Predicting Electoral Results Using Census and RDH Data

Deven Hagen and Vibby Janardhan

Dr. Yilmaz - Machine Learning 1

21 October 2024

Statement/Project Goal
Elections are the backbone of representative democracy—they determine the officials who make the important decisions for our country. Electoral analysis has many applications, from understanding demographic behavior to identifying thematic shifts in public opinion. Accordingly, our project aims to utilize demographics, housing data, past electoral results, and other related attributes to improve our understanding of how different groups vote; using Virginia precinct-level data, we will predict the party alignment of voting districts in the 2020 Presidential Election. This data will be useful for politicians to better understand the desires of different constituencies they represent, and for citizens to obtain information about the state of various races.

Description of Dataset
In order to facilitate proper analysis, one must use a comprehensive dataset. We acquired our dataset from the Redistricting Data Hub (RDH), which has various datasets based on the Census and other official government projects (see source 1).
In particular, we used the “va_pl2020_official_vtds.csv” file from the download above. This file includes various critical variables, such as racial makeup, land area, housing units, and previous electoral results. The final dataset, before preprocessing, includes 463 attributes in total. We created a class attribute titled “WINPARTY” from the “PRES2020D” and “PRES2020R” columns, which represent the party that received the most votes in the 2020 election—each precinct will have a value of either “D”, representing the Democrats, “R”, representing the Republicans, or “T”, representing a tie. The dataset has 2465 instances, corresponding to each precinct in Virginia. In preprocessing, we will cut down the number of attributes significantly using attribute selection techniques such as Correlation and others to avoid the curse of dimensionality.

The dimension of our dataset currently is 462, because there are 463 attributes but the class does not count. The class distribution is as follows: 1747 “R”, 697 “D”, and 21 “T”. A number of columns are entirely missing values, and some precincts have disguised missing values for key attributes as well, although Weka does not show these. The dataset is skewed towards Republicans, although most of the “D” precincts are heavily Democratic-voting and thus outliers. In terms of the definitions of the attributes, a complete catalogue can be found if you click on the “Census and AutoBound Edge field definitions (.xlsx file)” link (see source 2). Although it is quite unfeasible to provide a detailed explanation of all 463 attributes, here is a representative sample:
GEOID20: The unique Census geo ID that corresponds to each precinct.
ALAND20: The land area of the precinct.
TAPERSONS: The total population of the precinct.
TAWHITEALN: The total White population in the precinct.
TABLACKALN: The total Black population in the precinct.
TAAMINDALN: The total American Indian and Alaskan Native population in the precinct.
TAASIANALN: The total Asian population in the precinct.
TANHPOALN: The total Hawaiian and Pacific Islander population in the precinct.
TAOTHERALN: The total population of all other races not previously mentioned in the precinct.
TAHOUSING: The total number of homes in the precinct.
PRES16D: The number of votes in the precinct for 2016 Democratic nominee Hillary Clinton.

PRES16R: The number of votes in the precinct for 2016 Republican nominee Donald Trump. There are more combinations and variations of attributes like these, but this provides a general idea of the attributes contained in our dataset.

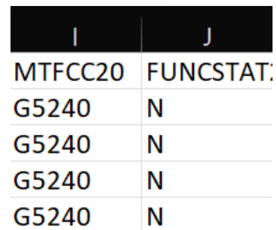
Intermediate Report - 10/7/24

Data Preprocessing

The next steps were to pare down our 463 attributes into a manageable number and remove instances with many missing values. As stated in our proposal, many of the columns in our dataset had either missing values or identical values for each instance. It would take up too much space to enumerate each attribute removed, but below are examples:

BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ
ECTA	NMEMI	CNECTA	NECTAC	CBSAPC	NECTAF	UA	UATYPE	UR	CD116	CD118

An example of columns missing every value



I	J
MTFCC20	FUNCSTAT:
G5240	N
G5240	N
G5240	N
G5240	N

An example of columns with identical values

After getting rid of these, we were down to 369 attributes—still a lot! Each election in our dataset was split into two columns: the number of Democratic votes and the number of Republican votes. The next step was normalization, and many of our methods came from ML Slides 4 (see source 3). In order to normalize the data, since some precincts have more people than others, we transformed each of these into the percentage of votes for each party. For example, we transformed USSEN18D, USSEN18R, and USSEN18L into USSEN18DPERC, which represents the percentage of total votes received by Democrats. We then repeated this process on the remaining election results.

