**CS 418: Project-2 Report**

**Team 18: Ashwin Deshpande, Apurva Raghunath, Nandana Shimoga Prasad**

**Task 1 (5 pts).** 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?

**Solution**: Hold-out method is used to partition the data into training and validation sets by using 75% of the data in the training set and 25% of the data into test set.

**Task 3 (25 pts).** 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.

**Solution**:

**Regression models to predict Democratic values**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Number | Selected variables | R2 | Adjusted R2 | RMSE | Conclusions |
| 1. | Linear Regression using all variables | 0.934 | 0.933 | 14771.994 | Since the Adj. R2 value is high it is a good model |
| 2. | Linear Regression using ‘Total Population’, ‘Percent White, not Hispanic or Latino’,’Percent Black, not Hispanic or Latino’,’Percent Hispanic or Latino’,’Percent Foreign Born'. | 0.927 | 0.927 | 14592.862 | Adj. R2 value is slightly lower than the previous model. So this is not a better model |
| 3. | Linear Regression using ‘Total Population’,’Percent Black, not Hispanic or Latino’,’Percent Less than Bachelor's Degree’ | 0.950 | 0.950 | 12456.892 | This is the best model as it has the highest R2 value |
| 4. | LASSO regression with all variables | 0.934 | 0.933 | 14768.885 | Not the best model as its Adj. R2 is less than the previous model |

**Conclusion:**   
The Linear Regression using ‘Total Population’,’Percent Black, not Hispanic or Latino’,’Percent Less than Bachelor's Degree’ is the best model as it has highest Adjusted R2 and lower RMSE value than other models.

**Regression models to predict Democratic values**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model # | Selected variables | R2 | Adjusted R2 | RMSE | Conclusions |
| 1. | Linear Regression using all variables | 0.724 | 0.712 | 15962.431 | It has an average Adj. R2 and RMSE values |
| 2. | Linear Regression using ‘Total Population’, ‘Percent White, not Hispanic or Latino’,’Percent Black, not Hispanic or Latino’,’Percent Hispanic or Latino’,’Percent Foreign Born'. | 0.670 | 0.668 | 17111.714 | This model is not good as it has low Adj. R2 value but a high RMSE value |
| 3. | Linear Regression using 'Total Population', 'Percent White, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Age 65 and Older', 'Percent Unemployed', 'Median Household Income', 'Percent Rural' | 0.730 | 0.730 | 15749.245 | This is the best model with high Adj. R2 value and a lower RMSE value. |
| 4. | LASSO regression with all variables | 0.724 | 0.712 | 15962.567 | This model has an average Adj. R2 and RMSE values |

**Conclusion:**

The Linear Regression using ‘Total Population’, ‘Percent White, not Hispanic or Latino’,’Percent Hispanic or Latino’, ’Percent Foreign Born', 'Percent Age 65 and Older', 'Percent Unemployed', 'Median Household Income', 'Percent Rural' is the best model as it has highest Adjusted R2 and lower RMSE value than other models.

**Task 4 (25 pts).** 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?

**Solution**:

**Model Performances**

1. **Decision Tree**

Table

Description automatically generated

1. **K Nearest Neighbors**

Table

Description automatically generated

1. **Naïve Bayes**

Table

Description automatically generated

1. **Support Vector machine**

Table

Description automatically generated

* K-Nearest Neighbour with k = 3 using the predictor variables ‘Percent White, not Hispanic or Latino’, ‘Percent black, not Hispanic or Latino’, ‘Percent Hispanic or Latino’, ‘Percent Age 29 and Under’, ‘Percent Less than High School Degree’ and ‘Percent Less than Bachelor's Degree’ gives the best F1 Score.
* Accuracy: 0.85, Precision: 0.72, Recall: 0.66, F1 Score: 0.69

