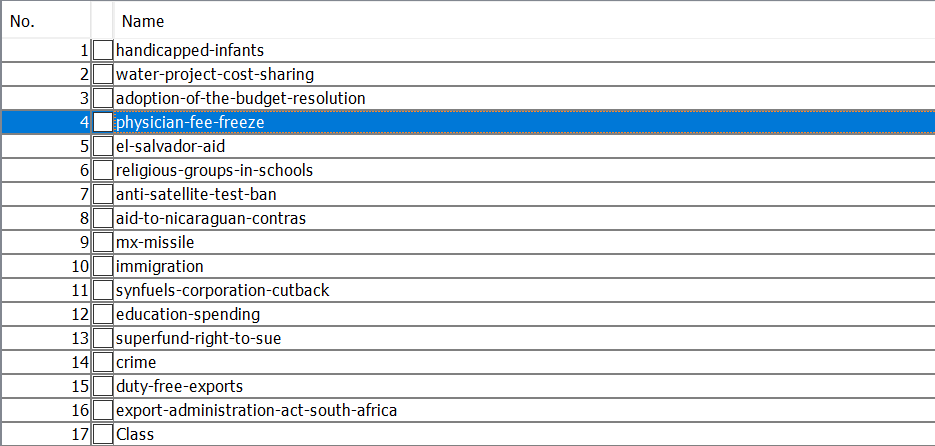
**21L-6269 Eesha Tariq Data Mining Lab BDS-6A2 Lab-04**

**Dataset used: vote.arff**

****

|  |  |  |  |
| --- | --- | --- | --- |
| **SNo** | **Applied Formula** | **Result** | **Reasoning** |
| 1 | ReplaceMissingValues | Replaced missing values in most of the attriutes. | Replaced the missing values with the above value (data is nominal) |

1. **What is the size of the training set?**

435 Rows/Instances

1. **How many attributes exist in the training set?**

17 attributes

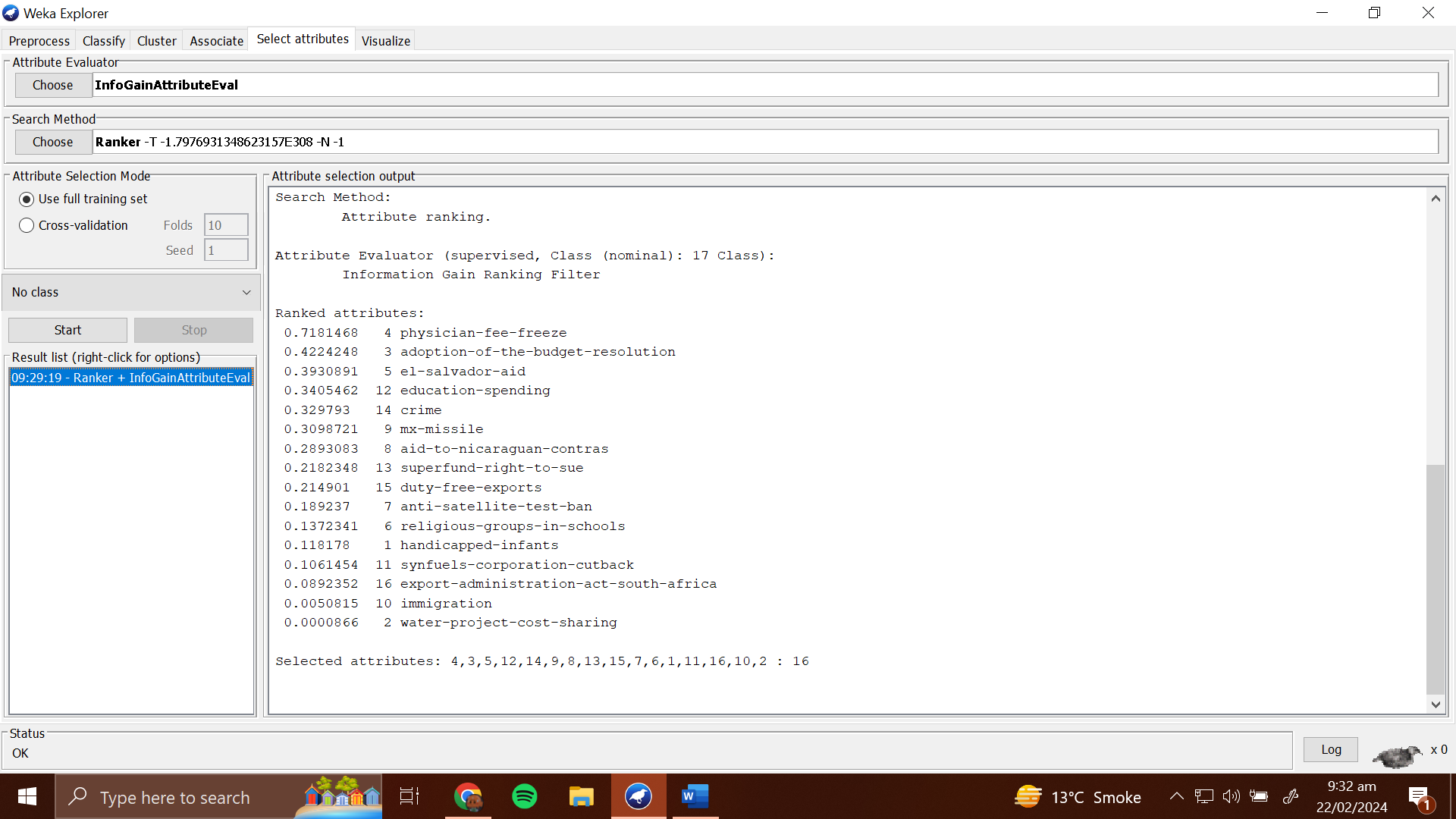
1. **How many instances are republican and how many democrat?**

