**Project 2 Report:**

**Data mining with Weka**

Allen Martillano

[martillanoallen@gmail.com](mailto:martillanoallen@gmail.com)

Abstract – In this report, the purpose of this project is to perform data mining tasks using a collection of machine learning algorithms from WEKA. Furthermore, using a real-world dataset from the 1984 United States congressional voting records database. As it pertains to the file; vote\_arff.arff.

Tasks

- provide a context of use of the dataset by providing the classification task, and dataset characteristics.

- Use exploratory data analysis to examine the categorical variables, and the target class variable.

- Then we compared evaluation protocols for classification running a 2-fold cross validation, and 10-fold cross validation, as well as reporting the performance.

- Show the list of attributes by order of importance.

- 3 runs using the 10–fold cross validation algorithm and different number of attributes.

- Decision tree visualization, and classification of testing instances.

Table of Contents

[Data analytics Task and dataset characteristics 2](#_Toc131451000)

[Classification task 2](#_Toc131451001)

[Dataset characteristics 2](#_Toc131451002)

[Exploratory Data Analysis 3](#_Toc131451003)

[Categorical variables 3](#_Toc131451004)

[Numerical variables 4](#_Toc131451005)

[*Target class variable* 4](#_Toc131451006)

[Data visualization 5](#_Toc131451007)

[Comparing evaluation protocols for classification 7](#_Toc131451008)

[2-Fold cross validation 7](#_Toc131451009)

[10-fold cross validation 9](#_Toc131451010)

[Comparing 2-fold cross validation and 10-fold cross validation 10](#_Toc131451011)

[Feature selection 11](#_Toc131451012)

[Ranking attributes 11](#_Toc131451013)

[Show the list of the attributes by order of importance (based on average merit) 11](#_Toc131451014)

[Classification using decision tree 12](#_Toc131451015)

[Classification using all 16 attributes and 1 Class 12](#_Toc131451016)

[Classification using best 3 attributes and 1 Class 13](#_Toc131451017)

[Classification using best 5 attributes and 1 Class 14](#_Toc131451018)

[Comparison 15](#_Toc131451019)

[Decision tree visualization and testing 16](#_Toc131451020)

[Decision tree model visualization 16](#_Toc131451021)

[Classification of testing instances 17](#_Toc131451022)

# Data analytics Task and dataset characteristics

## Classification task

*Since the UCI machine learning repository provides a task description like regression, classification, and clustering. We need to first provide a context of use (description of the classification task) of the dataset if the dataset publisher does not make this information available.*

The vote dataset represents the number of votes each one of the 16 bills received in favour (“yes”) and against (“no”) by the 435 members (267 Democrats and 168 Republicans) of the House of Representatives in the U.S.

The dataset gives the following information:

|  |  |  |
| --- | --- | --- |
| **Category** | **Qty** | **Description** |
| Instances (rows) | 435 | Each instance represents one vote for each Democrat seat and each Republican seat in the House of Representatives. |
| Division of seats: | 267 | Democrat votes |
|  | 168 | Republican votes |
|  |  |  |
| Attributes | 17 | 16 bills that were voted on and 1 Class (democrat, republican) |

## Dataset characteristics

*The number of instances:*

435 (267 Democrat + 168 Republican)

*The list of variable names and their types in a table:*

|  |  |  |
| --- | --- | --- |
| Variable Names | Type |  |
| 1. Handicapped infants | Nominal | 1. No, (2) Yes |
| 1. Water project cost sharing | Nominal | 1. No, (2) Yes |
| 1. Adoption of the budget resolution | Nominal | 1. No, (2) Yes |
| 1. Physician fee freeze | Nominal | 1. No, (2) Yes |
| 1. El Salvador aid | Nominal | 1. No, (2) Yes |
| 1. Religious groups in schools | Nominal | 1. No, (2) Yes |
| 1. Anti-Satellite test ban | Nominal | 1. No, (2) Yes |
| 1. Aid to Nicaraguan contras | Nominal | 1. No, (2) Yes |
| 1. Mx-missile | Nominal | 1. No, (2) Yes |
| 1. Immigration | Nominal | 1. No, (2) Yes |
| 1. Synfuels corporation cutback | Nominal | 1. No, (2) Yes |
| 1. Education spending | Nominal | 1. No, (2) Yes |
| 1. Superfund right-to-use | Nominal | 1. No, (2) Yes |
| 1. Crime | Nominal | 1. No, (2) Yes |
| 1. Duty-free exports | Nominal | 1. No, (2) Yes |
| 1. Export administration act South Africa | Nominal | 1. No, (2) Yes |
| 1. Class (dependent variable/target) | Nominal | 1. Democrats, (2) Republicans |

*the dependent variable/target variable and the independent variable.*

Dependent variable/target: Class

Independent variable:

# Exploratory Data Analysis

## Categorical variables

*Examining the categorical variables and interpreting the frequency tables. Frequency tables are displayed once you select the attribute within the preprocess tab in Weka.*

Each categorical variable represents 1 bill. The frequency table corresponding to each categorical variable (each bill) shows the number of votes against (“no”) and the number of votes in favour (“yes”) that each bill received. With this information, we can determine the relative frequency of the number of “no” and “yes” votes for each bill in relation to the total number of possible votes (435 total possible votes, 267 Democrat and 168 Republican) as well as the relative frequency of the total number of votes each bill received in relation to the total number of possible votes (435 total possible votes, 267 Democrat and 168 Republican). This also provides insight on the number/percentage of abstentions.