**Task 5 (25 pts).** 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?

**Solution**:

**Model Performances**

1. **Hierarchical Clustering – Single Linkage Method**

Graphical user interface, text, application, chat or text message

Description automatically generated

1. **Hierarchical Clustering – Complete Linkage Method**

Graphical user interface, text, application, chat or text message

Description automatically generated

1. **KMeans Clustering**

Graphical user interface, text, application, chat or text message

Description automatically generated

1. **DBSCAN Clustering – Eps: 1.5, MinPts: 5**

Graphical user interface, text, application, chat or text message

Description automatically generated

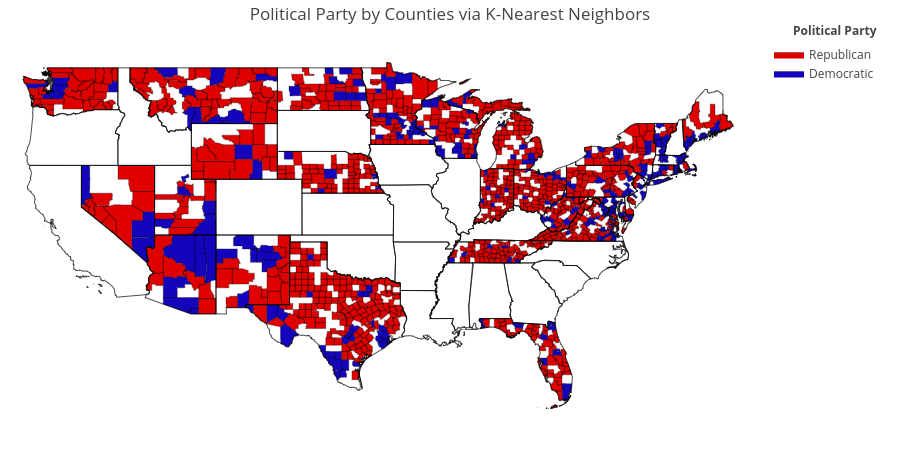
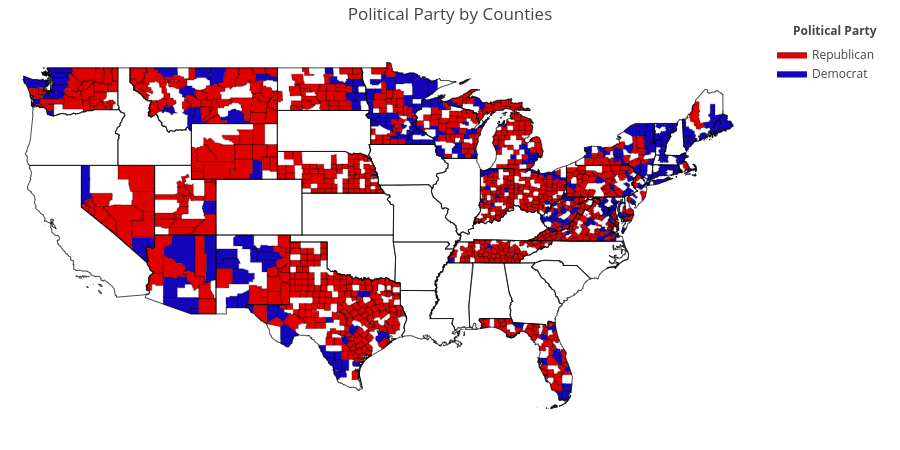
1. **DBSCAN Clustering – Eps: 1.0, MinPts: 3**Graphical user interface, text, application, chat or text message

   Description automatically generated

**Conclusion**: For the optimum variable: ‘Total Population’ the Clustering score is highest for all clustering algorithms except for Kmeans. Best performing clustering algorithms do not however give best true clusters, i.e. their classification performance is poor.

**Task 6 (10 pts).**

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?

**Conclusion:**  We see that our prediction is different from some of the true labels. Because we are using K-Nearest Neighbors classification we find that some of the counties have nearest neighbors of the other party based on Race, Age and Education properties. Following is true label map:  ****

**Task 7 (5 pts).** 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.

**Solution**:

Table

Description automatically generated

This is the output generated. Output is generated in ouput.csv.