Republican: 168

Democrat: 267

1. **Which attribute best separates the data?**

|  |  |  |  |
| --- | --- | --- | --- |
| **SNo** | **Applied Formula** | **Result** | **Reasoning** |
| 2 | InfoGainAttributeEval | Ranked attributes by information gain | Information gain determines the best split in a decision tree |



1. **How many elements from the data set have physician fee freeze as yes?**

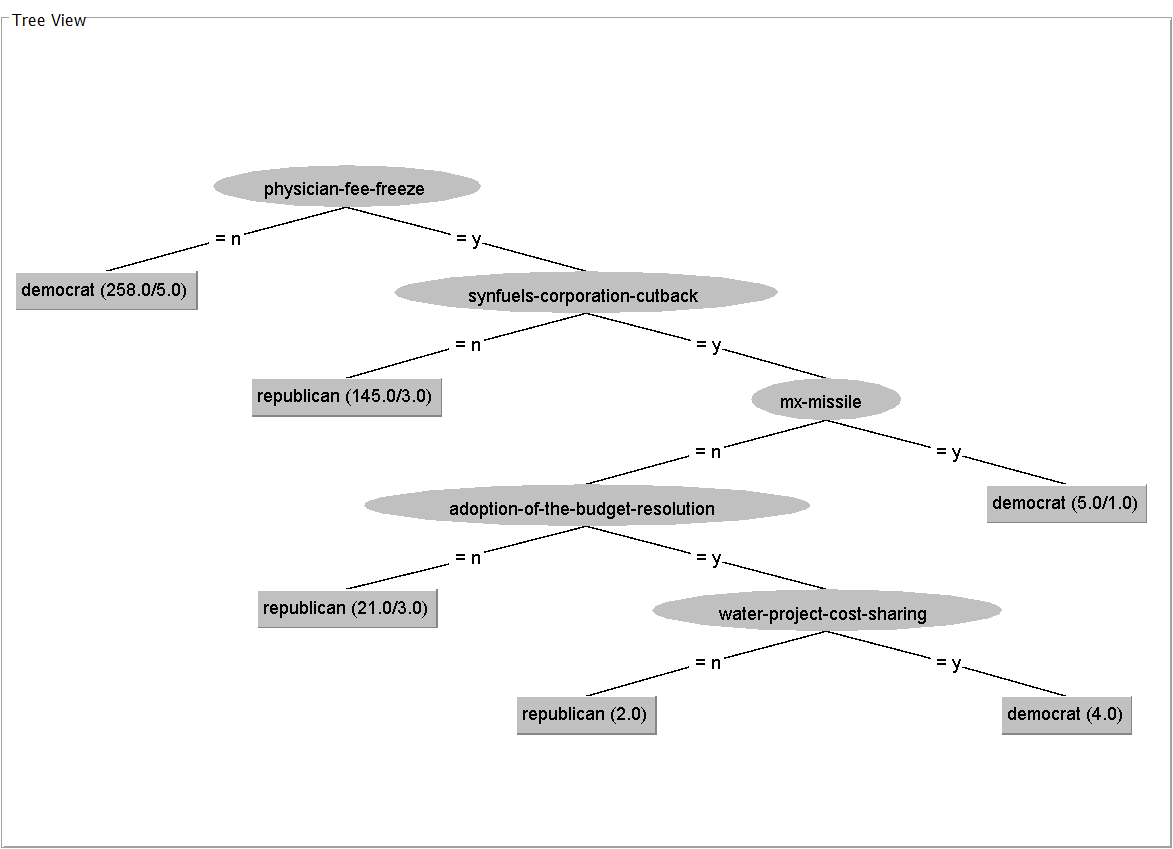
177 yes

**Apply J48:**

|  |  |  |  |
| --- | --- | --- | --- |
| **SNo** | **Applied Formula** | **Result** | **Reasoning** |
| 3 | J48 -C 0.25 -M 2 | Screenshot | Information gain determines the best split in a decision tree |

