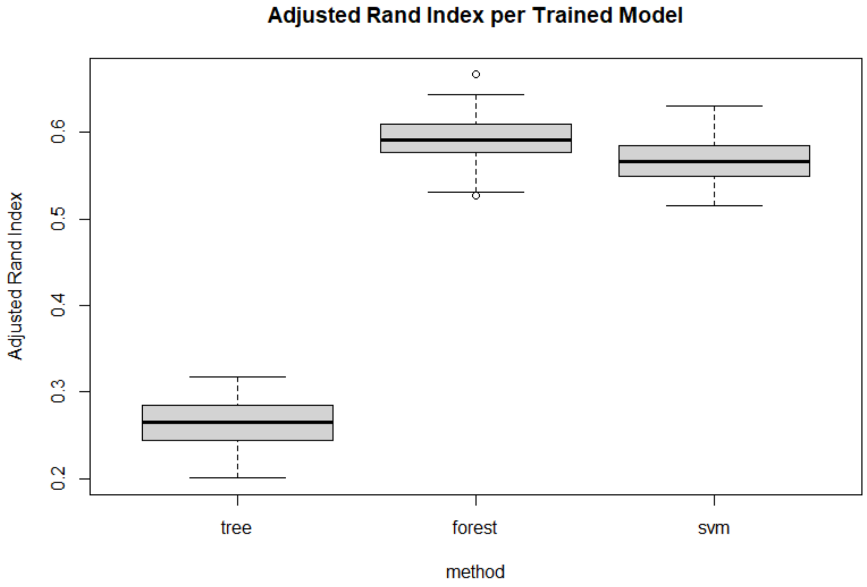

Figure 7 - Adjusted Rand Index of the three models for each of the 100 bootstrap samplings

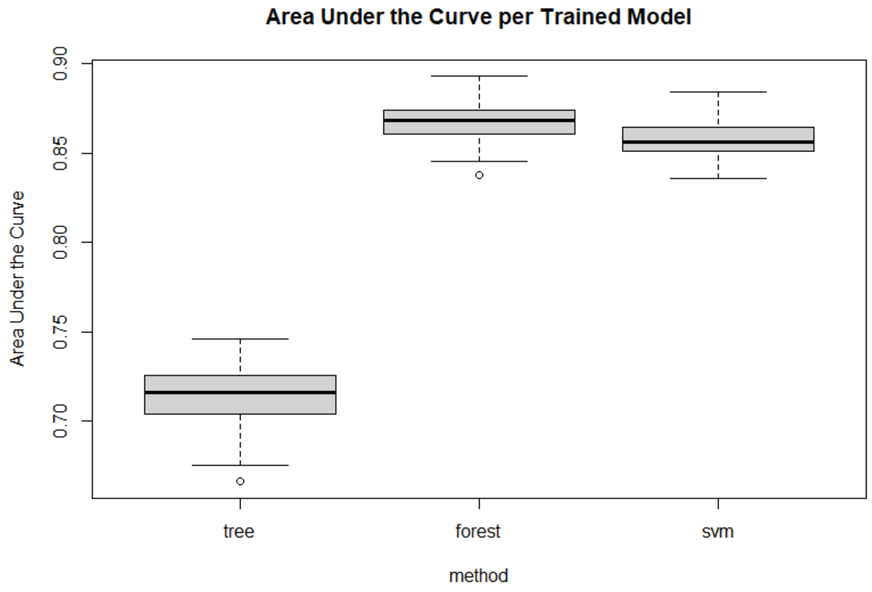

Figure 8 - Area Under the Curve of the three models for each of the 100 bootstrap samplings

Clustering Model

Then, we tried to create some clusters of the counties based on their demographic characteristics and then use the economic characteristics for the description of the forementioned clusters. Firstly, we separated the demographic variables from the dataset in order to use them for the creation of the clusters. Because the demographic variables are numeric variables which describe different metrics, mainly number of people and percentages, and in order to gain better performance we scaled our data. The scaling of the data was done for each variable by subtracting the column mean and then dividing each column by its standard deviation.

Then we conducted a hierarchical clustering of the dataset. We started by assigning each observation to its own cluster, each containing just one observation. Then, iteratively, by using an agglomeration method provided, we were merging the closest (most similar) pair of clusters into a single cluster. The distances (similarities) between different clusters were being computed via a provided distance measure. We tried all possible combinations of 'Euclidean', 'Maximum' and 'Manhattan' distances with 'single-link', 'complete-link', 'average-link', 'Ward' and 'Centroid' linkage functions. From the forementioned combinations, only the combinations of the 'Ward' linkage function with any of the forementioned measured distances provided acceptable clusters. The assessment of the clustering solution was done by plotting their dendrogram and unrooted phylogenetic tree.

