MACHINE LEARNING PROJECT
Naive - Bayes Classifier

PES1201700003 → Ashwin R Bharadwaj PES1201701566 → Hemanth C

Abstract:

Create a Naive Bayes Classifier for given House_votes_data, The task is to predict whether the given voter is a republican or democrat. Also calculate the Accuracy, Precision, F-score of the Model

About the Dataset:

The given dataset contains 17 columns, the first 16 columns depicting whether a voter has supported a cause or not, and the 17th column depicting which category the voter belongs to. The Dataset has 435 rows, with 267 Democrats and 168 Republicans, the dataset is not clean.

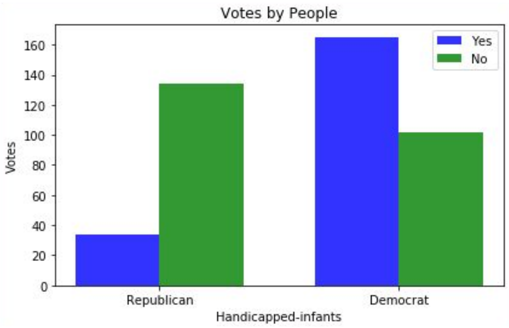
Cleaning the Dataset:

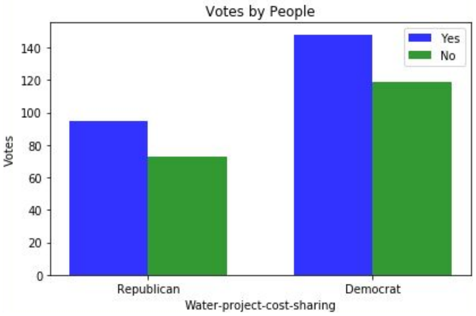
There are many ways appropriate to clean the dataset. Few of them are:

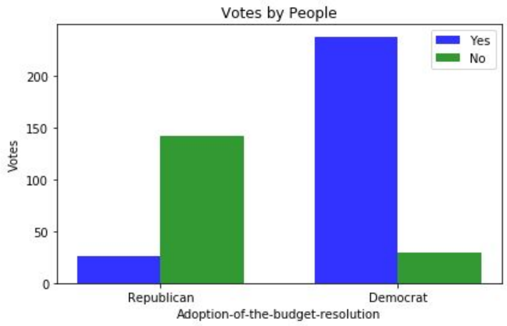
- Fill in the missing values with the mode of the attribute of that column
- Fill in the missing values depending on the type of voter he/she is. So if the value missing for an attribute is regarding immigration and the type of Voter is Republican, then fill in that value based on the mode of the attribute for republicans.
- KNN (using K Nearest Neighbours)

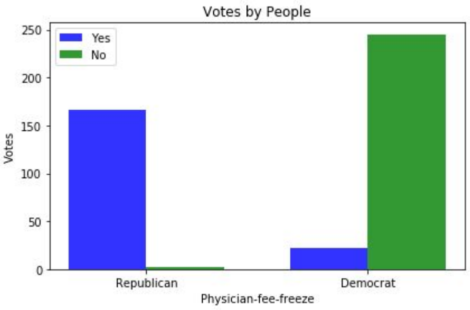
Visualization:

Visualization of the type of vote cast by each type of voter for some of the variable.









It can be observed that there is skewness in each decision type, hence this can be used for data-filling

Navie Bayes:

Training the Algorithm:

We split the dataset in the ratio of 80:20 (80% of the data for training and 20% of the data for testing). We do this 5 times, to perform 5-fold testing on the cleaned dataset, and take the average values for Accuracy, Precision, and F-score. Every time the dataset is trained, print the accuracy and the time taken to predict the values

Working of the Algorithm:

As mentioned above we are using Naive-Bayes for the classification, as we know each attribute is Assumed to be independent of the other attributes of the given data-row, as shown in the diagram Below.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

We use Bayes Therom of trying to find the probability of the given attribute.

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

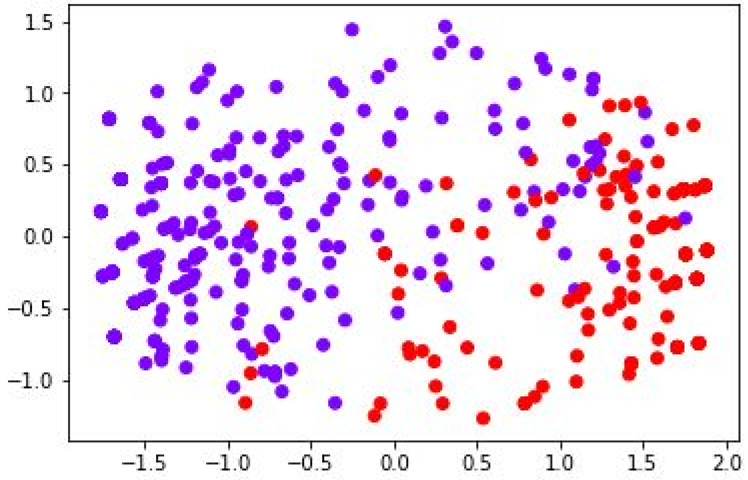
Accuracy of the model → as taken as the average of the 5 fold method is 91.53% on average
Precision of the model → 0.9774
Recall of the model → 0.7514
F-Score of the model → 0.8497

Simple KNN:

Algorithm:

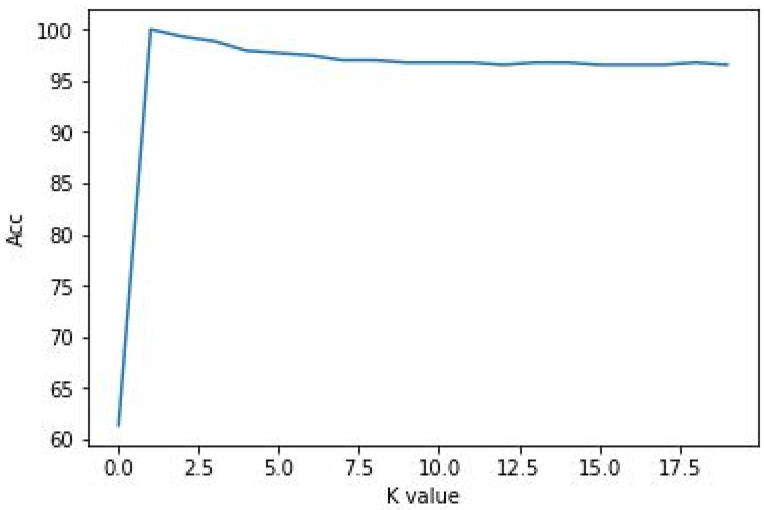
All the values in the data set are binary and so we have converted into numeric value (0,1). In this sample, we test all the data points. Starting with a k value of 5 all the way up to 10. For each of the values of K, the accuracy has been printed.

Visualization:



→ Blue is Democrat

→ Red is Republican



Variation of acc wrt to K value