CS418 Project 2
Regression, Classification and Clustering

Frank Errichiello, Tomasz Hulka, Ankit Kumar Singh

Overview

The purpose of this report is to build models using regression, classification, and clustering techniques on datasets of election data and demographics for a large sample of US counties and find these best performing models. Using these models we project a map of the counties with their predicted political party.

Task 1

We partitioned the dataset using the holdout method, with a 75% training set and a 25% test set.

Task 2

We used the sklearn library's standard scaler to scale our numbers to standardize the training and validation set.

Task 3

Prediction of Democratic votes

- Simple linear regression using "Population" as the predictor variable.

```
Model coefficient [74711.50206856]
R_squared value:  0.9436415220931651
Adjusted R_squared value:  0.9435784812901373
```

- Simple linear regression using "Percent Less than High School Degree" as the predictor variable.

```
Model coefficient [-8137.73810376]
R_squared value:  0.022734480001930003
Adjusted R_squared value:  0.02164134183638411
```

- Multiple linear regression using "Population", "Median Household Income" as the predictor variables.

```
Model coefficient [73067.37334453  6279.76422366]
R_squared value:  0.939337563328897
Adjusted R_squared value:  0.939201701208693
```

- Multiple linear regression using all the predictor variables.

```
Model coefficient [ 69224.38708039  -3209.1591268   -1023.23488454  -6931.14708179
   3973.74580741    194.19056985  -5299.5676761   -1853.22320472
   1471.25963216   1467.0213699    4037.7699931  -10519.02638282
   -158.13004477]
R_squared value:  0.9338361960241593
Adjusted R_squared value:  0.9326318491941099
```

- Multiple linear regression using "Population", “Median Household Income”,”Percent white, not hispanic or latino”, “Percent Less than Bachelor's degree” as the predictor variables.

```
Model coefficient [71012.84796525  -345.05366382  1157.04687807 -8608.17042826]
R_squared value:  0.947734113056962
Adjusted R_squared value:  0.9474994738338731
```

The best performing model for the prediction of democratic votes is the multiple linear regression model using “Population”, “Median Household Income”, “Percent white, not hispanic or latino”, “Percent Less than Bachelor's degree” as the predictor variables. The model performs well with these four predictors with an adjusted R^2 value of 0.947. Selection of the variables are consistent with Project 1’s conclusion and we also see the adjusted R^2 value decrease if all the variables are used as predictors.

Prediction of Republican votes

- Simple linear regression using 'Population' as the predictor variable.

```
Model coefficient [45306.87897032]
R_squared value:  0.6718468162068597
Adjusted R_squared value:  0.6714797544800217
```

- Simple linear regression using 'Percent Less than High School Degree' as the predictor variable.

```
Model coefficient [-6381.7748349]
R_squared value:  0.03593599340559725
Adjusted R_squared value:  0.03485762203356779
```

- Multiple linear regression using "Population", "Median Household Income" as the predictor variables.

```
Model coefficient [44042.16950014  4830.56902305]
R_squared value:  0.6841236214388341
Adjusted R_squared value:  0.6834161715428404
```

- Multiple linear regression using "Population", "Median Household Income","Percent white, not hispanic or latino", "Percent Less than Bachelor's degree" as the predictor variables.

```
Model coefficient [44609.62027579  3068.87458444  3337.02252553 -2140.80688346]
R_squared value:  0.6837837434980034
Adjusted R_squared value:  0.6823641418975455
```

- Multiple linear regression using all predictor variables.

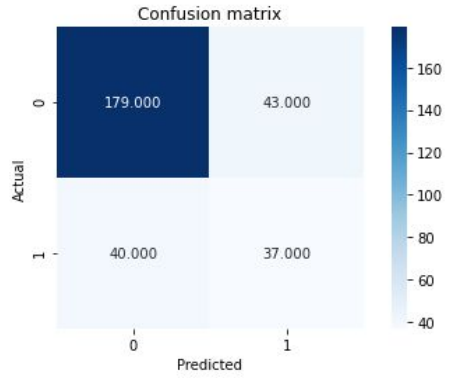
```
Model coefficient [45467.5097118   1769.95034533 -3141.42063749  1167.17323402
 -6463.65917143 -1121.73432851  -955.67013341  2580.74056065
  5910.97457236  2037.10575397  3530.42010898 -3156.11275644
 -5992.05181735]
R_squared value:  0.7239014362949742
Adjusted R_squared value:  0.7188757514038702
```

The best performing model for the prediction of republican votes is the multiple linear regression model using all variables for the prediction. The model does not perform too well with an adjusted R^2 value of 0.719. But using all the values for the model as it gives the best adjusted R^2 value out of the other models.