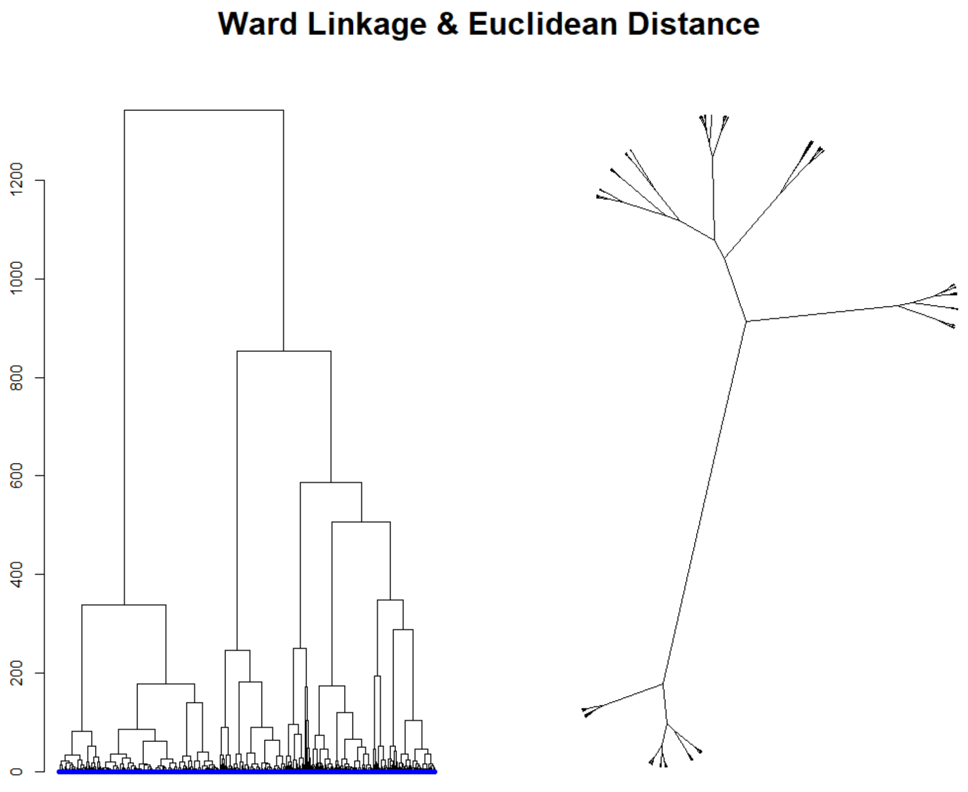

Figure 9 - Ward Linkage & Euclidean Distance Clustering Dendrogram and Phylogenic Tree

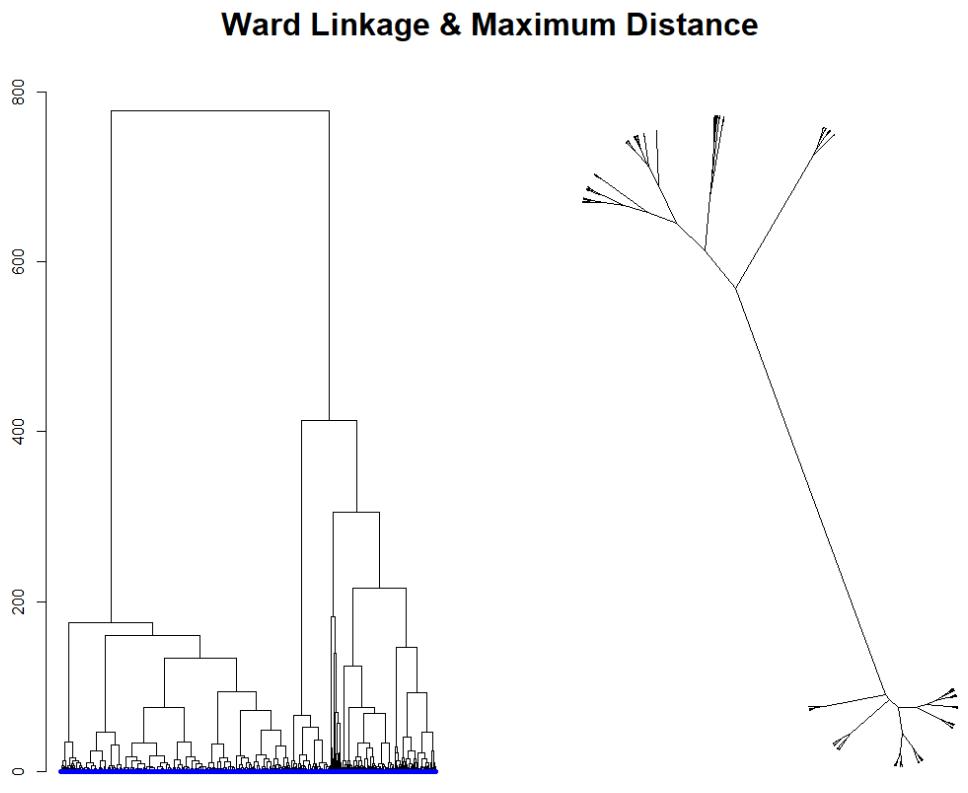

Figure 10 - Ward Linkage & Maximum Distance Clustering Dendrogram and Phylogenic Tree

