Predicting Presidency impact on work Visa

Content

- Problem Statement
- About the Dataset
- Data Pre-processing
- Algorithm used
- Prediction
- Analysis
- Result

Problem Statement

How Work Visa approval and denial Rates has been affected by Presidency in US from 2016 to 2017.

Dataset

The dataset we selected is from kaggle website :

a) Source: Office of Foreign Labor Certification, U.S. Department of Labor Employment and Training Administration

b) List Link: https://www.foreignlaborcert.doleta.gov/performancedata.cfm

c) Dataset Type: Record – Transaction Data

d) Number of Attributes: 27

e) Number of Instances: 528,147

Data Preprocessing

1. Dividing the dataset into 2016 and 2017 timeline.

2. Dealing with missing values.

    a. Omitting Missing values.

    b. Replacing by Mode for factors.

    c. Replacing by mean for continuous variable.

3. Dealing with unbalanced dataset.

    a. The dataset was not a balanced dataset the output with output as certified are 95% whereas the one with denied are 5%.

    b. The observation for visa approved were far more than the denied.

    c. Hence in order for our models to work better we used smote package which deals with undersampling and oversampling and use cross validation

Algorithm used

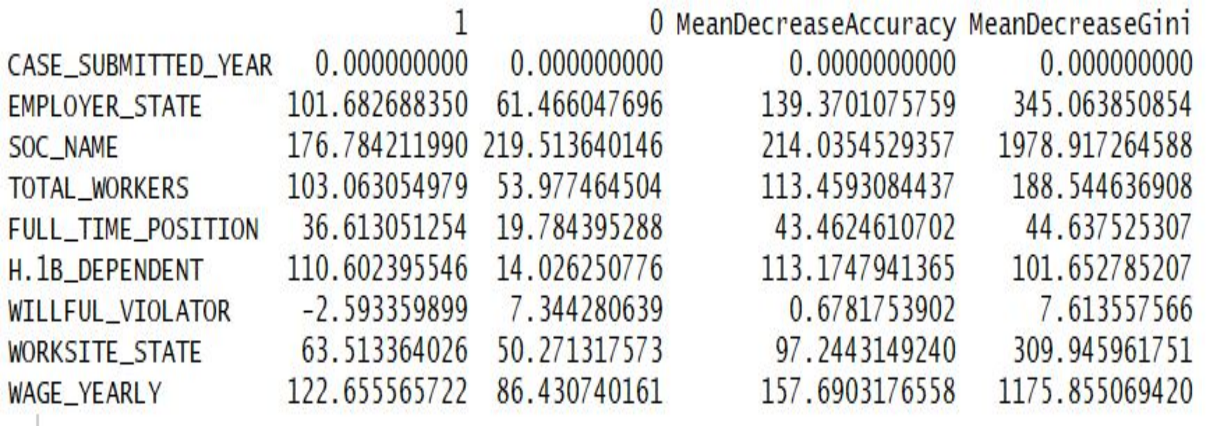
The ALgorithm used are:

1. Random Forest(RF)
2. Neural Network(ANN)
3. Naive Bayes
4. CART
5. KNN

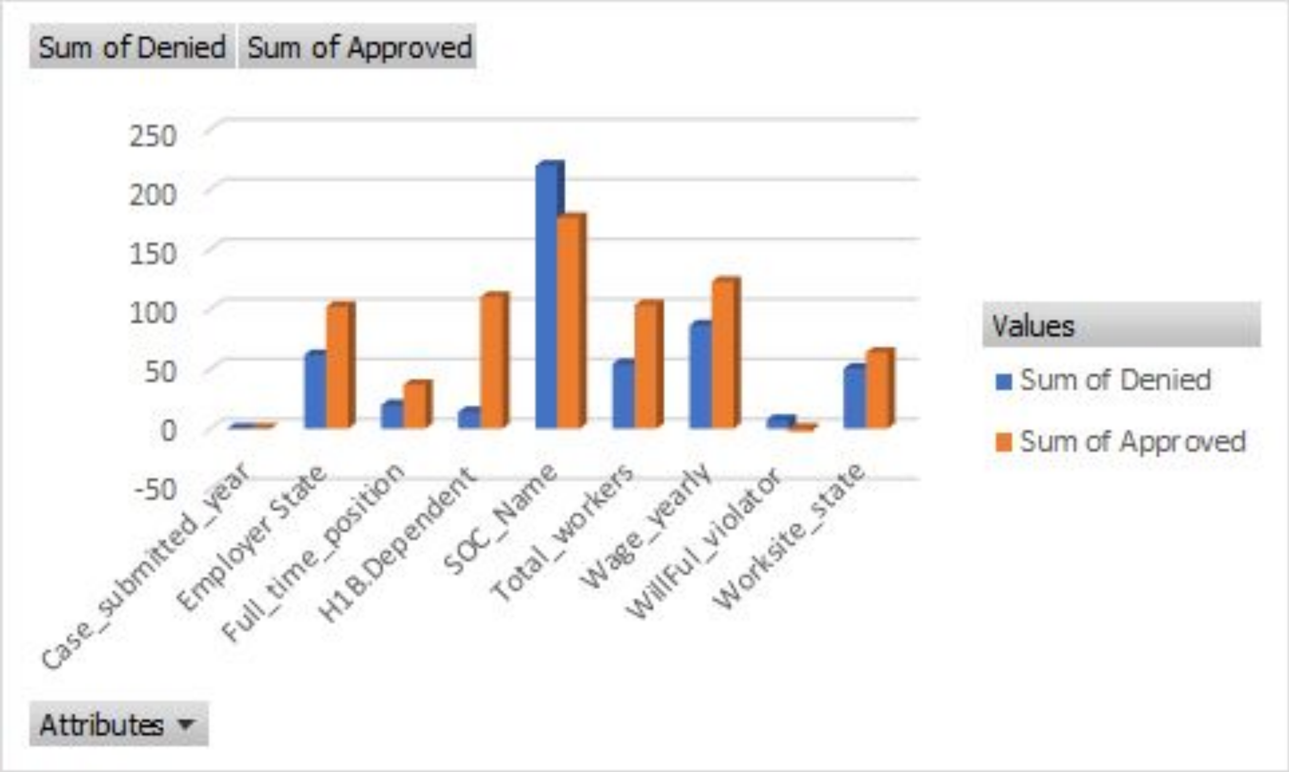
Random Forest

Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction

Importance Of Each Variable



```
                                1             0 MeanDecreaseAccuracy MeanDecreaseGini
CASE_SUBMITTED_YEAR   0.000000000   0.000000000         0.0000000000      0.000000000
EMPLOYER_STATE      101.682688350  61.466047696       139.3701075759    345.063850854
SOC_NAME            176.784211990 219.513640146       214.0354529357   1978.917264588
TOTAL_WORKERS       103.063054979  53.977464504       113.4593084437    188.544636908
FULL_TIME_POSITION   36.613051254  19.784395288        43.4624610702     44.637525307
H.1B_DEPENDENT      110.602395546  14.026250776       113.1747941365    101.652785207
WILLFUL_VIOLATOR     -2.593359899   7.344280639         0.6781753902      7.613557566
WORKSITE_STATE       63.513364026  50.271317573        97.2443149240    309.945961751
WAGE_YEARLY         122.655565722  86.430740161       157.6903176558   1175.855069420
```



Random Forest Advantages and Disadvantages

Advantages

Variable importance

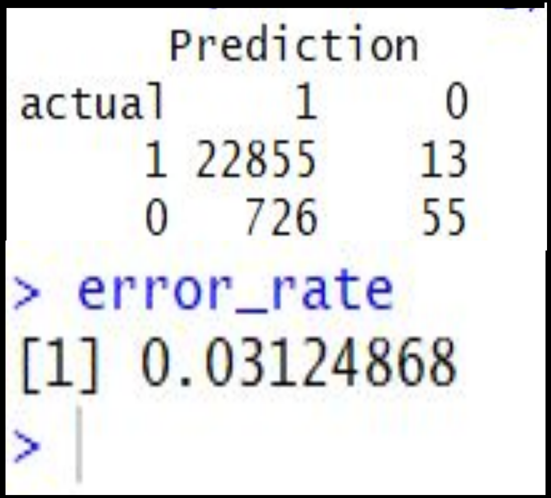
Converts weak classifiers into a good one.

Disadvantages

Can not handle factor with more than 53 levels

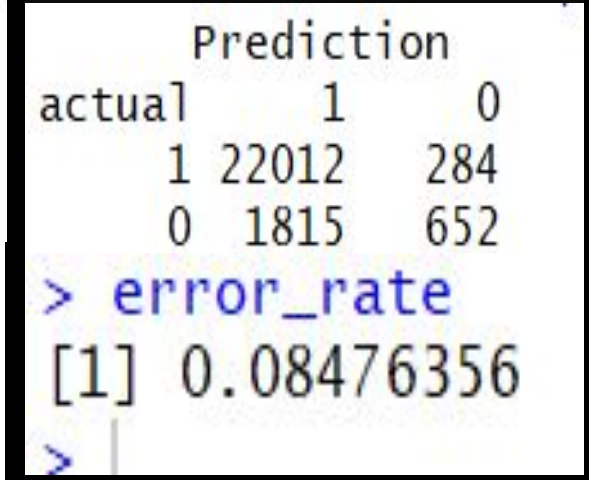
Output

H1B_2016 Confusion Matrix



```
      Prediction
actual     1     0
     1 22855    13
     0   726    55
> error_rate
[1] 0.03124868
> 
```

H1B_2017 Confusion Matrix

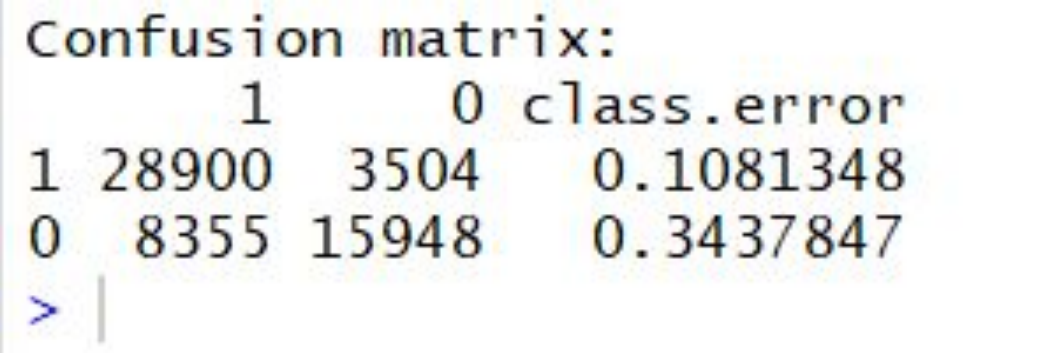


```
      Prediction
actual     1     0
     1 22012   284
     0  1815   652
> error_rate
[1] 0.08476356
> 
```

Dealing with Unbalanced dataset

We have used Synthetic Minority Over-sampling Technique to overcome this unbalance and the result were better as compared to previous ones