Task 4

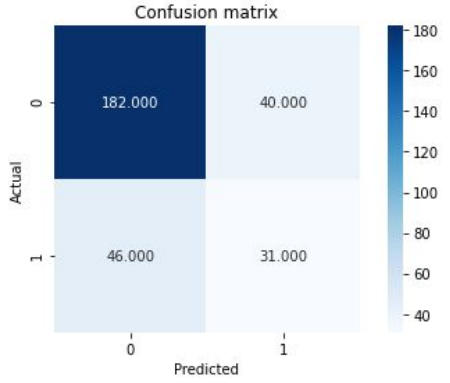
Decision tree classifier with no random state and variables 'Total Population', 'Percent White', and 'Percent Rural'

```
Number of decision tree nodes:  347
Accuracy:  0.7224080267558528
Error:  0.2775919732441472
Precision:  [0.8173516 0.4625   ]
Recall:  [0.80630631 0.48051948]
F1 score:  [0.81179138 0.47133758]
```



Decision tree classifier with random state and variables 'Median Household Income', 'Percent Unemployed', 'Percent Less than High School Degree' and 'Percent Less than Bachelor's Degree'

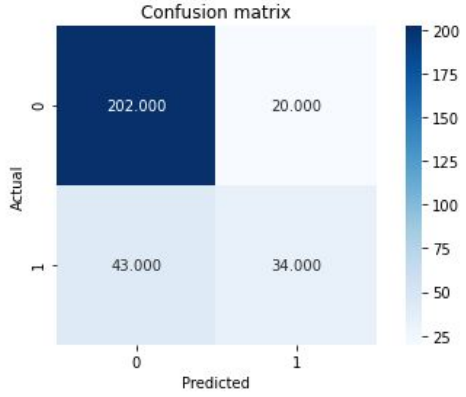
```
Number of decision tree nodes:  309
Accuracy:  0.7123745819397993
Error:  0.2876254180602007
Precision:  [0.79824561 0.43661972]
Recall:  [0.81981982 0.4025974 ]
F1 score:  [0.80888889 0.41891892]
```



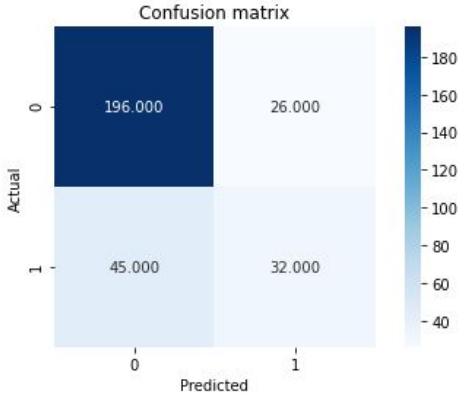
Naive Bayes classifier with variables 'Total Population', 'Percent White', and 'Percent Rural'

Naive Bayes classifier with variables 'Median Household Income', 'Percent Unemployed', 'Percent Less than High School Degree' and Percent Less than Bachelor's Degree'

```
Accuracy:  0.7892976588628763
Error:  0.21070234113712372
Precision:  [0.8244898  0.62962963]
Recall:  [0.90990991 0.44155844]
F1 score:  [0.86509636 0.51908397]
```




```
Accuracy:  0.7625418060200669
Error:  0.23745819397993306
Precision:  [0.81327801 0.55172414]
Recall:  [0.88288288 0.41558442]
F1 score:  [0.84665227 0.47407407]
```