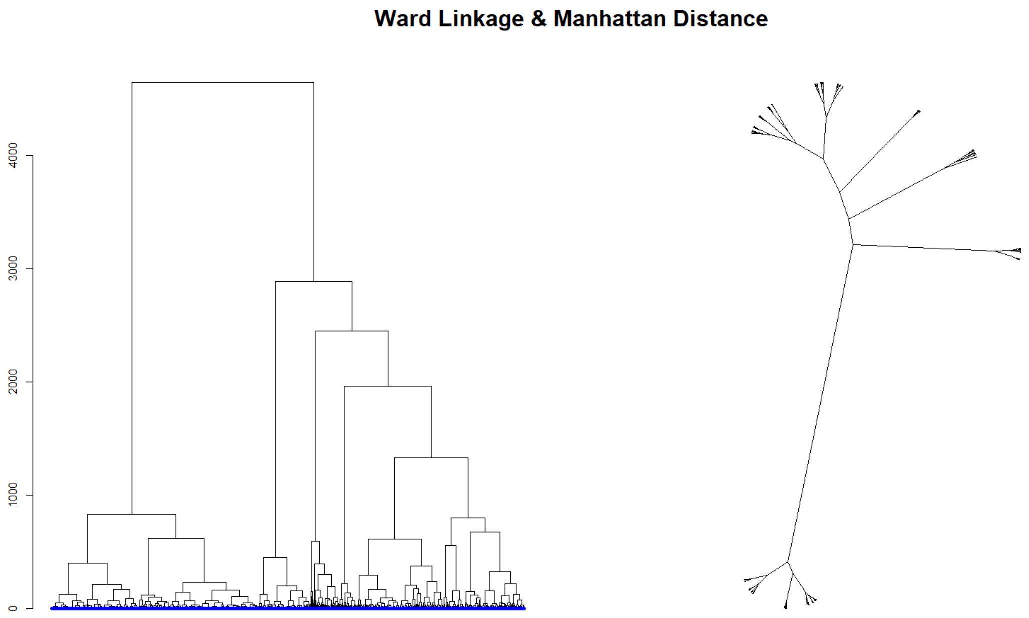

Figure 11 - Ward Linkage & Manhattan Distance Clustering Dendrogram and Phylogenic Tree

We then plotted the silhouette information using Ward's linkage and a combination of Euclidean, Maximum and Manhattan distance with 2, 3, 4, 5, and 6 clusters. Because the initial Silhouette values were low, we took the best combination of the forementioned parameters and at each iteration we were adding a group variable of the clusters proposed to the dataset. We were then training a Random Forest to provide us with the Variable Importance and at each iteration we were dropping the least important variable together with the grouping variable and were calculating again the Silhouette values. After a few iterations and the deletion of the variables holding information about the percents of Native Hawaiian and Other Pacific Islander alone, American Indian and Alaska Native alone and Two or More Races, we ended up with the following Silhouette value indicating a number of two clusters.

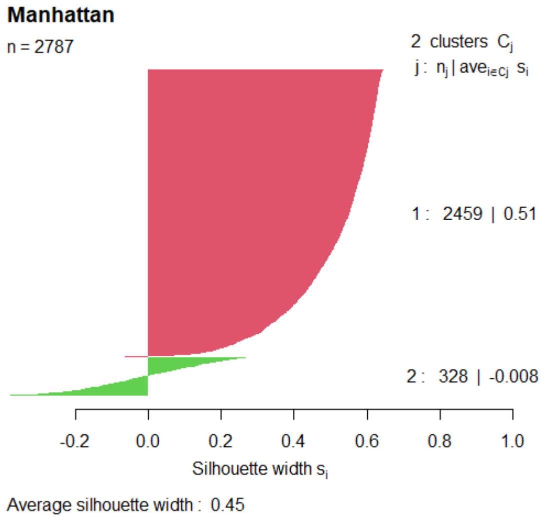

Figure 12 - Silhouette Information about Ward Linkage Clustering with Manhattan distance and 2 clusters