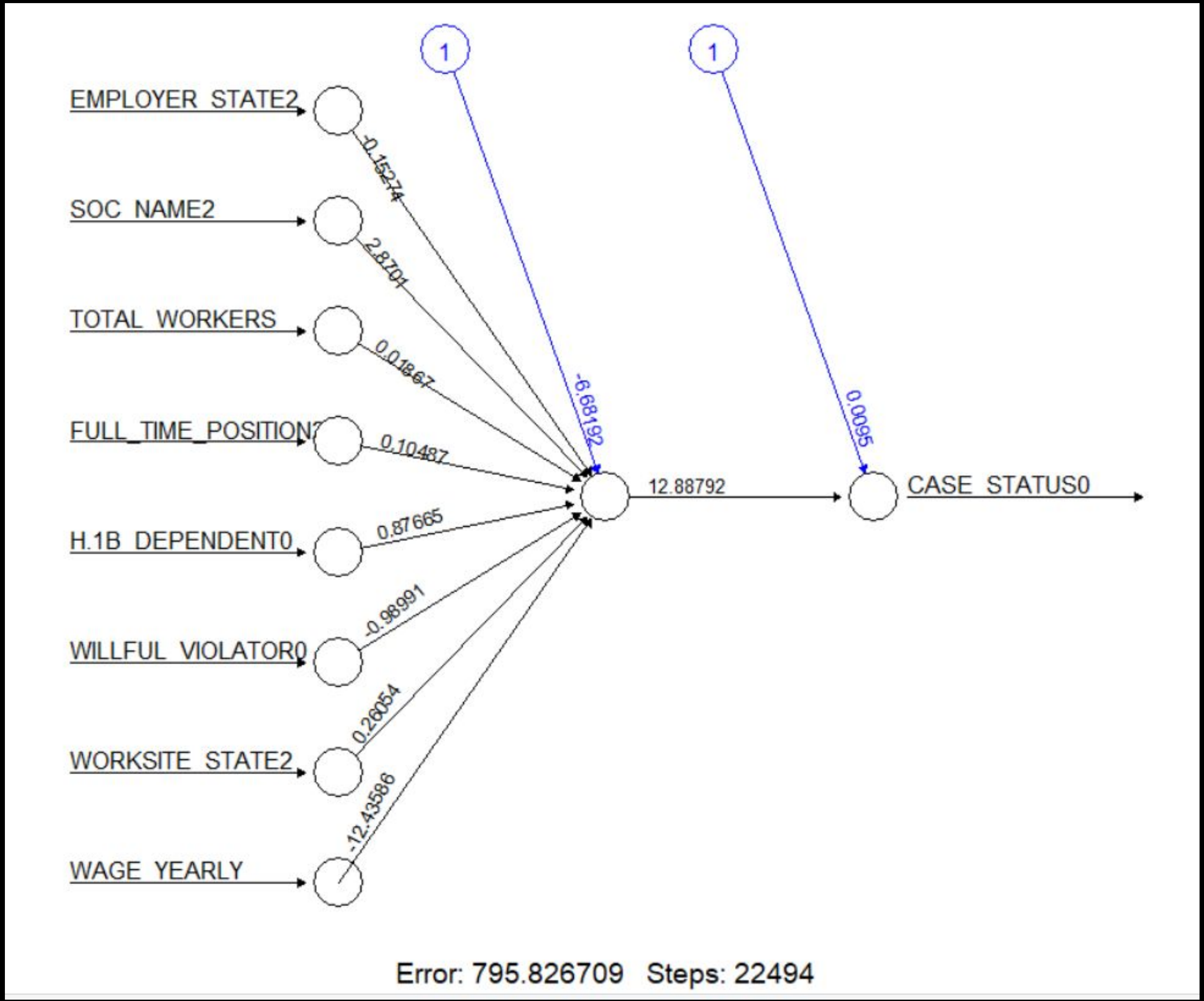
H1B_2017 Confusion Matrix



```
Confusion matrix:
      1     0 class.error
1 28900  3504   0.1081348
0  8355 15948   0.3437847
> 
```

Neural Net(ANN)

H1B_2016



Advantages

- Great for complex/abstract problems like image recognition.

Disadvantages

- Requires a shit load of training and cases
- Black box that not much can be gleaned from