For example:

* The Handicapped-infants bill received 187 votes in favour (which represents 43% of the total possible votes, 435) and 236 votes against (which represents 54% of the total possible votes,435). The total number of votes the bill received (both in favour and against) was 423 (187+236) (which represents 97% of the total possible votes, 435). There were 12 abstentions.
* The Water-project-cost-sharing bill received 195 votes in favour (which represents 45% of the total possible votes, 435) and 192 votes against (which represents 44% of the total possible votes, 435). The total number of votes the bill received (both in favour and against) was 387 (195+192) (which represents 89% of the total possible votes, 435). There were 48 abstentions.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Categorical Variables | Yes | Relative Freq. of 435 | No | Relative Freq. of 435 | Total Votes | Relative Freq. of 435 |
| 1. Handicapped infants | 187 | 43% | 236 | 54% | 423 | 97% |
| 1. Water project cost sharing | 195 | 45% | 192 | 44% | 387 | 89% |
| 1. Adoption of the budget resolution | 253 | 58% | 171 | 39% | 424 | 97% |
| 1. Physician fee freeze | 177 | 41% | 247 | 57% | 424 | 97% |
| 1. El Salvador aid | 212 | 49% | 208 | 48% | 420 | 96% |
| 1. Religious groups in schools | 272 |  | 152 |  | 424 | 97% |
| 1. Anti-Satellite test ban | 239 | 63% | 182 | 42% | 421 | 97% |
| 1. Aid to Nicaraguan contras | 242 | 56% | 178 | 41% | 420 | 96% |
| 1. Mx-missile | 207 | 48% | 206 | 47% | 413 | 94% |
| 1. Immigration | 216 | 50% | 212 | 49% | 428 | 98% |
| 1. Synfuels corporation cutback | 150 | 34% | 264 | 61% | 414 | 95% |
| 1. Education spending | 171 | 39% | 233 | 54% | 404 | 93% |
| 1. Superfund right-to-use | 209 | 48% | 201 | 46% | 410 | 94% |
| 1. Crime | 248 | 57% | 170 | 39% | 418 | 96% |
| 1. Duty-free exports | 174 |  | 233 |  | 407 | 93% |
| 1. Export admin. act South Africa | 269 |  | 62 |  | 331 | 76% |

## Numerical variables

*Examining the numerical variables and interpret the summary statistics obtained from Weka (min, max, mean, and standard deviation). Basic summary statistics are displayed once you select the attribute within preprocess tab.*

Although this dataset doesn’t have numerical variables.

## *Target class variable*

*Show the frequency distribution of each of the possible values. Interpret. Is the dataset balanced? Any other comment?*

The dataset is not balanced.

Democrats (blue) represent approximately 61% (267/435) of the total seats, while Republican (red) represent approximately 39% (168/435) of the total seats.



## Data visualization

*Clicking on Visualize all button, we comment on the graphs shown there. Considering at least 2 variables and interpreting its power/range of values that help discriminating between classes.*

Each cluster column chart provides a visual representation of the vote frequency (yes vs no) for each categorical variable by showing:

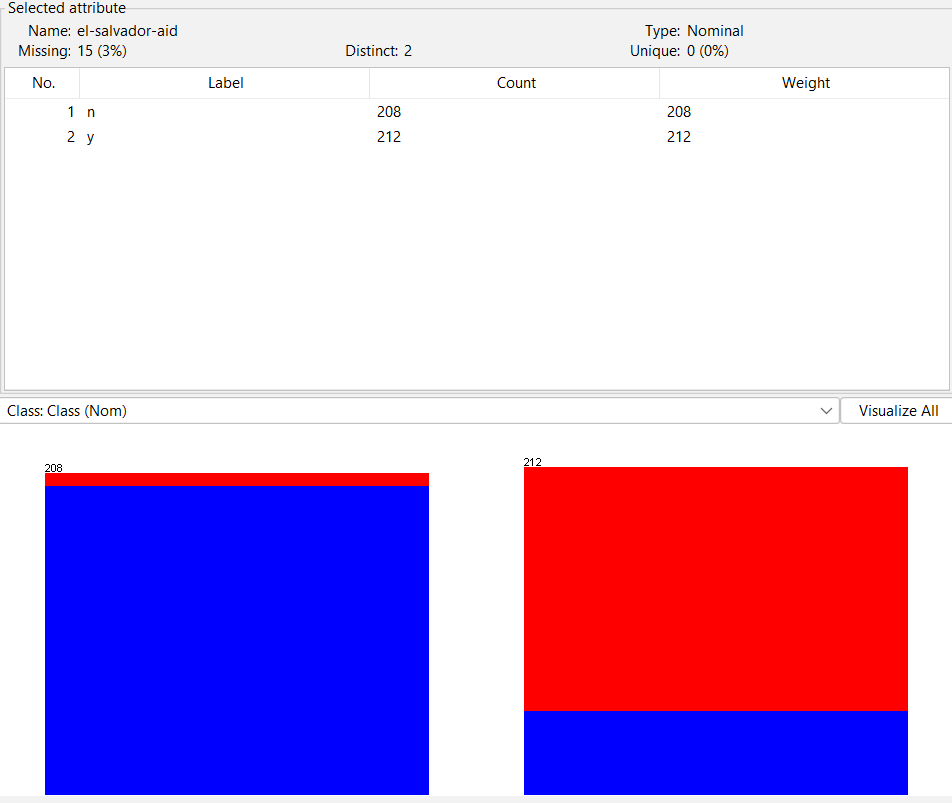
* vote quantity (left bar for “no” votes and right bar “yes” votes)
* vote attribution (blue for Democrats and red for Republicans)