The vast majority of the attributes remaining were related to race - particularly obscure combinations of races. To address the excess columns for race data, we combined the multiracial columns into a new attribute representing multiracial people. This consolidation of columns helped avoid the curse of dimensionality significantly. We had both the Voting Age and Total numbers for racial demographics, and we kept both to ascertain whether key patterns would emerge. Similarly to what we did with electoral results, we normalized the race data as percentages of the total population.

After normalizing race and electoral attributes, some null values became apparent due to a lack of data. Since these instances were missing very key values, particularly for past election data, we removed them from the dataset. After this removal, there were 2423 instances. We also deleted all instances with a class label of "T" (tied). Voters do not vote for a tie but instead choose a preferred candidate, so trying to predict precincts that will end in a tie is not a useful exercise. This left us with 2411 instances. When it came to the columns regarding housing, most were missing many values, so we kept it to just TAHOUSING, TAHOCCUPIED, and TAHVACANT, removing the others. Since TAHOCCUPIED + TAHVACANT = TAHOUSING, we removed TAHVACANT, since it can be derived from the other attributes. We then normalized the housing data on a per-capita basis.

We used decimal scaling for AREALAND and AREAWATR to ensure their influence was not exaggerated, since some of the values for these attributes approached 10^11. Additionally, we used min-max normalization to scale VAPERSONS (voting-age population) and TAPERSONS (Total population). Finally, we changed our class labels from "R" and "D" to 0 and 1, where "R" is 0 and "D" is 1. Changing each value to a number will help when we use our models on the dataset. To conclude, after preprocessing, there were 32 attributes and 2411 instances left in our modified dataset.

Attribute Selection

To identify which attributes should be used in a classification model, four attribute selection methods were employed: OneRAttributeEval, CorrelationAttributeEval, InfoGainAttributeEval, and ReliefFAttributeEval. The results of each selection technique allowed us to accurately choose our attribute set. Each previous election result (USSEN18DPERC, AG18DPERC, GOV17DPERC, LTGOV17DPERC, PRES16DPERC, AG13DPERC, GOV13DPERC, LTGOV13DPERC, PRES12DPERC) consistently ranked high, so we will keep them. Beyond that, VANWHTALN (total only white voting-age population), VANBLKALN (total voting-age black population), and AREALAND (total land area) also consistently ranked high. Therefore, the above 13 attributes were chosen for the model. The results for each attribute selection method can be seen below.

Ranked attributes:

94.8569	27	PRES16DPERC
94.7325	23	USSEN18DPERC
93.3637	26	LTGOV17DPERC
93.1564	24	AG18DPERC
92.7001	25	GOV17DPERC
91.7876	31	PRES12DPERC
91.6217	29	GOV13DPERC
91.4144	28	AG13DPERC
89.5064	30	LTGOV13DPERC
85.9394	5	VANWHTALN
84.6537	14	TNWHALN
79.4276	15	TNBLKALN
77.6856	6	VANBLKALN
77.022	1	AREALAND
76.1095	4	VAHISPANIC
75.4873	13	TAHISPANIC
75.3214	2	AREAWATR
73.8698	17	TNASIANALN
72.5425	8	VANASANALN
71.2982	12	TAPERSONS
71.2153	19	TNOTHRALN
70.7175	22	TAHOCCUPID
69.8465	3	VAPERSONS
68.8096	9	VANNHPOALN
68.6437	7	VANAIANALN
68.3119	18	TNNHPOALN
67.8142	16	TNAIANALN
67.7727	10	VANORALN
67.4824	11	VANM2RACES
67.3165	21	TAHOUSING
66.1551	20	TN2MRACES

OneR

Ranked attributes:

0.7888	27	PRES16DPERC
0.7865	26	LTGOV17DPERC
0.7824	24	AG18DPERC
0.7807	31	PRES12DPERC
0.7802	28	AG13DPERC
0.7797	29	GOV13DPERC
0.7765	25	GOV17DPERC
0.7758	23	USSEN18DPERC
0.7473	30	LTGOV13DPERC
0.6876	5	VANWHTALN
0.6648	14	TNWHALN
0.5316	15	TNBLKALN
0.4888	6	VANBLKALN
0.4346	4	VAHISPANIC
0.4191	13	TAHISPANIC
0.3609	1	AREALAND
0.3413	17	TNASIANALN
0.3145	19	TNOTHRALN
0.3082	8	VANASANALN
0.3075	3	VAPERSONS
0.3009	12	TAPERSONS
0.2394	20	TN2MRACES
0.2107	10	VANORALN
0.1813	11	VANM2RACES
0.1606	18	TNNHPOALN
0.1324	9	VANNHPOALN
0.0742	21	TAHOUSING
0.0579	2	AREAWATR
0.0387	7	VANAIANALN
0.0287	22	TAHOCCUPID
0.0174	16	TNAIANALN

Correlation

0.6916	23	USSEN18DPERC
0.6796	27	PRES16DPERC
0.6722	25	GOV17DPERC
0.6669	24	AG18DPERC
0.6639	26	LTGOV17DPERC
0.6253	31	PRES12DPERC
0.6228	28	AG13DPERC
0.6164	29	GOV13DPERC
0.5746	30	LTGOV13DPERC
0.4144	5	VANWHTALN
0.3728	14	TNWHALN
0.2569	1	AREALAND
0.2382	15	TNBLKALN
0.2034	6	VANBLKALN
0.1572	4	VAHISPANIC
0.1487	2	AREAWATR
0.1407	13	TAHISPANIC
0.1061	17	TNASIANALN
0.101	8	VANASANALN
0.0831	19	TNOTHRALN
0.0673	3	VAPERSONS
0.0639	12	TAPERSONS
0.0427	10	VANORALN
0.0384	20	TN2MRACES
0.038	22	TAHOCCUPID
0.0259	18	TNNHPOALN
0.0237	9	VANNHPOALN
0.0226	11	VANM2RACES
0	21	TAHOUSING
0	7	VANAIANALN
0	16	TNAIANALN

InfoGain

0.206555	23	USSEN18DPERC
0.183269	27	PRES16DPERC
0.180663	26	LTGOV17DPERC
0.178971	25	GOV17DPERC
0.175165	24	AG18DPERC
0.123077	28	AG13DPERC
0.113753	29	GOV13DPERC
0.110577	30	LTGOV13DPERC
0.10999	31	PRES12DPERC
0.082153	5	VANWHTALN
0.075585	14	TNWHALN
0.061844	15	TNBLKALN
0.056405	6	VANBLKALN
0.04872	1	AREALAND
0.041586	20	TN2MRACES
0.038933	17	TNASIANALN
0.037806	22	TAHOCCUPID
0.036286	8	VANASANALN
0.027004	11	VANM2RACES
0.02451	13	TAHISPANIC
0.024192	4	VAHISPANIC
0.020693	12	TAPERSONS
0.01809	3	VAPERSONS
0.014764	10	VANORALN
0.014508	9	VANNHPOALN
0.012837	21	TAHOUSING
0.009908	18	TNNHPOALN
0.009466	19	TNOTHRALN
0.000832	16	TNAIANALN
0.00083	7	VANAIANALN
-0.000177	2	AREAWATR

ReliefF

Train-Validation-Test vs K-fold Cross-Validation

We used Pandas to read our new dataset into Python as a Dataframe. We decided to make our training set 70% of the instances, our validation set 15%, and our testing set 15%. We then used Scikit-Learn's train-test split twice to create our sets, making sure to use the stratify parameter to keep the balanced class distributions. We took some of this code from the ML Slides 2 presentation (see source 4). These new datasets can be found in our drive: each attribute selection algorithm has a folder with the three datasets. Here is the code we used:

```
from sklearn.model_selection import train_test_split
import pandas as pd
import os

for folder in ("CORRELATION", "INFOGAIN", "ONER", "RELIEFF"):
    os.chdir(folder)
    df = pd.read_csv(f"{folder}_DATASET.csv")
    df_train, df_test_val = train_test_split(df, test_size=0.3, stratify =
df["WIN_PARTY"], random_state=42)
    df_test, df_val = train_test_split(df_test_val, test_size=0.5,
stratify = df_test_val["WIN_PARTY"], random_state=42)

    df_train.to_csv('train.csv', index=False)
    df_test.to_csv('test.csv', index=False)
    df_val.to_csv('val.csv', index=False)
    os.chdir('..')
```

However, we also tested implementations of K-fold cross-validation through Weka and decided that it was superior to simply creating train-validation-test sets due to its more comprehensive validation methods and simple implementation.