Naive Bayes Algorithm

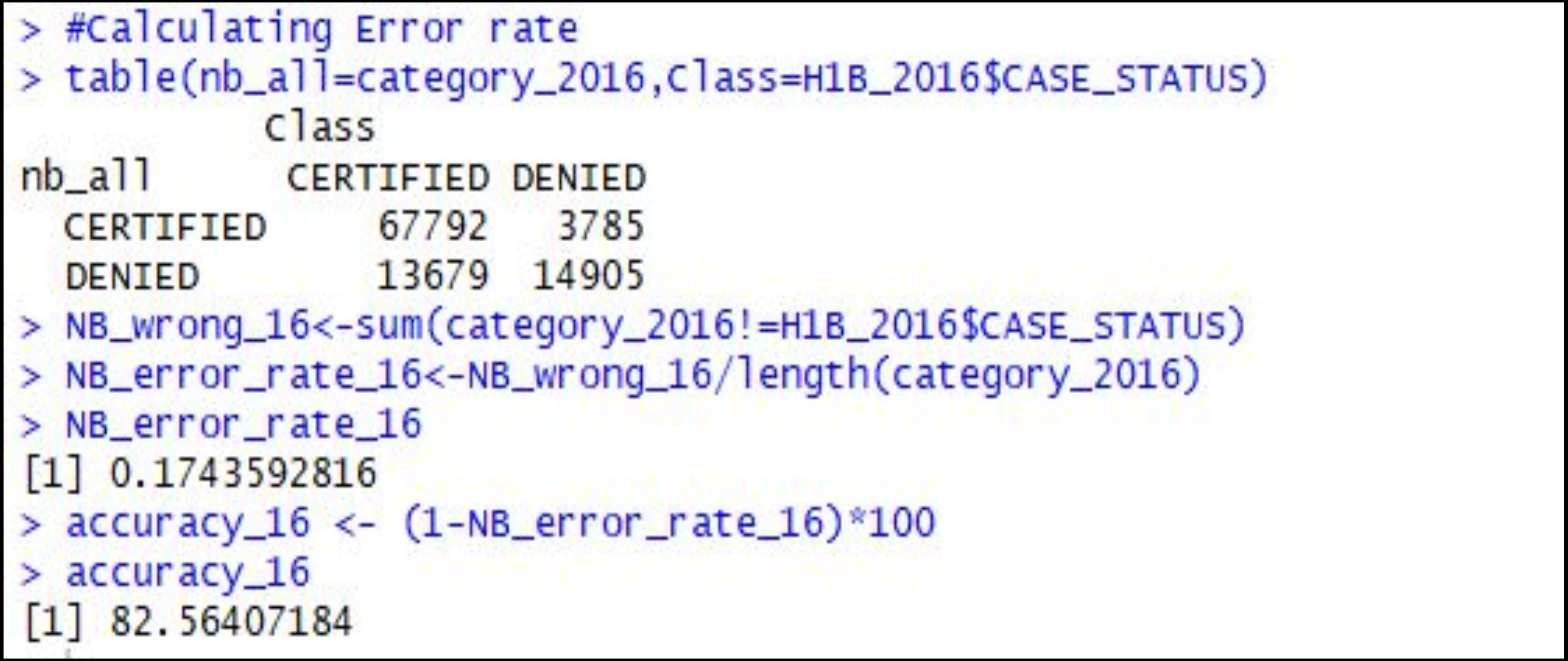
Advantages

- Fast
- Can make probabilistic predictions

Disadvantages

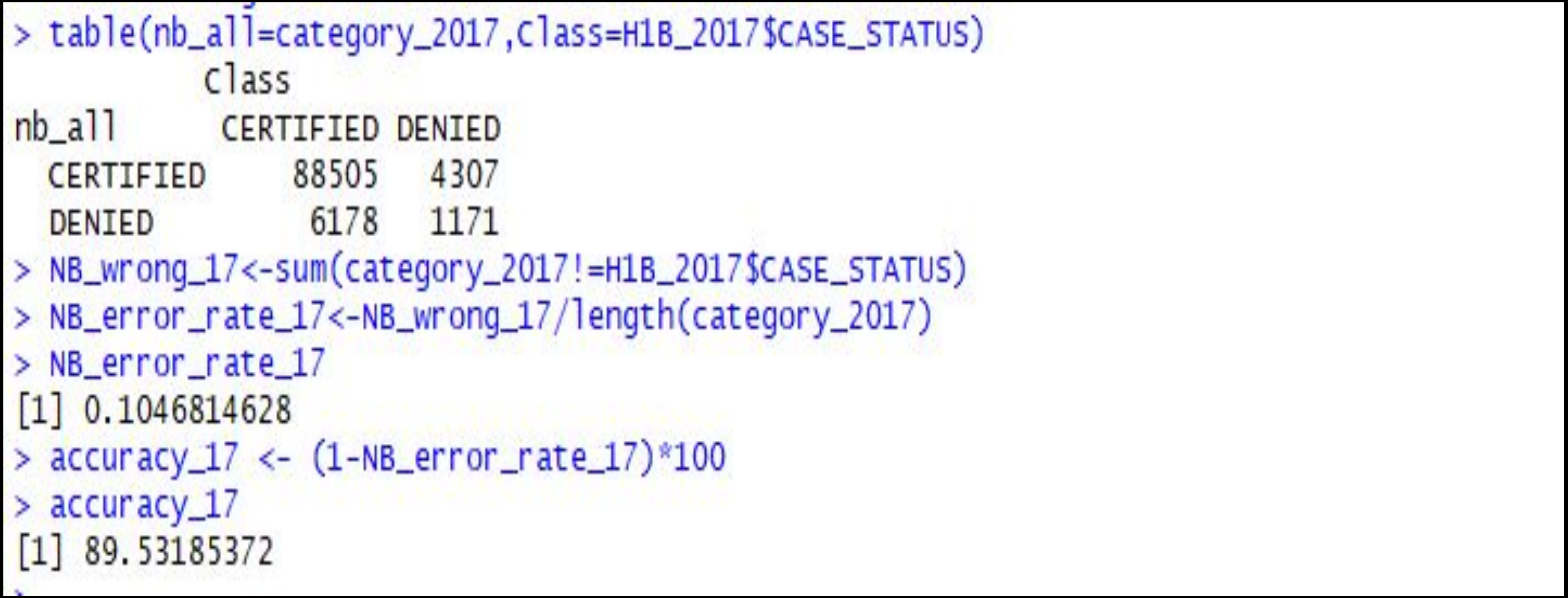
- Initialization is a bit time consuming

Dataset Result of year 2016



```
> #Calculating Error rate
> table(nb_all=category_2016,Class=H1B_2016$CASE_STATUS)
           Class
nb_all      CERTIFIED DENIED
  CERTIFIED     67792   3785
  DENIED        13679  14905
> NB_wrong_16<-sum(category_2016!=H1B_2016$CASE_STATUS)
> NB_error_rate_16<-NB_wrong_16/length(category_2016)
> NB_error_rate_16
[1] 0.1743592816
> accuracy_16 <- (1-NB_error_rate_16)*100
> accuracy_16
[1] 82.56407184
```