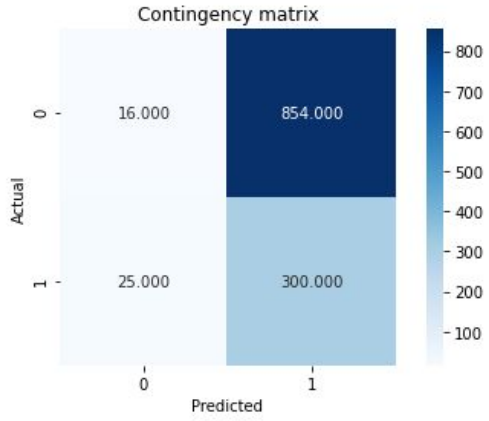


The best performing classification model is the Naive Bayes classifier using variables 'Total Population', 'Percent White', and 'Percent Rural'. It had an accuracy of 78.93% and the most True Positive values out of all 4 classifiers used. It also was the best in terms of evaluation metrics with precision of 62.96%, recall of 44.15%, and F1 Score of 51.91%. The variables were selected because they gave the best indicator of a Republican county (a rural place with smaller populations and a white demographic). The model had no parameters because we did not want to use var_smoothing and there was nothing else to change.

Task 5

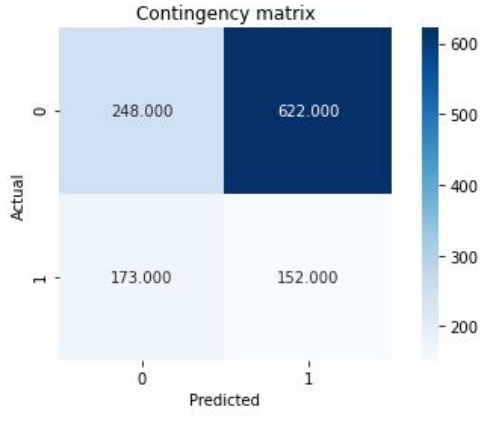
Hierarchical clustering with complete linkage, euclidean distance metric and using variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born'

```
Supervised metric:  0.05057355502877359
Unsupervised metric:  0.64412187209008
```



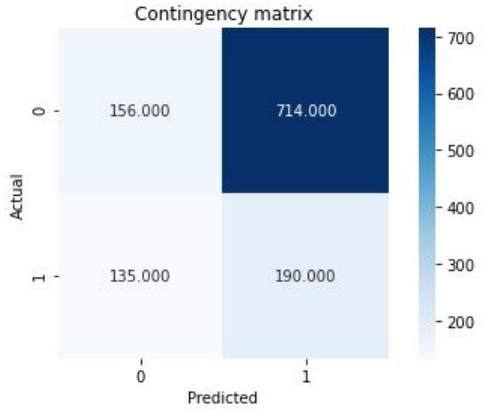
Hierarchical clustering with complete linkage, jaccard distance metric and using variables 'Percent Female', 'Percent Age 29 and Under', 'Percent Age 65 and Older', 'Median Household Income', 'Percent Unemployed'

```
Supervised metric:  0.09223396339355758
Unsupervised metric:  0.37214725289194917
```



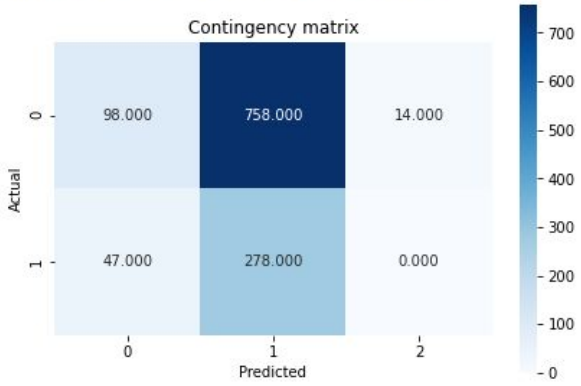
DBSCAN clustering with eps = .5, minimum samples = 15 and using variables 'Percent White, not Hispanic or Latino','Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born'

```
Supervised metric:  0.12908114450483138
Unsupervised metric:  0.5786850122739186
```



DBSCAN clustering with eps = 1, minimum samples = 10 and using variables 'Percent Female', Percent Age 29 and Under', 'Percent Age 65 and 'Older', 'Median Household Income', 'Percent Unemployed'