We soon realized when we opened WEKA that we had to switch our class labels from "0" and "1" to "R" and "D" because WEKA was reading the 0 and 1 as floats and hence the Naive Bayes classifier was not functioning.

Results and Evaluation

For each dataset—OneR, Correlation, InfoGain, ReliefF, and our own chosen one—we applied four models to classify the class variable: J48 decision tree, Naïve Bayes, OneRClassification, and K-nearest neighbor. We used a K-fold value of 10 for each model, as we found changing it from this value did not significantly improve the model. The screenshots of the results are below.

OneR Dataset

J48 decision tree: Accuracy = 0.954791, TP = 0.955, FP = 0.074, ROC = 0.961

```
Correctly Classified Instances        2302               95.4791 %
Incorrectly Classified Instances       109                4.5209 %
Kappa statistic                          0.8875
Mean absolute error                      0.0589
Root mean squared error                  0.1951
Relative absolute error                 14.5386 %
Root relative squared error             43.3486 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.093    0.964      0.973    0.969      0.888    0.961     0.966     R
                 0.907    0.027    0.931      0.907    0.919      0.888    0.961     0.905     D
Weighted Avg.    0.955    0.074    0.955      0.955    0.955      0.888    0.961     0.949

=== Confusion Matrix ===

    a    b   <-- classified as
 1685   46 |    a = R
   63  617 |    b = D
```

OneRClassifier: Accuracy = 0.948569, TP = 0.949, FP = 0.074, ROC = 0.937

```
Correctly Classified Instances        2287               94.8569 %
Incorrectly Classified Instances       124                5.1431 %
Kappa statistic                          0.8732
Mean absolute error                      0.0514
Root mean squared error                  0.2268
Relative absolute error                 12.6966 %
Root relative squared error             50.3972 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.088    0.965      0.963    0.964      0.873    0.937     0.956     R
                 0.912    0.037    0.906      0.912    0.909      0.873    0.937     0.851     D
Weighted Avg.    0.949    0.074    0.949      0.949    0.949      0.873    0.937     0.927

=== Confusion Matrix ===

    a    b   <-- classified as
 1667   64 |    a = R
   60  620 |    b = D
```

K-nearest neighbors: Accuracy = 0.941518, TP = 0.942, FP = 0.092, ROC = 0.927

```
Correctly Classified Instances        2270               94.1518 %
Incorrectly Classified Instances       141                5.8482 %
Kappa statistic                          0.8548
Mean absolute error                      0.0589
Root mean squared error                  0.2417
Relative absolute error                 14.5377 %
Root relative squared error             53.7163 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.113    0.956      0.963    0.959      0.855    0.927     0.950     R
                 0.887    0.037    0.904      0.887    0.895      0.855    0.927     0.842     D
Weighted Avg.    0.942    0.092    0.941      0.942    0.941      0.855    0.927     0.919

=== Confusion Matrix ===

    a    b   <-- classified as
 1667   64 |    a = R
   77  603 |    b = D
```

Naïve Bayes: Accuracy = 0.92866, TP = 0.929, FP = 0.046, ROC = 0.985

```
Correctly Classified Instances        2239               92.866  %
Incorrectly Classified Instances       172                7.134  %
Kappa statistic                          0.8337
Mean absolute error                      0.071
Root mean squared error                  0.2571
Relative absolute error                 17.5318 %
Root relative squared error             57.1262 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.912    0.029    0.987      0.912    0.948      0.841    0.985     0.993     R
                 0.971    0.088    0.813      0.971    0.885      0.841    0.984     0.959     D
Weighted Avg.    0.929    0.046    0.938      0.929    0.930      0.841    0.985     0.983

=== Confusion Matrix ===

    a    b   <-- classified as
 1579  152 |    a = R
   20  660 |    b = D
```