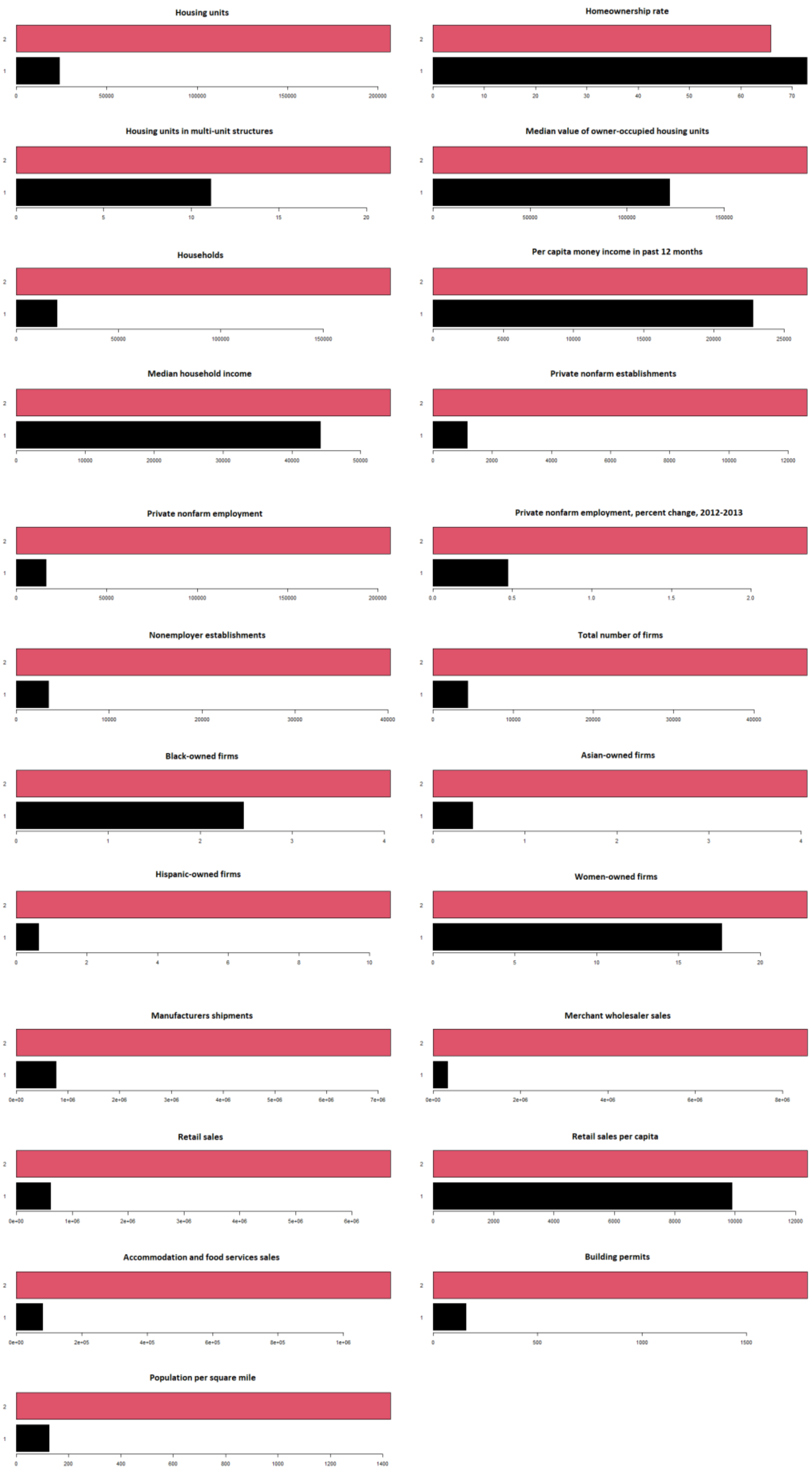

Figure 13 - Average Values of the 2 clusters

Conclusions

To sum up, from the classification metrics and plots, we can observe that on average Random Forest and Support Vector Machine techniques can classify whether Trump got more than 50% of the votes or not (group '1' and '0' respectively) a lot better than the Decision Tree model trained. SVM model has more compact behavior than the Random Forest model having approximately the same accuracy in the In-Sample and Out-Of-Sample datasets. However, the Random Forest model has overall the best score on all metrics measured above. Thus, we select the Random Forest model as our predictive model for the classification problem described above.

From the clustering perspective of the USA Counties, we can observe that on average, the two clusters relate to the different levels of economic characteristics measured at a county level. Cluster 2 is the cluster with the more housing units, non-farm and non-employer establishments, median income, total number of firms, retail sales and building permits. The counties of the second cluster are also denser in population than the counties of the first cluster, that also being a sign of an economic thrive of them related to the counties of the first cluster.

References

Karlis, D., 2022. *Classification.* Athens: s.n.

Karlis, D., 2022. *Clustering.* Athens: s.n.