Dataset Result of year 2017



```
> table(nb_all=category_2017,Class=H1B_2017$CASE_STATUS)
           Class
nb_all      CERTIFIED DENIED
  CERTIFIED     88505   4307
  DENIED         6178   1171
> NB_wrong_17<-sum(category_2017!=H1B_2017$CASE_STATUS)
> NB_error_rate_17<-NB_wrong_17/length(category_2017)
> NB_error_rate_17
[1] 0.1046814628
> accuracy_17 <- (1-NB_error_rate_17)*100
> accuracy_17
[1] 89.53185372
```

CART

Advantages

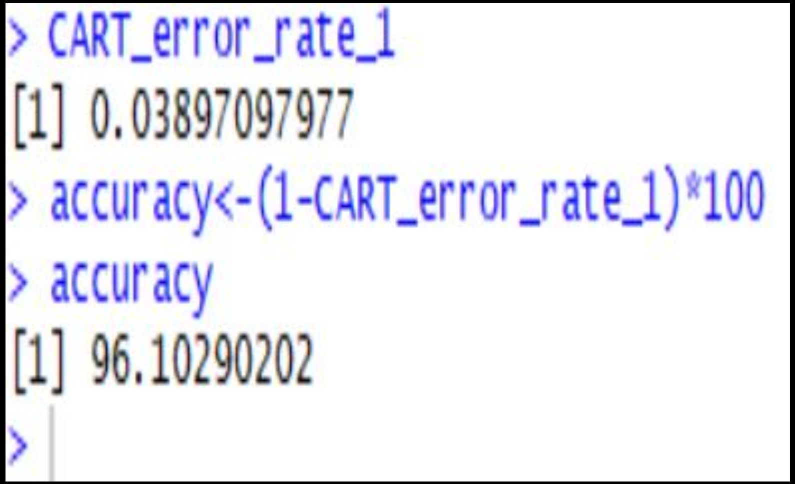
1.Automatically performs variable selection

2.Uses any combination of continuous/discrete variable.

Disadvantages

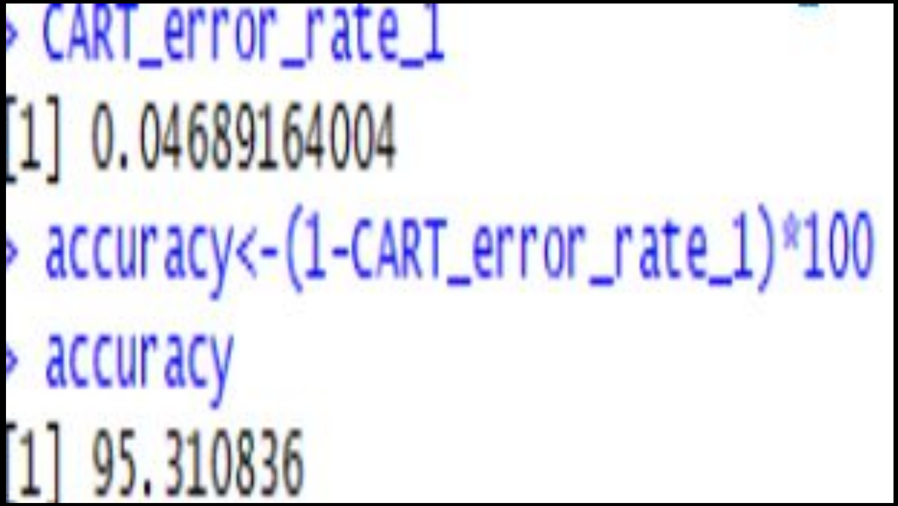
Instability of model structure

Dataset Result of year 2017



```
> CART_error_rate_1
[1] 0.03897097977
> accuracy<-(1-CART_error_rate_1)*100
> accuracy
[1] 96.10290202
> 
```

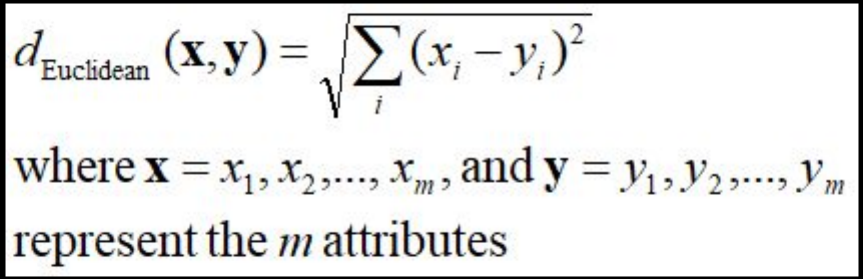
Dataset Result of year 2016



```
> CART_error_rate_1
[1] 0.04689164004
> accuracy<-(1-CART_error_rate_1)*100
> accuracy
[1] 95.310836
```

K-NN

- K-NN uses distance function to calculate the distance between points from the center



$$d_{\text{Euclidean}}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

where $\mathbf{x} = x_1, x_2, \ldots, x_m$, and $\mathbf{y} = y_1, y_2, \ldots, y_m$ represent the m attributes

Analysis

- The comparisons used to classify are as follows:
  - Case Status Vs. Region (Employer State)
  - Case Status Vs. Department (Soc Name)
  - Case Status Vs. Wage (Yearly)