ReliefF Dataset:

J48 decision tree: Accuracy = 0.942762, TP = 0.943, FP = 0.088, ROC = 0.942

```
Correctly Classified Instances        2273               94.2762 %
Incorrectly Classified Instances       138                5.7238 %
Kappa statistic                          0.8582
Mean absolute error                      0.0664
Root mean squared error                  0.2206
Relative absolute error                 16.4011 %
Root relative squared error             49.0302 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.962    0.107    0.958      0.962    0.960      0.858    0.942     0.947     R
                 0.893    0.038    0.903      0.893    0.898      0.858    0.942     0.878     D
Weighted Avg.    0.943    0.088    0.943      0.943    0.943      0.858    0.942     0.927

=== Confusion Matrix ===

    a    b   <-- classified as
 1666   65 |    a = R
   73  607 |    b = D
```

OneRClassifier: Accuracy = 0.948569, TP = 0.949, FP = 0.074, ROC = 0.937

```
Correctly Classified Instances        2287               94.8569 %
Incorrectly Classified Instances       124                5.1431 %
Kappa statistic                          0.8732
Mean absolute error                      0.0514
Root mean squared error                  0.2268
Relative absolute error                 12.6966 %
Root relative squared error             50.3972 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.088    0.965      0.963    0.964      0.873    0.937     0.956     R
                 0.912    0.037    0.906      0.912    0.909      0.873    0.937     0.851     D
Weighted Avg.    0.949    0.074    0.949      0.949    0.949      0.873    0.937     0.927

=== Confusion Matrix ===

    a    b   <-- classified as
 1667   64 |    a = R
   60  620 |    b = D
```

K-nearest neighbors: Accuracy = 0.941103, TP = 0.941, FP = 0.093, ROC = 0.927

```
Correctly Classified Instances        2269               94.1103 %
Incorrectly Classified Instances       142                5.8897 %
Kappa statistic                          0.8537
Mean absolute error                      0.0593
Root mean squared error                  0.2426
Relative absolute error                 14.64   %
Root relative squared error             53.9064 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.115    0.955      0.963    0.959      0.854    0.927     0.949     R
                 0.885    0.037    0.904      0.885    0.895      0.854    0.927     0.842     D
Weighted Avg.    0.941    0.093    0.941      0.941    0.941      0.854    0.927     0.919

=== Confusion Matrix ===

    a    b   <-- classified as
 1667   64 |    a = R
   78  602 |    b = D
```

Naïve Bayes: Accuracy = 0.934467, TP = 0.934, FP = 0.044, ROC = 0.988

```
Correctly Classified Instances        2253               93.4467 %
Incorrectly Classified Instances       158                6.5533 %
Kappa statistic                          0.8463
Mean absolute error                      0.0676
Root mean squared error                  0.2491
Relative absolute error                 16.6789 %
Root relative squared error             55.3672 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.920    0.029    0.988      0.920    0.953      0.852    0.988     0.995     R
                 0.971    0.080    0.827      0.971    0.893      0.852    0.988     0.970     D
Weighted Avg.    0.934    0.044    0.942      0.934    0.936      0.852    0.988     0.988

=== Confusion Matrix ===

    a    b   <-- classified as
 1593  138 |    a = R
   20  660 |    b = D
```

InfoGain Dataset:

J48 decision tree: Accuracy = 0.951058, TP = 0.951, FP = 0.061, ROC = 0.967

```
Correctly Classified Instances        2293               95.1058 %
Incorrectly Classified Instances       118                4.8942 %
Kappa statistic                          0.8804
Mean absolute error                      0.0728
Root mean squared error                  0.2033
Relative absolute error                 17.9733 %
Root relative squared error             45.1696 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.959    0.069    0.972      0.959    0.966      0.881    0.967     0.973     R
                 0.931    0.041    0.899      0.931    0.915      0.881    0.967     0.917     D
Weighted Avg.    0.951    0.061    0.952      0.951    0.951      0.881    0.967     0.957

=== Confusion Matrix ===

    a    b   <-- classified as
 1660   71 |    a = R
   47  633 |    b = D
```