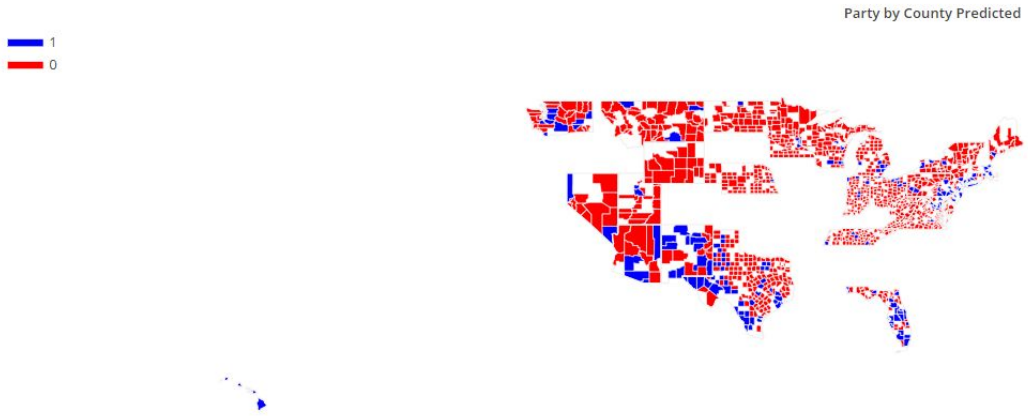
```
Supervised metric:  0.008381757165266535
Unsupervised metric:  0.2213462398681104
```



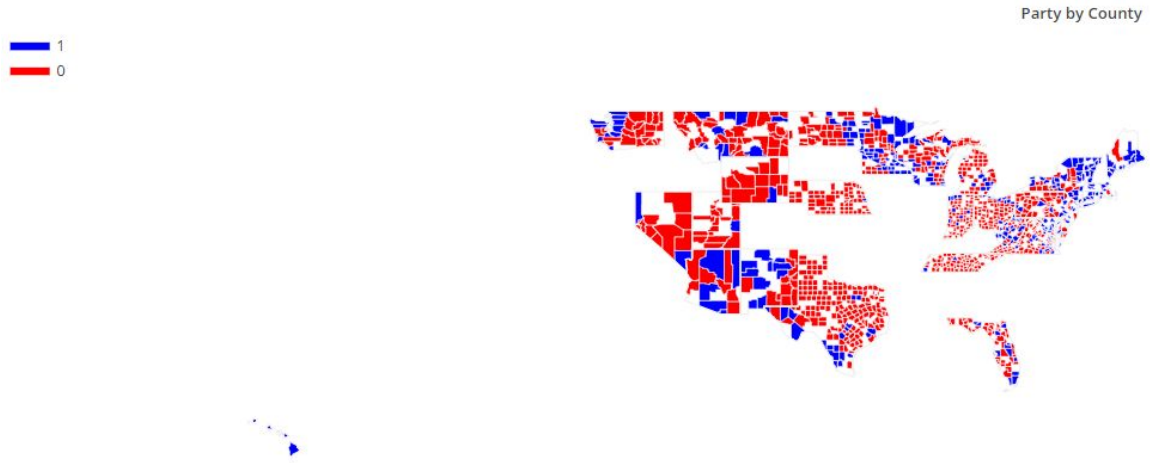
The best performing model was Hierarchical clustering with complete linkage and variables 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born'. It had an unsupervised metric of 64.41% and supervised metric of 5.06. The parameters selected were complete linkage and euclidean distance because single linkage did not accurately cluster the data. The variables were selected because they all describe the demographic of a given county and provide a rough estimation of diversity.

Task 6

The plot of the predicted counties voting on the “merge_train” data set using our best classification method from task 4.



The plot of the original parties by county given in the “merge_train” data set



The conclusion made by looking at the plots is that the classifier is getting most but not all of the classifications correct, which correlates to the accuracy of the classifier being around 78%.

Task 7

For the prediction of the number of votes for the democratic party we used the multiple linear regression model using “Population”, “Median Household Income”, “Percent white, not hispanic or latino”, “Percent Less than Bachelor's degree” as the predictor variables. For the prediction of the number of votes for the republican party we used the multiple linear regression model using all variables for the prediction. Finally, for the prediction on the party we used the Naive Bayes classifier with the variables 'Total Population', 'Percent White', and 'Percent Rural'.