The chart for the Immigration bill (variable), which shows an almost even distribution of the total votes (212 “no” votes and 216 “yes” votes) also shows an almost even distribution of Democrat (blue) and Republican (red) votes in each column. The almost even split of votes makes it difficult to predict the class label as it makes it difficult to discriminate between the number of Democrats and Republicans that voted “no” and the number of Democrats and Republicans that voted “yes”.



On the other hand, the chart for the El-Salvador-Aid bill (variable), although it shows an almost even distribution of the total votes (208 “no” votes vs. 212 “yes” votes), it makes it easy to predict the class label as it makes it easy to discriminate between Democrat (blue) and Republican (red) votes in both the “no” and the “yes” votes columns. We can clearly see that almost all “no” votes were casted by Democrats (blue) and almost all “yes” votes were casted by Republicans (red) for this bill.



Likewise, the chart for the Education-spending bill (variable) makes it easy to predict the class label as it makes it easy to discriminate between Democrat (blue) and Republican (red) votes in both the “no” and the “yes” votes columns. We can clearly see that almost all “no” votes were casted by Democrats (blue) and almost all “yes” votes were casted by Republicans (red) for this bill.



# Comparing evaluation protocols for classification

## 2-Fold cross validation

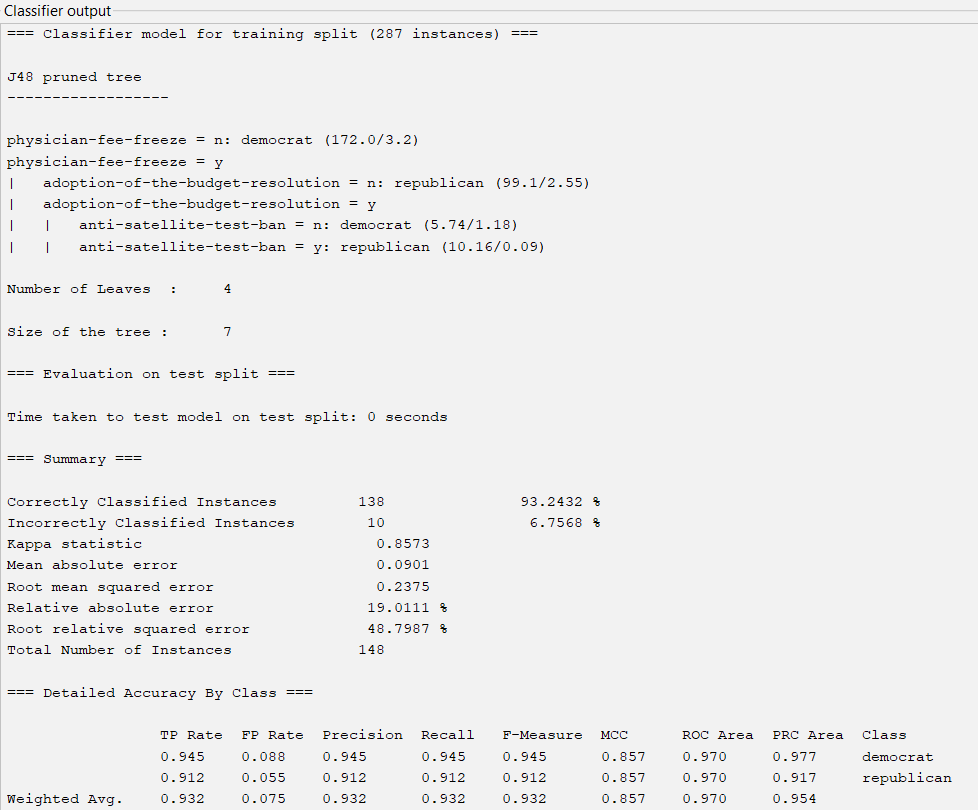
*Run 2-fold cross validation on the chosen dataset. Considering the option “percentage split: 66% training and the rest for testing”. reporting the performance of the classifier based on the following measures:*