KNN Advantages and Disadvantages

Advantages

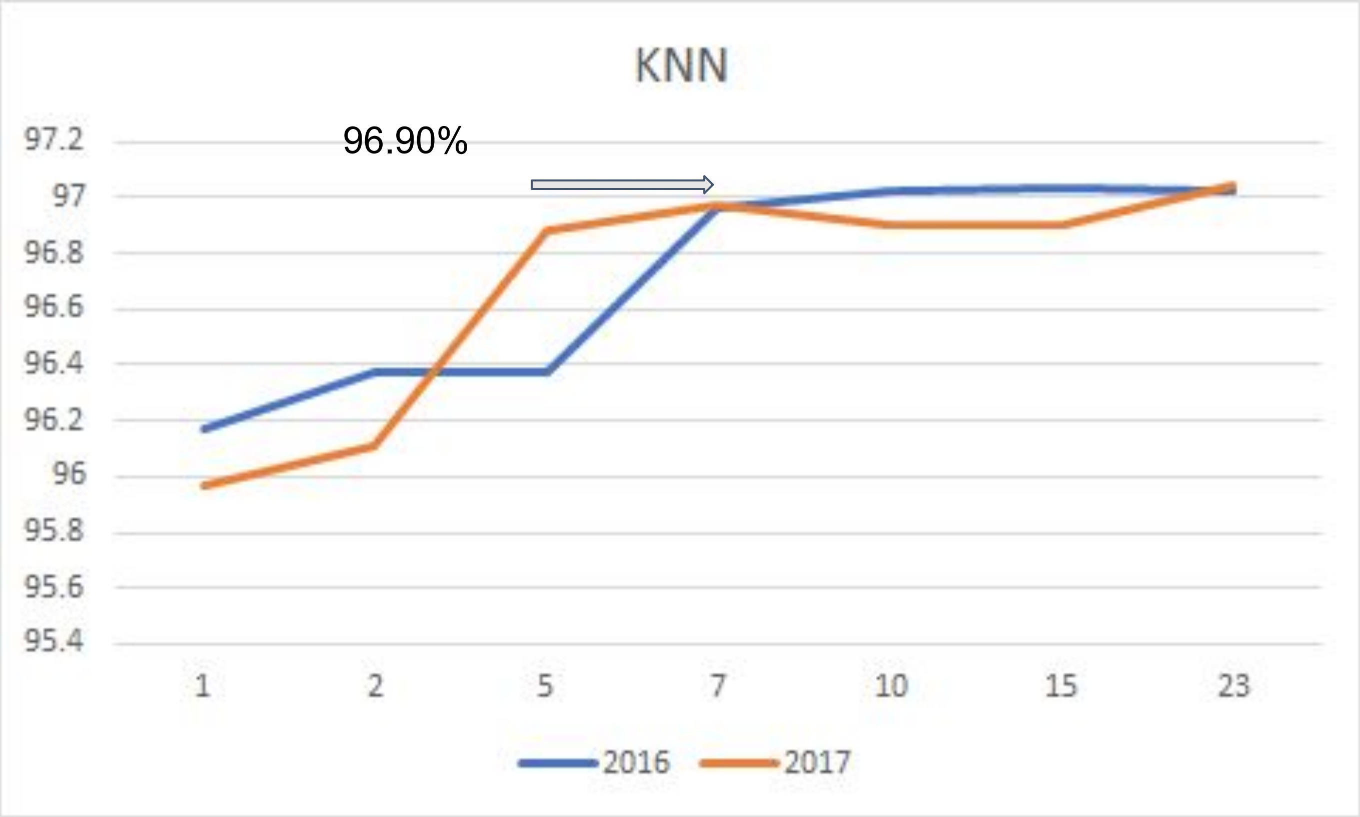
- Cost of learning process is zero
- Effective if training data is large

Disadvantages

- KNN algorithm is lazy learner
- Need to determine value of parameter K

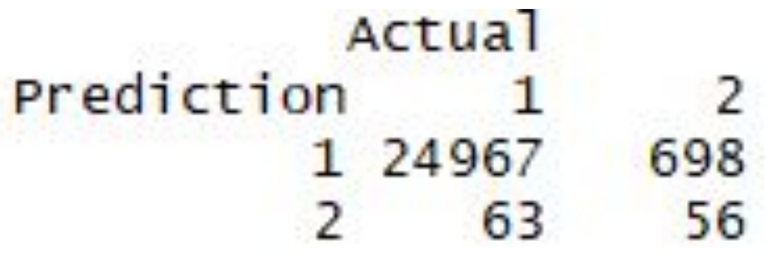
KNN Algorithm

Accuracy of KNN using various K



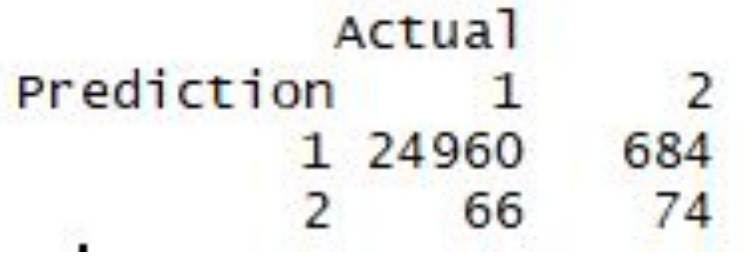
96.90%

H1B_2016 Confusion Matrix



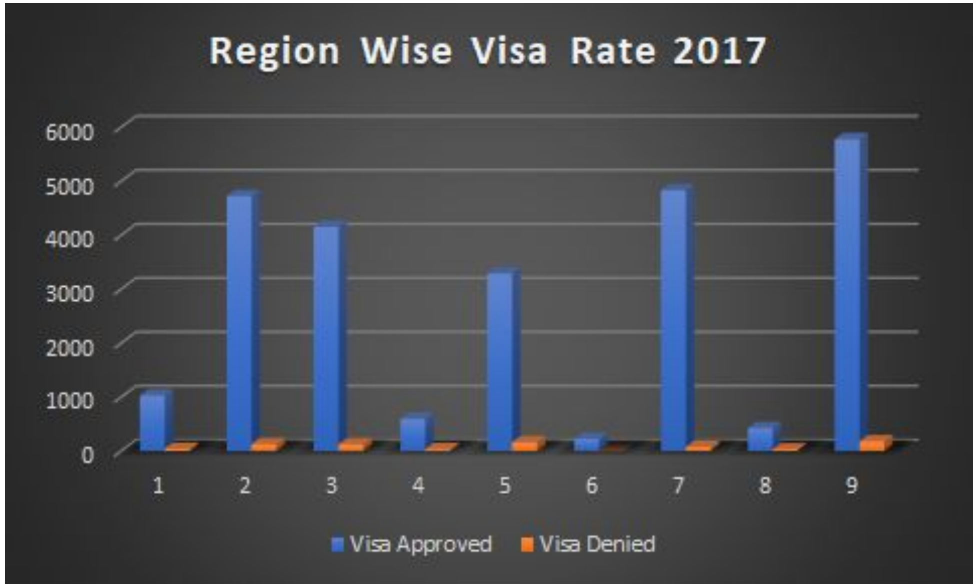
```
          Actual
Prediction     1     2
         1 24967   698
         2    63    56
```