OneRClassifier: Accuracy = 0.948569, TP = 0.949, FP = 0.074, ROC = 0.937

```
Correctly Classified Instances        2287               94.8569 %
Incorrectly Classified Instances       124                5.1431 %
Kappa statistic                          0.8732
Mean absolute error                      0.0514
Root mean squared error                  0.2268
Relative absolute error                 12.6966 %
Root relative squared error             50.3972 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.088    0.965      0.963    0.964      0.873    0.937     0.956     R
                 0.912    0.037    0.906      0.912    0.909      0.873    0.937     0.851     D
Weighted Avg.    0.949    0.074    0.949      0.949    0.949      0.873    0.937     0.927

=== Confusion Matrix ===

    a    b   <-- classified as
 1667   64 |    a = R
   60  620 |    b = D
```

K-nearest neighbors: Accuracy = 0.933637, TP = 0.934, FP = 0.104, ROC = 0.919

```
Correctly Classified Instances        2251               93.3637 %
Incorrectly Classified Instances       160                6.6363 %
Kappa statistic                          0.8351
Mean absolute error                      0.0668
Root mean squared error                  0.2575
Relative absolute error                 16.4813 %
Root relative squared error             57.2211 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.958    0.128    0.950      0.958    0.954      0.835    0.919     0.944     R
                 0.872    0.042    0.890      0.872    0.881      0.835    0.919     0.820     D
Weighted Avg.    0.934    0.104    0.933      0.934    0.933      0.835    0.919     0.909

=== Confusion Matrix ===

    a    b   <-- classified as
 1658   73 |    a = R
   87  593 |    b = D
```

Naïve Bayes: Accuracy = 0.92949, TP = 0.929, FP = 0.045, ROC = 0.989

```
Correctly Classified Instances        2241               92.949  %
Incorrectly Classified Instances       170                7.051  %
Kappa statistic                          0.8356
Mean absolute error                      0.0713
Root mean squared error                  0.253
Relative absolute error                 17.594   %
Root relative squared error             56.216   %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.913    0.028    0.988      0.913    0.949      0.842    0.989     0.996     R
                 0.972    0.087    0.814      0.972    0.886      0.842    0.989     0.971     D
Weighted Avg.    0.929    0.045    0.939      0.929    0.931      0.842    0.989     0.989

=== Confusion Matrix ===

    a    b   <-- classified as
 1580  151 |    a = R
   19  661 |    b = D
```

Correlation Dataset:

J48 decision trees: Accuracy = 0.944836, TP = 0.945, FP = 0.072, ROC = 0.951

```
Correctly Classified Instances        2278               94.4836 %
Incorrectly Classified Instances       133                5.5164 %
Kappa statistic                          0.8651
Mean absolute error                      0.0644
Root mean squared error                  0.2086
Relative absolute error                 15.9016 %
Root relative squared error             46.3614 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.956    0.082    0.967      0.956    0.961      0.865    0.951     0.954     R
                 0.918    0.044    0.890      0.918    0.904      0.865    0.951     0.906     D
Weighted Avg.    0.945    0.072    0.946      0.945    0.945      0.865    0.951     0.940

=== Confusion Matrix ===

    a    b   <-- classified as
 1654   77 |    a = R
   56  624 |    b = D
```

OneRClassifier: Accuracy = 0.948569, TP = 0.949, FP = 0.074, ROC = 0.937

```
Correctly Classified Instances        2287               94.8569 %
Incorrectly Classified Instances       124                5.1431 %
Kappa statistic                          0.8732
Mean absolute error                      0.0514
Root mean squared error                  0.2268
Relative absolute error                 12.6966 %
Root relative squared error             50.3972 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.088    0.965      0.963    0.964      0.873    0.937     0.956     R
                 0.912    0.037    0.906      0.912    0.909      0.873    0.937     0.851     D
Weighted Avg.    0.949    0.074    0.949      0.949    0.949      0.873    0.937     0.927

=== Confusion Matrix ===

    a    b   <-- classified as
 1667   64 |    a = R
   60  620 |    b = D
```