* + *overall classification accuracy:*

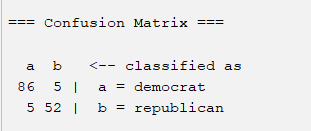
Split dataset 2-fold cross validation accuracy: **93.2432%**

* + *for each class value, we report the FP rate, TP rate, Precision, recall, F-measure.*
  + *the overall weighted average over class levels for each of the of the FP rate, TP rate, Precision, recall, F-measure.*

**2-fold cross validation**



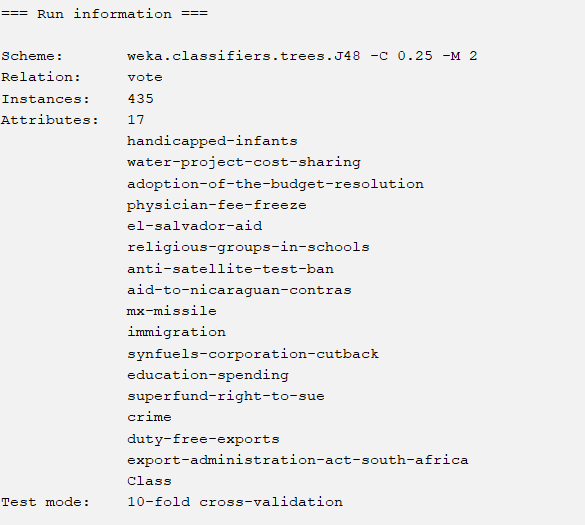
**Split dataset Confusion Matrix**



|  |  |  |
| --- | --- | --- |
|  | Classified as Democrat | Classified as Republican |
| Actual Democrats | TP = 86 | FN = 5 |
| Actual Republican | FP = 5 | TN = 52 |

## 10-fold cross validation

*Run 10-fold cross validation on the chosen dataset. Considering the option “Cross-validation Folds: 10”. Reporting the performance of the classifier as done for the 2-fold cross validation.*

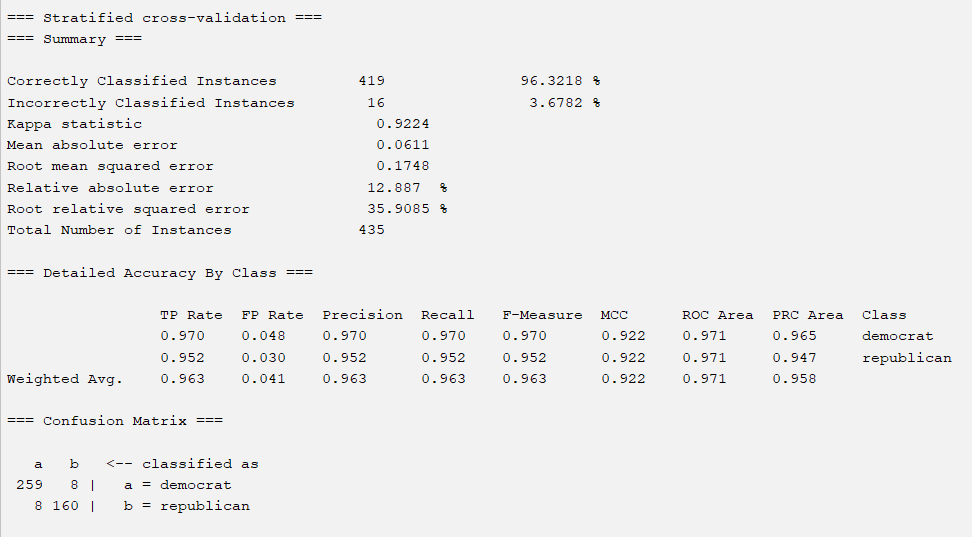


* + *overall classification accuracy:*

Split dataset 10-fold cross validation accuracy: **96.3218%**

* + *for each class value, reporting the FP rate, TP rate, Precision, recall, F-measure.*
  + *the overall weighted average over class levels for each of the of the FP rate, TP rate, Precision, recall, F-measure.*

**10-fold cross validation**



## Comparing 2-fold cross validation and 10-fold cross validation

*Comparing the performance of the classifier between 2-fold and 10-fold cross validation.*

* + *Which evaluation protocol provided higher performance?*

**As expected, the 10-fold cross validation resulted in a more accurate model:**

|  |  |
| --- | --- |
| Model | Accuracy |
| 2-fold cross validation | 93.2432% |
| 10-fold cross validation | 96.3218% |

*Interpret why?*

The 10-fold cross validation provided higher performance as it is based on a small dataset.

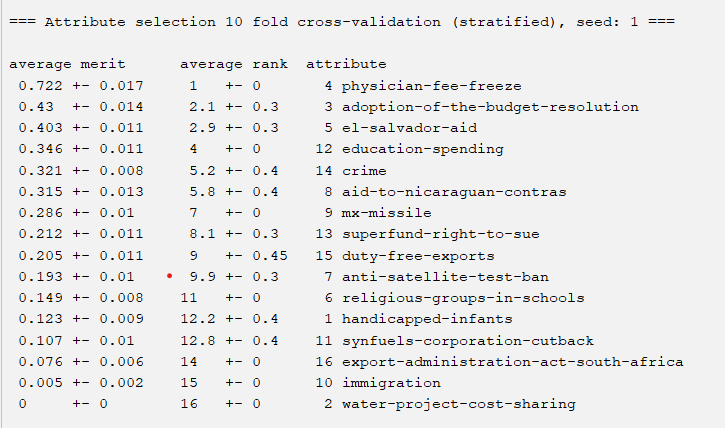
For big datasets, a 2-fold cross validation is enough.

