Project-2

1) Partition the merged dataset into a training set and a validation set using the holdout method or the cross-validation method. How did you partition the dataset?
   Holdout method was used to partition the dataset.

   Train set has 80% of the data while validation set has 20%. Random_state = 1

   For the Regression problems the response variable was the column Democratic and Republican.

   For classification and clustering the response variable was Party.

   The Predictor variables included all columns except State, County and FIPS since these are not demographic information.

2) Standardize the training set and the validation set
   Used StandardScaler to standardize the training and validation data.

3) Build a linear regression model to predict the number of votes cast for the Democratic party in each county. Consider multiple combinations of predictor variables. Compute evaluation metrics for the validation set and report your results. ***What is the best performing linear regression model? What is the performance of the model? How did you select the variables of the model?***
   • Repeat this task for the number of votes cast for the Republican party in each county.
   The best linear regression model for Democratic party came out with the following predictors:

   Total Population, Percent White, not Hispanic or Latino, Percent Less than High School Degree, Percent Less than Bachelor's Degree.

   The variables were selected by using the conclusion of the Project -1 (exploratory analysis). This gave an idea as to what variables affected the response variable. To gain more confidence about the variable selection, a correlation matrix was generated which can be seen in the code file. Total Population has a higher and positive correlation which is evident from this matrix. Also experimented with a LASSO regression model. The surprising part was that the coefficients weren't reduced to 0. But it did provide some idea about the significance of variables. Based on these inferences many regression models were run with different variables. The best model was chosen with the highest adjusted R-squared value.

   Coefficients of variables = [65395.60625571 1398.31650451 2079.40174926 -8779.75132538]
   Intercept = 25935.015690376567
   R_squared = 0.9202820206063544
   Adjusted R_squared = 0.918919320103899
   Root Mean squared error = 20410.414822848878

   Following are the predictors which resulted in the best linear regression model for Republican party:
   Total Population
   Percent White, not Hispanic or Latino
   Percent Foreign Born
   Percent Age 65 and Older
   Median Household Income
   Percent Less than High School Degree
   Percent Less than Bachelor's Degree
   Percent Rural

   Coefficients of variables = [38912.28550654 3759.63248713 2393.50255217 -4216.97161912 2419.75441205 -2027.0167158 -5791.66501256 2734.08540567   4611.74495947]
   Intercept = 20875.02510460251

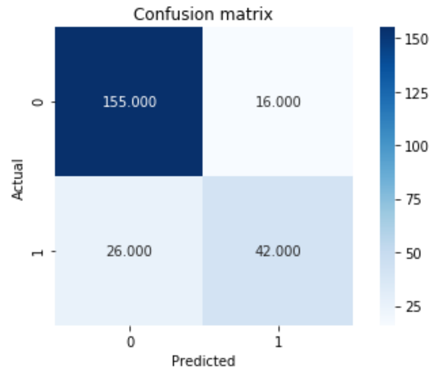
R_squared = 0.8276299094243877
Adjusted R_squared = 0.8208555390524205
Root Mean squared error = 21847.363645967314

4) Build a classification model to classify each county as Democratic or Republican. Consider at least two different classification techniques with multiple combinations of parameters and multiple combinations of variables. Compute evaluation metrics for the validation set and report your results. ***What is the best performing classification model? What is the performance of the model? How did you select the parameters of the model? How did you select the variables of the model?***

The two classification techniques used are Decision Tree and K-Nearest Neighbor. As the problem states, numerous models were run with different combination of parameters and variables. Here also, we selected variables based on our inference from the first project. However, after linear regression models, we had a fair idea as to which variables contribute more on the target variable. Different models were run by incrementing the parameter and keeping the variables constant. The same was done but in a vice-versa manner to see if the results improved. The following metrics were computed the find out the best model:
Accuracy, Error, Precision, Recall, F1 score, F1 score(macro) and F1 score(weighted).
However, the conclusion was drawn by considering the F1 score(macro). By seeing the map from Project-1, it’s evident that majority of the counties favor Republicans hence, the dataset is not a balanced data. This is also evident from the F1-scores. The F1-score for both the classes should be good. If it classifies one class label correctly then, it doesn’t mean that the model will generalize better. Looking at the **average** F1-scores tells us which model performs better for both the classes. Since our data is not highly imbalanced hence, we will consider macro-averaging. This does not take class imbalance into account.
Best performing model: KNN with n-neighbors = 3