K-nearest neighbors: Accuracy = 0.941103, TP = 0.941, FP = 0.088, ROC = 0.930

```
Correctly Classified Instances        2269               94.1103 %
Incorrectly Classified Instances       142                5.8897 %
Kappa statistic                          0.8543
Mean absolute error                      0.0593
Root mean squared error                  0.2426
Relative absolute error                 14.64   %
Root relative squared error             53.9064 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.960    0.107    0.958      0.960    0.959      0.854    0.930     0.952     R
                 0.893    0.040    0.898      0.893    0.895      0.854    0.930     0.840     D
Weighted Avg.    0.941    0.088    0.941      0.941    0.941      0.854    0.930     0.920

=== Confusion Matrix ===

    a    b   <-- classified as
 1662   69 |    a = R
   73  607 |    b = D
```

Naïve Bayes: Accuracy = 0.929905, TP = 0.930, FP = 0.048, ROC = 0.989

```
Correctly Classified Instances        2242               92.9905 %
Incorrectly Classified Instances       169                7.0095 %
Kappa statistic                          0.8359
Mean absolute error                      0.0692
Root mean squared error                  0.2515
Relative absolute error                 17.0798 %
Root relative squared error             55.8959 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.916    0.034    0.986      0.916    0.949      0.842    0.989     0.996     R
                 0.966    0.084    0.818      0.966    0.886      0.842    0.989     0.972     D
Weighted Avg.    0.930    0.048    0.938      0.930    0.932      0.842    0.989     0.989

=== Confusion Matrix ===

    a    b   <-- classified as
 1585  146 |    a = R
   23  657 |    b = D
```

Dataset with Our Chosen Attributes:

J48 decision tree: Accuracy = 0.944836, TP = 0.945, FP = 0.078, ROC = 0.939

```
Correctly Classified Instances        2278               94.4836 %
Incorrectly Classified Instances       133                5.5164 %
Kappa statistic                          0.8642
Mean absolute error                      0.0668
Root mean squared error                  0.2154
Relative absolute error                 16.5    %
Root relative squared error             47.8779 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.960    0.093    0.963      0.960    0.962      0.864    0.939     0.941     R
                 0.907    0.040    0.898      0.907    0.903      0.864    0.939     0.893     D
Weighted Avg.    0.945    0.078    0.945      0.945    0.945      0.864    0.939     0.928

=== Confusion Matrix ===

    a    b   <-- classified as
 1661   70 |    a = R
   63  617 |    b = D
```

OneRClassifier: Accuracy = 0.948569, TP = 0.949, FP = 0.074, ROC = 0.937

```
Correctly Classified Instances        2287               94.8569 %
Incorrectly Classified Instances       124                5.1431 %
Kappa statistic                          0.8732
Mean absolute error                      0.0514
Root mean squared error                  0.2268
Relative absolute error                 12.6966 %
Root relative squared error             50.3972 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.088    0.965      0.963    0.964      0.873    0.937     0.956     R
                 0.912    0.037    0.906      0.912    0.909      0.873    0.937     0.851     D
Weighted Avg.    0.949    0.074    0.949      0.949    0.949      0.873    0.937     0.927

=== Confusion Matrix ===

    a    b   <-- classified as
 1667   64 |    a = R
   60  620 |    b = D
```

K-nearest neighbors: Accuracy = 0.9382, TP = 0.938, FP = 0.099, ROC = 0.922

```
Correctly Classified Instances        2262               93.82   %
Incorrectly Classified Instances       149                6.18   %
Kappa statistic                          0.8461
Mean absolute error                      0.0622
Root mean squared error                  0.2485
Relative absolute error                 15.3561 %
Root relative squared error             55.2191 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.962    0.124    0.952      0.962    0.957      0.846    0.922     0.946     R
                 0.876    0.038    0.902      0.876    0.889      0.846    0.922     0.832     D
Weighted Avg.    0.938    0.099    0.938      0.938    0.938      0.846    0.922     0.913

=== Confusion Matrix ===

    a    b   <-- classified as
 1666   65 |    a = R
   84  596 |    b = D
```