# Feature selection

## Ranking attributes

*Use the select attributes tab in Weka to rank attributes by importance. Choose InfoGainAttributeEval and Ranker for the attribute evaluator and search method respectively and 10-fold cross validation for selecting attributes.*

## Show the list of the attributes by order of importance (based on average merit)

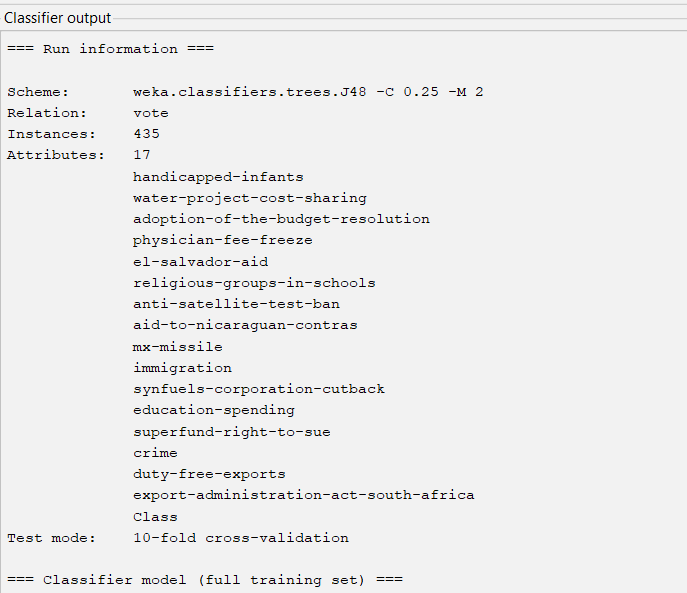


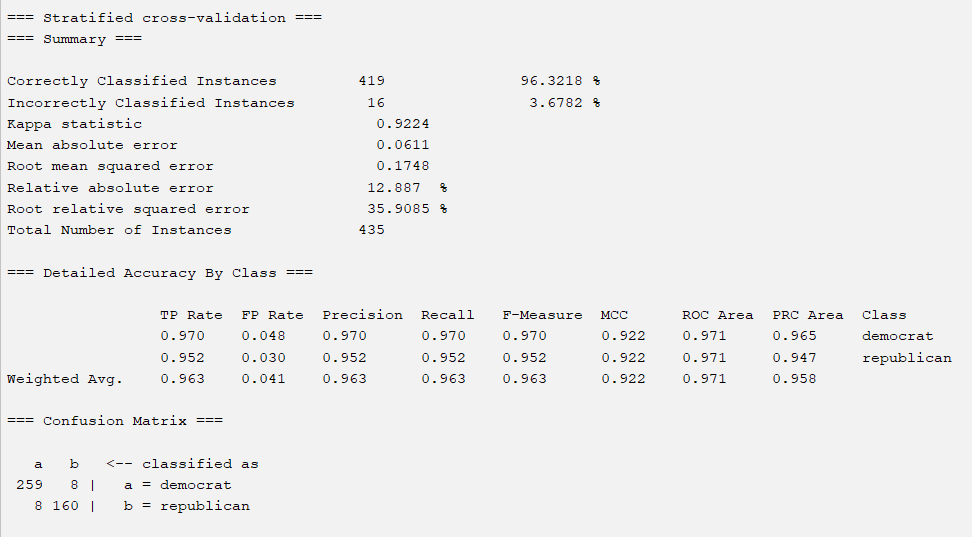
This ranking shows that the most important attribute to best predict the target variable (Class) is Physician-fee-freeze as it has the highest average merit (72.2%). In second place is Adoption-of-the-budget-resolution (43%) and in third place is El-Salvador-aid (40.3%). The least important attribute is Water-project-cost-sharing with an average merit of 0%.

# Classification using decision tree

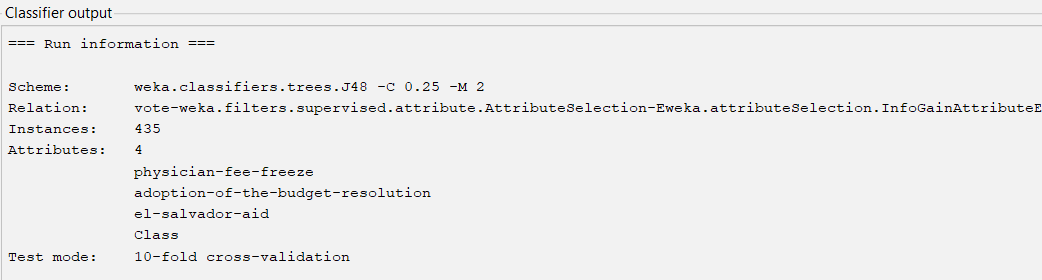
*For all the following runs, use 10-fold cross validation and reporting the FP rate, TP rate, Precision, recall, F-measure.*