Predictors:
Total Population
Percent White, not Hispanic or Latino
Percent Black, not Hispanic or Latino
Percent Age 65 and Older
Median Household Income
Percent Less than High School Degree
Percent Less than Bachelor's Degree

Accuracy = 0.8242677824267782
Error = 0.17573221757322177
Precision = array([0.85635359, 0.72413793])
Recall = array([0.90643275, 0.61764706])
F1_score = array([0.88068182, 0.66666667])]
F1 score (average: macro) = 0.7736742424242424
F1 score (average: weighted) = 0.8197904780017751



5) Build a clustering model to cluster the counties. Consider at least two different clustering techniques with multiple combinations of parameters and multiple combinations of variables. Compute unsupervised and supervised evaluation metrics for the validation set with the party of the counties (Democratic or Republican) as the true cluster and report your results. What is the best performing clustering model? What is the performance of the model? How did you select the parameters of model? How did you select the variables of the model?

The clustering techniques used are K-Means and DBSCAN. Since our target variable here is Party and it can have only two values hence, n-clusters for K-means is 2. The parameters were such as n_init and random_state was randomly chosen. Multiple models were run with different parameters till no performance improvement was seen. The various parameter values on which the models were experimented on were n_init = 1, 10, 20; random_state = 0, 1, 2. For DBSCAN, the parameters were eps = 2, 5, 10; min_samples = 10, 15, 20; metric = "euclidean".

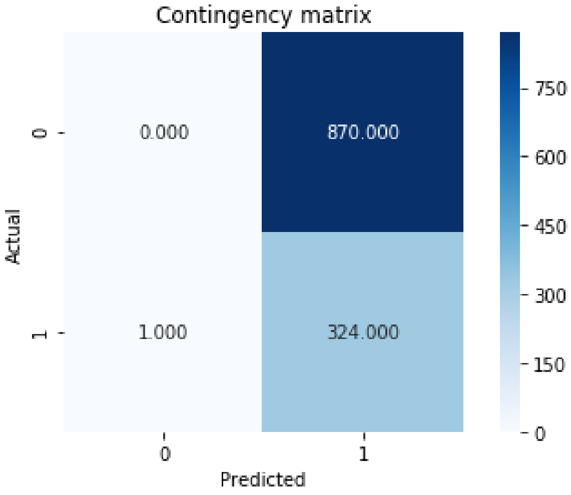
The models were run on different combination of variables. In this task also we used the inference from previous project and, we could see the trend and correlation between the variable selection and performance.

The best performing clustering model is DBSCAN.

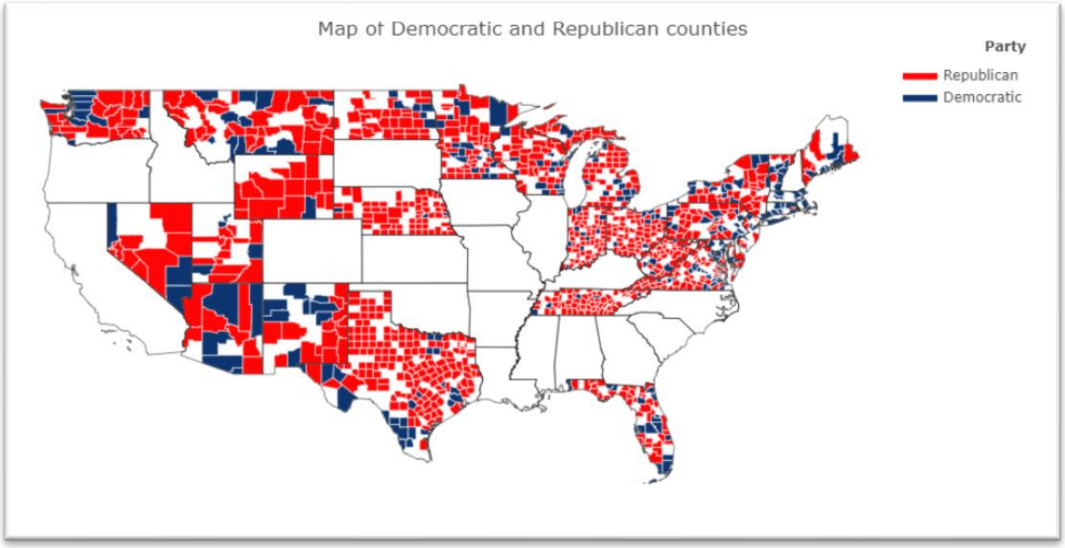
Predictors: “Total Population”, “Percent White, not Hispanic or Latino”, “Percent Less than Bachelor’s Degree”, “Percent Less than High School Degree”, “Percent Age 65 and Older”