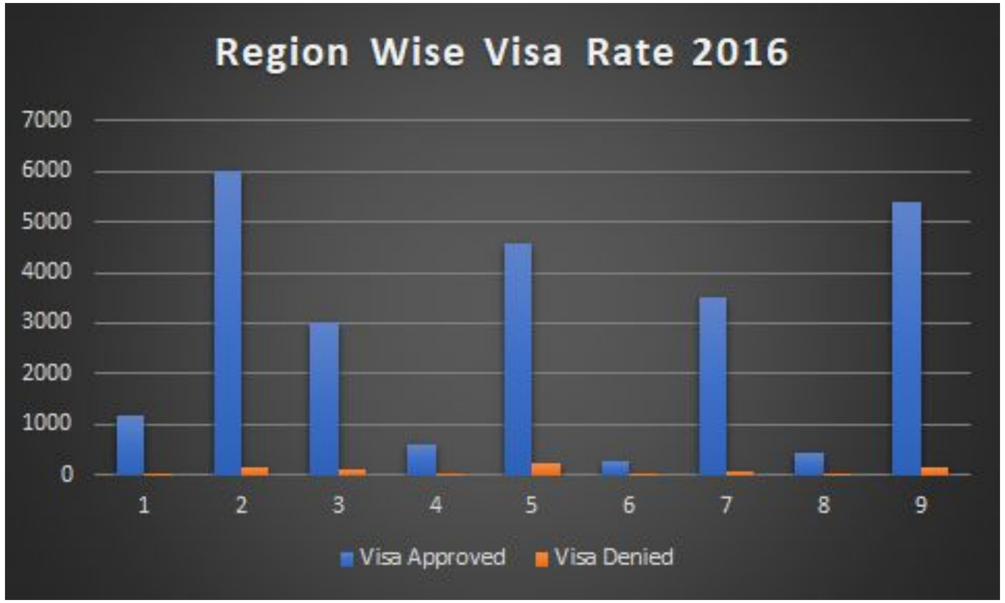
H1B_2017 Confusion Matrix

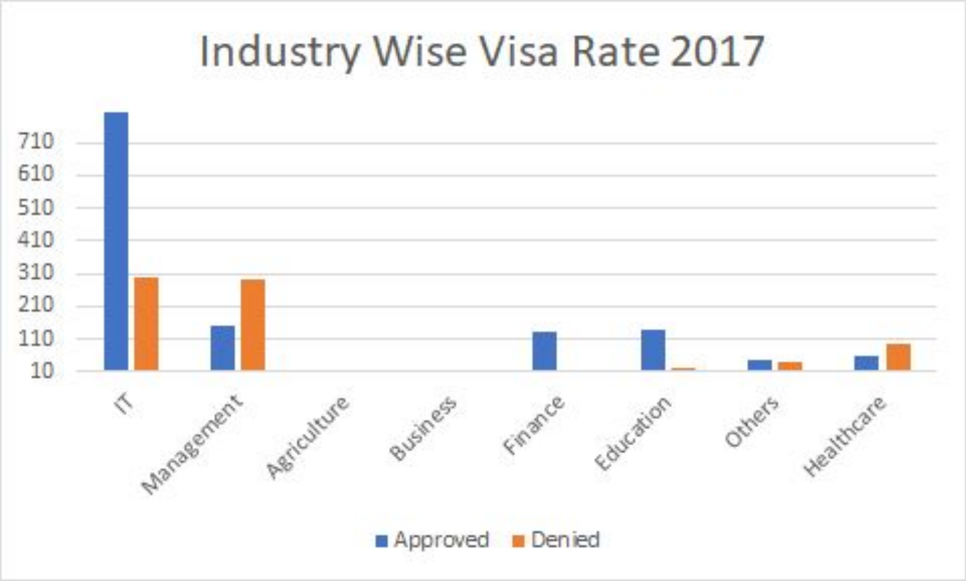


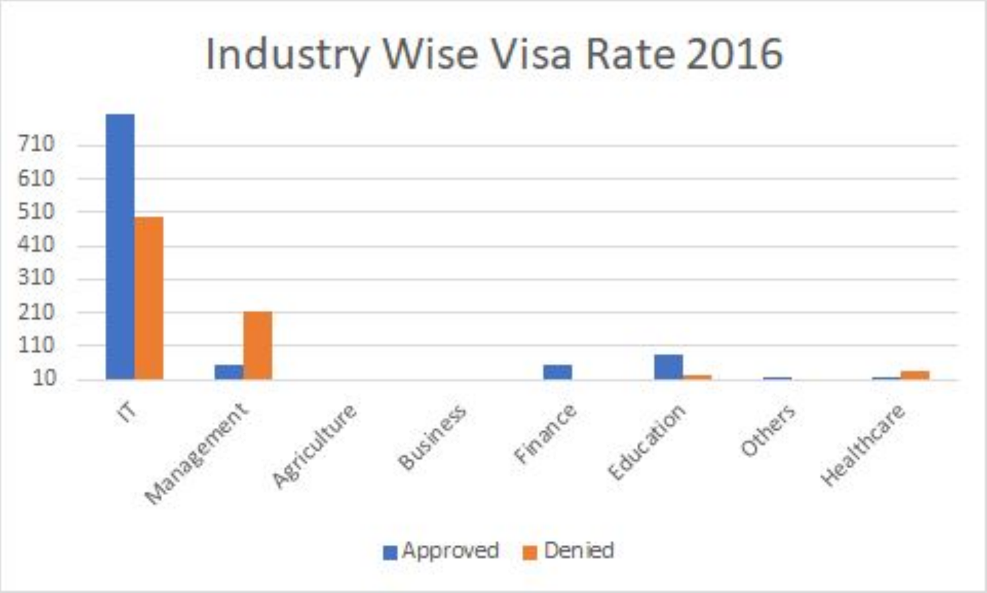
```
          Actual
Prediction     1     2
         1 24960   684
         2    66    74
.
```

Analysis Using KNN

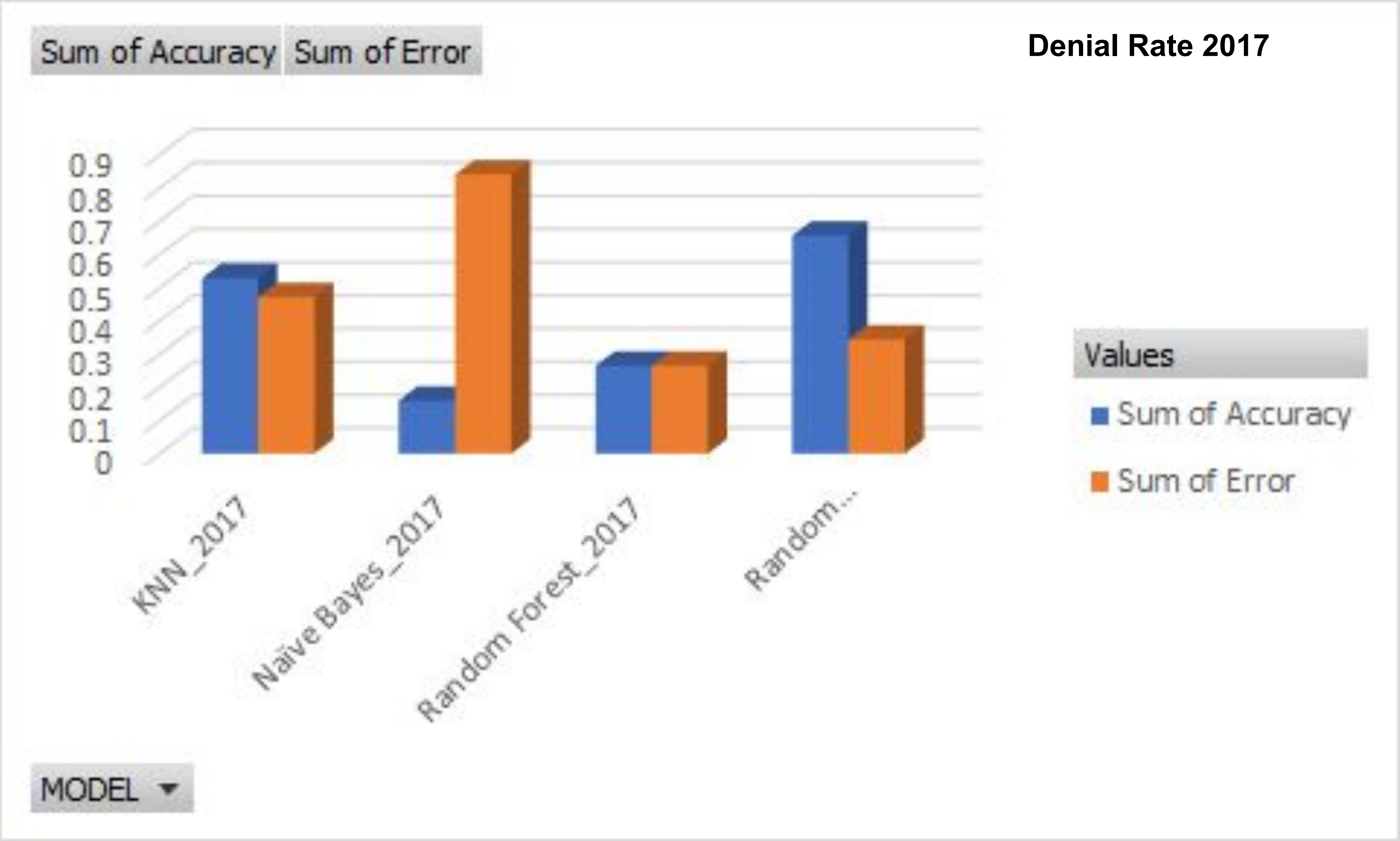








Comparing Algorithm Used



Dataset Results

The Visa Denial rate have increase according to our analysis and the same have indicated by the USCIS

U.S. Citizenship and Immigration Services

	2016	2017
RECEIPTS	399,349	404,087
APPROVALS	348,162	298,445
DENIALS	51,187	105,642

References

1. https://www.kaggle.com/trivedicharmi/h1b-disclosure-dataset
2. https://visualstudiomagazine.com/articles/2013/07/01/neural-network-data-normalization-and-encoding.aspx
3. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
4. https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/
5. https://www.geeksforgeeks.org/naive-bayes-classifiers/