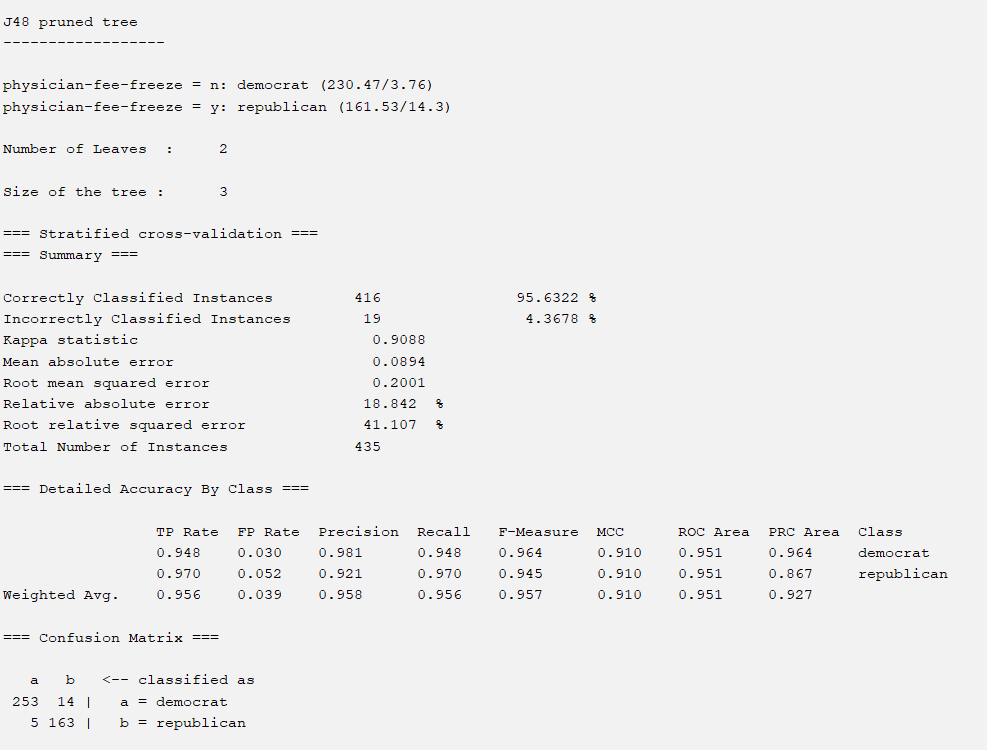
## Classification using all 16 attributes and 1 Class