Parameters: eps = 5, min_samples = 10, metric = "euclidean"

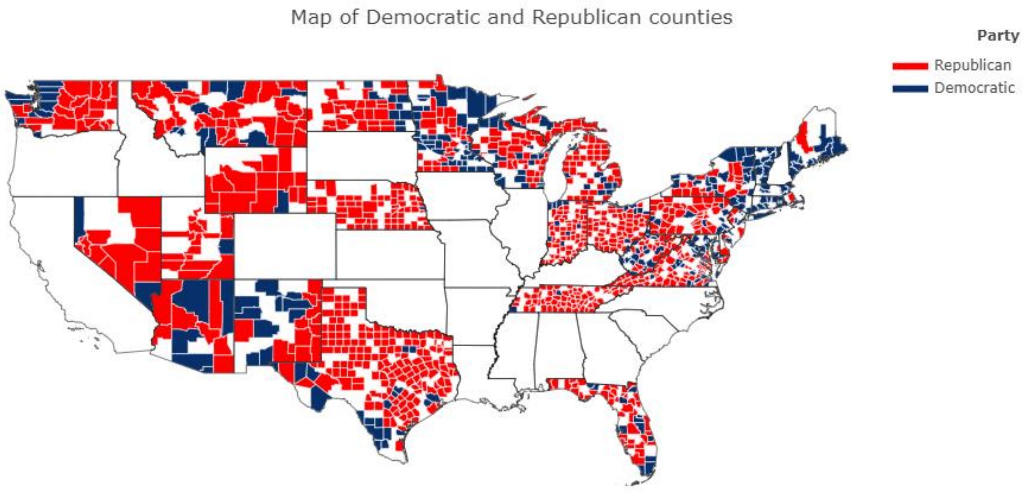
The evaluation metrics used were Adjusted Rand Index (Supervised) and Silhouette coefficient (Unsupervised). To select the best model, we considered Silhouette coefficient as we were emphasizing on finding clusters that are cohesive and well-separated. The silhouette_coefficient for the best model is 0.8014334712935183. However, the adjusted Rand Index for this model is very low which is 0.0028041107323011935. In general, all the metrics should be considered to select the best model. But in all the scenarios, the Adjusted Rand Index was coming around ~0.3 and Silhouette coefficient around ~0.4.
In the Scikit documentation it’s mentioned that even if Adjusted Rand Index is 1 that doesn’t guarantee that the labels are correctly classified. Hence, we settled with DBSCAN model which yielded the highest Silhouette coefficient.



6) Create a map of Democratic counties and Republican counties using the counties' FIPS codes and Python's Plotly library (plot.ly/python/county-choropleth/). Compare with the map of Democratic counties and Republican counties created in Project 01. What conclusions do you make from the plots?



Project-2: Figure



Project-1: Figure

It is evident from Project-2 Figure that most of the counties prefer favor Republicans. This map is plotted using the best classifier from Task-4. If we compare with the Project-1 Figure, then we'll see that most of the counties have been correctly classified as Democratic and Republican barring few.

7) Use your best performing regression and classification models to predict the number of votes cast for the Democratic party in each county, the number of votes cast for the Republican party in each county, and the party (Democratic or Republican) of each county for the test dataset (demographics_test.csv). Save the output in a single CSV file. For the expected format of the output, see sample_output.csv.

The output is saved in the file named output.csv.