Naïve Bayes: Accuracy = 0.92949, TP = 0.929, FP = 0.045, ROC = 0.989

```
Correctly Classified Instances        2241               92.949 %
Incorrectly Classified Instances       170                7.051 %
Kappa statistic                          0.8356
Mean absolute error                      0.0701
Root mean squared error                  0.2526
Relative absolute error                 17.3174 %
Root relative squared error             56.1243 %
Total Number of Instances             2411

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.913    0.028    0.988      0.913    0.949      0.842    0.989     0.995     R
                 0.972    0.087    0.814      0.972    0.886      0.842    0.989     0.974     D
Weighted Avg.    0.929    0.045    0.939      0.929    0.931      0.842    0.989     0.989

=== Confusion Matrix ===

    a    b   <-- classified as
 1580  151 |    a = R
   19  661 |    b = D
```

The first thing to note is that the OneRClassifier for each dataset achieved the same results. This is likely because each dataset includes all of the previous election results and thus, the OneRClassifier picked the same attribute for classification for each dataset (probably either the 2018 US Senate election or the 2016 US Presidential election).

As our model simply predicts whether a precinct is Democratic or Republican, false positives and false negatives do not have any significant meaning like they may in a field like medicine. In addition, except for Naive Bayes, which classified a large number of Republican precincts as Democratic, none of the models had a particularly high number of false positives or false negatives. Thus, we used accuracy as the primary metric to assess each model. For this reason, the best model is the one using OneR for attribute selection and the J48 decision tree for classification. This combination achieved an accuracy of 95.4791%, almost four tenths of a percent higher than the second best model, the InfoGain/J48 combination, which achieved an accuracy of 95.1048%. Beyond this, each of the 4 OneR classifier models achieved a reasonably high accuracy of 94.8569%, while the K-nearest neighbor models sat in the 93 to low 94 range for accuracy, and Naïve Bayes brought up the rear with accuracies in the 92s and low 93s. From an attribute selection standpoint, OneR attribute selector was the best, although none was consistently superior.

Discussion and Conclusion

The first and most obvious takeaway is that a combination of OneR attribute selector with a J48 decision tree produced the best results. However, there are a number of other important conclusions we can draw from our report.

First, it is evident that past electoral results are superb predictors of future results. The 9 past elections were ranked as the top attributes by each of our four attribute selection techniques, indicating that these results are better predictors of future results than any other data we tested.

In addition, we noticed that the most predictive racial attributes were the White and Black populations expressed as a percentage of the population. While the proportions of other racial/ethnic groups had some correlation with electoral results (Hispanics, for example), the attribute selection techniques consistently picked measures of both the White and Black populations.

We did not, however, observe a meaningful difference between Voting Age and Total populations for each group, indicating that these metrics can be used interchangeably in future models. There were some other surprises as well: the land area of a precinct ended up being a reasonably good predictor of partisan lean, while the total housing was not. Future investigators can build upon these findings when creating models.

Each of the models ran quite quickly, with none exceeding 0.1 seconds in runtime. In the future, this study could be improved upon by considering more classifiers and more attribute selection methods. Furthermore, a superior K-fold value could be discovered to potentially improve accuracy.

Division of Responsibilities between Group Members

Deven

- Found datasets on Redistricting Data Hub and interpreted attributes
- Created script for train/test/validation split
- Deleted empty columns and precincts with missing values/tied precincts
- Performed min/max and decimal scaling normalization to make precinct dataset completely normalized

Vibby

- Ran attribute selection algorithms in WEKA
- Created new datasets for each attribute selection technique by deleting unnecessary columns
- Performed the train/test/validation split for each dataset (although we ended up using k-fold cross-validation instead)
- Built and tested all 20 models using WEKA

Both

- Worked together to write this lab report
- Discussed, both at home and in class, each person's personal responsibilities to ensure that we understood all parts of the project

Links to Sources

1. https://redistrictingdatahub.org/dataset/virginia-block-county-and-vtd-pl-94-171-2020-official
2. https://www.virginiaredistricting.org/PageReader.aspx?page=2020DataDownload
3. https://docs.google.com/presentation/d/1C_lwPZ02HMdpuLqNiqIkTB4prtIW79b7LzonvXq731Y/edit#slide=id.g2fea41556c7_1_0
4. https://docs.google.com/presentation/d/1dsZEHmHqs0FmfawpmHd8xc15hyQLaaWrE6mB61p7ilA/edit#slide=id.g2f4ed5cce61_0_0