**

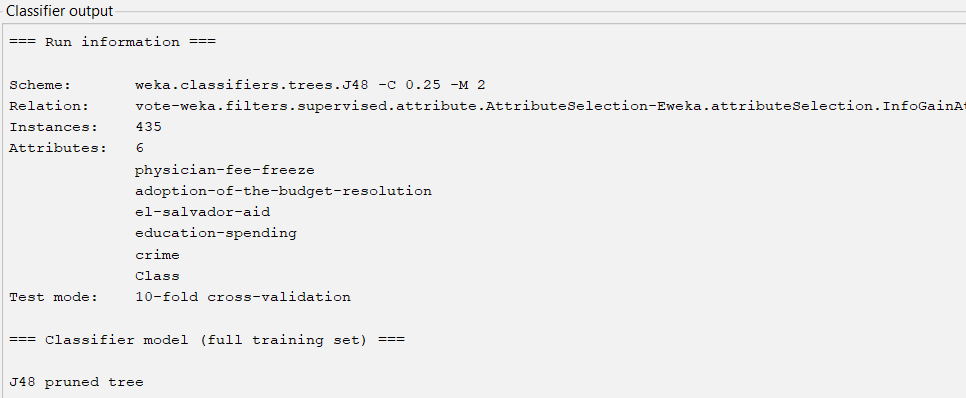


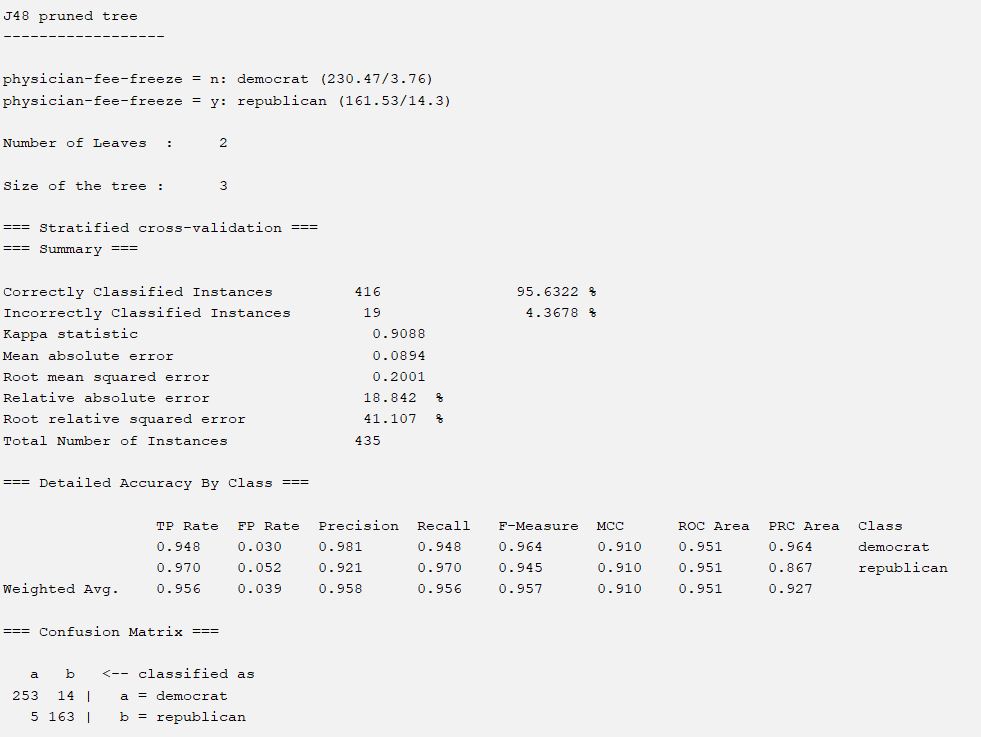
## Classification using best 3 attributes and 1 Class

**

**

## Classification using best 5 attributes and 1 Class

**

**

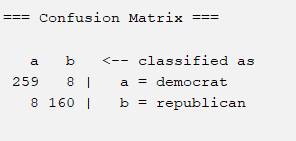
## Comparison

*Compare the classifier performance using all attributes, using best 3 attributes and using best 5 attributes.*

As seen in the comparison table below, the classifier with the best performance was the 10-fold cross validation using all the attributes.

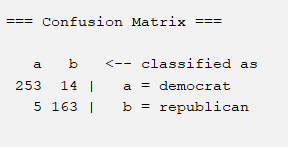
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Nr. of Attributes | Accuracy | TP Rate | FP Rate | Precision | Recall |
| All Attributes | 96.3218% | 96.3% | 4.1% | 96.3% | 963% |
| Best 3 attributes | 95.6322% | 95.6% | 3.9% | 95.8% | 95.6% |
| Best 5 attributes | 95.6322% | 95.6% | 3.9% | 95.8% | 95.6% |

**Confusion Matrix using all attributes:**



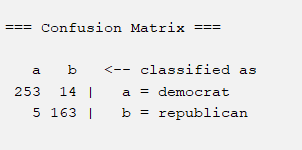
|  |  |  |
| --- | --- | --- |
|  | Classified as Democrat | Classified as Republican |
| Actual Democrats | TP = 259 | FN = 8 |
| Actual Republican | FP = 8 | TN = 160 |

**Confusion Matrix using best 3 attributes:**



|  |  |  |
| --- | --- | --- |
|  | Classified as Democrat | Classified as Republican |
| Actual Democrats | TP = 253 | FN = 14 |
| Actual Republican | FP = 5 | TN = 163 |

**Confusion Matrix using best 5 attributes:**



|  |  |  |
| --- | --- | --- |
|  | Classified as Democrat | Classified as Republican |
| Actual Democrats | TP = 253 | FN = 14 |
| Actual Republican | FP = 5 | TN = 163 |

*Does feature/attribute selection help increase the classifier performance?*

All three classifiers provided results of 95%+ accuracy. However, there was no difference between selecting 3 or 5 attributes. The number of True Positives remained the same for both selections (TP = 253) as well as the accuracy (95.6322%). Therefore, it is better to select 3 attributes as the training tree is smaller and this also helps prevent overfitting.

# Decision tree visualization and testing

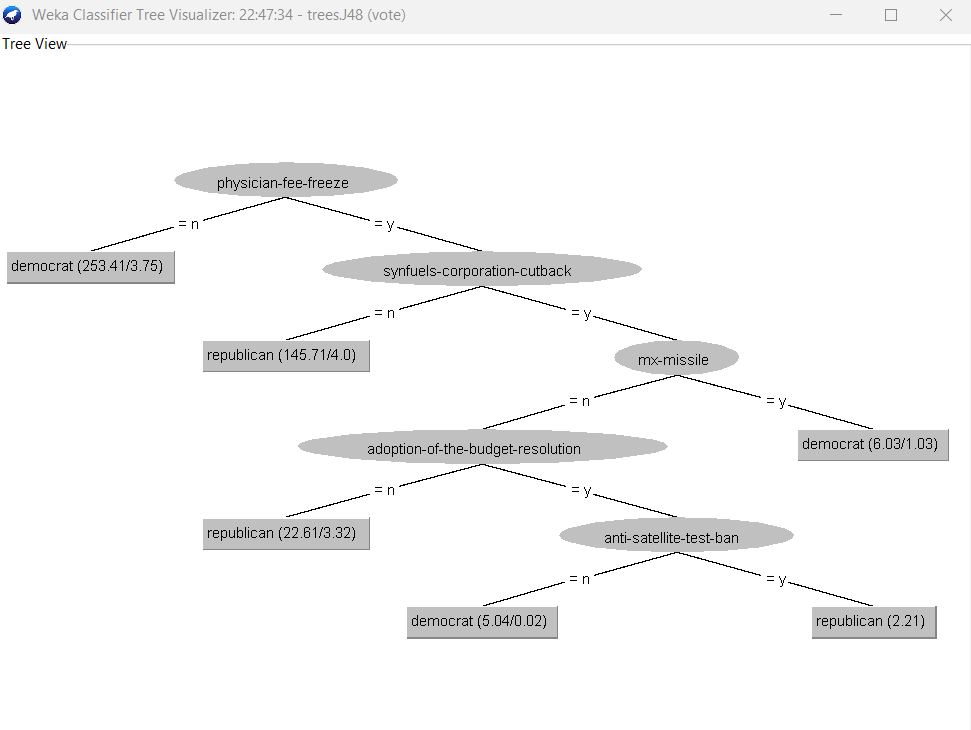
*Here, use 66% split for training and the rest for testing. Options should be as shown in the next column.*

## Decision tree model visualization

* + *Show the decision tree model.*
  + *What is the size of the tree?*

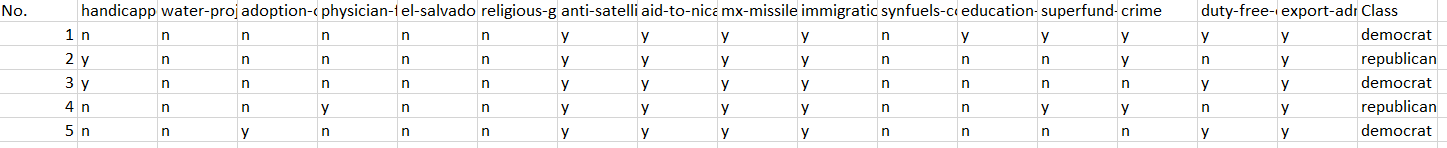
Size of the tree: 11

Number of leaves: 6



## Classification of testing instances

* + *Show the prediction of the first 5 instances of the testing set.* 
    - *Explain how they are classified by using the tree.*



|  |  |  |
| --- | --- | --- |
|  | Classified as | Actual |
| 1. Physician-fee-freeze | n = Democrat | Democrat |
| 1. Synfuels-corporation-cutback | n = Republican | Republican |
| 1. Mx-missile | y = Democrat | Democrat |
| 1. Adoption-of-the-budget-resolution | n = Republican | Republican |
| 1. Anti-satellite-test-ban | y = Republican | Democrat |

* The first node of the tree corresponds to the Physician-fee-freeze bill. On the dataset, the first instance classifies Physician-fee-freeze as “n”. On the tree, the “n” branch from the Physician-fee-free node has a “Democrat” leave. Therefore, the “n” classification for Physician-fee-freeze becomes “Democrat”. On the dataset, the actual Class for the first instance is “Democrat”. This means that Physician-fee-freeze is accurately classified.
* The second node of the tree corresponds to the Synfuels-corporation-cutback bill. On the dataset, the second instance classifies Synfuels-corporation-cutback as “n”. On the tree, the “n” branch from the Synfuels-corporation-cutback node has a “Republican” leave. Therefore, the “n” classification for Synfuels-corporation-cutback becomes “Republican”. On the dataset, the actual Class for the second instance is “Republican”. This means that Synfuels-corporation-cutback is accurately classified.
* The third node of the tree corresponds to the Mx-missile bill. On the dataset, the third instance classifies Mx-missile as “y”. On the tree, the “y” branch from the Mx-missile node has a “Democrat” leave. Therefore, the “y” classification for the Mx-missile becomes “Democrat”. On the dataset, the actual Class for the third instance is “Democrat”. This means that Mx-missile is accurately classified.
* The fourth node of the tree corresponds to the Adoption-of-the-budget-resolution bill. On the dataset, the fourth instance classifies Adoption-of-the-budget-resolution as “n”. On the tree, the “n” branch from the Adoption-of-the-budget-resolution node has a “Republican” leave. Therefore, the “n” classification for the Adoption-of-the-budget-resolution becomes “Republican”. On the dataset, the actual Class for the fourth instance is “Republican”. This means that Adoption-of-the-budget-resolution is accurately classified.
* The fifth node of the tree corresponds to the Anti-satellite-test-ban bill. On the dataset, the fifth instance classifies Anti-satellite-test-ban as “y”. On the tree, the “y” branch from the Anti-satellite-test-ban node has a “Republican” leave. Therefore, the “y” classification for the Anti-satellite-test-ban becomes “Republican”. On the dataset, the actual Class for the fifth instance is “Democrat”. This means that Anti-satellite-test-ban is not accurately classified.