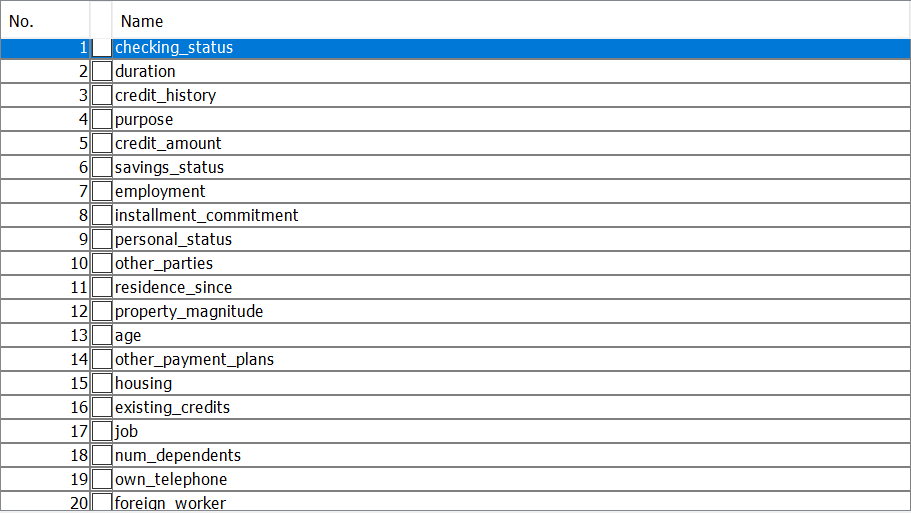
A screenshot of a computer

Description automatically generated

****

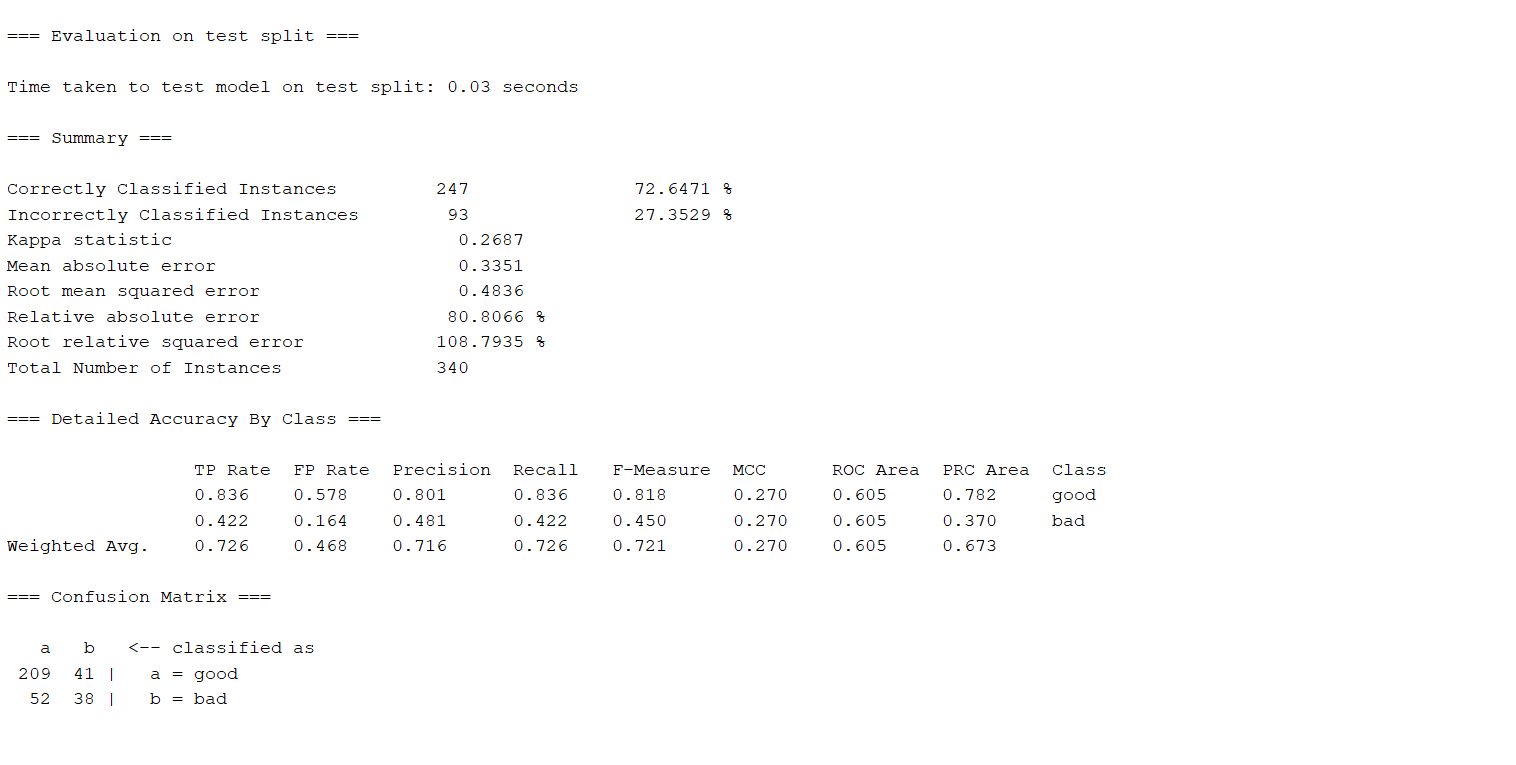
**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dataset used: credit-g.arff**



**Apply J48:**

|  |  |  |  |
| --- | --- | --- | --- |
| **SNo** | **Applied Formula** | **Result** | **Reasoning** |
| 1 | J48 -C 0.25 -M 2 | Screenshot | Creates a decision tree based of information gain (C4.5) classification. |

****

**Apply ZeroR:**

|  |  |  |  |
| --- | --- | --- | --- |
| **SNo** | **Applied Formula** | **Result** | **Reasoning** |
| 2 | ZeroR | Screenshot | Classifies data based on majority class of target variable. |

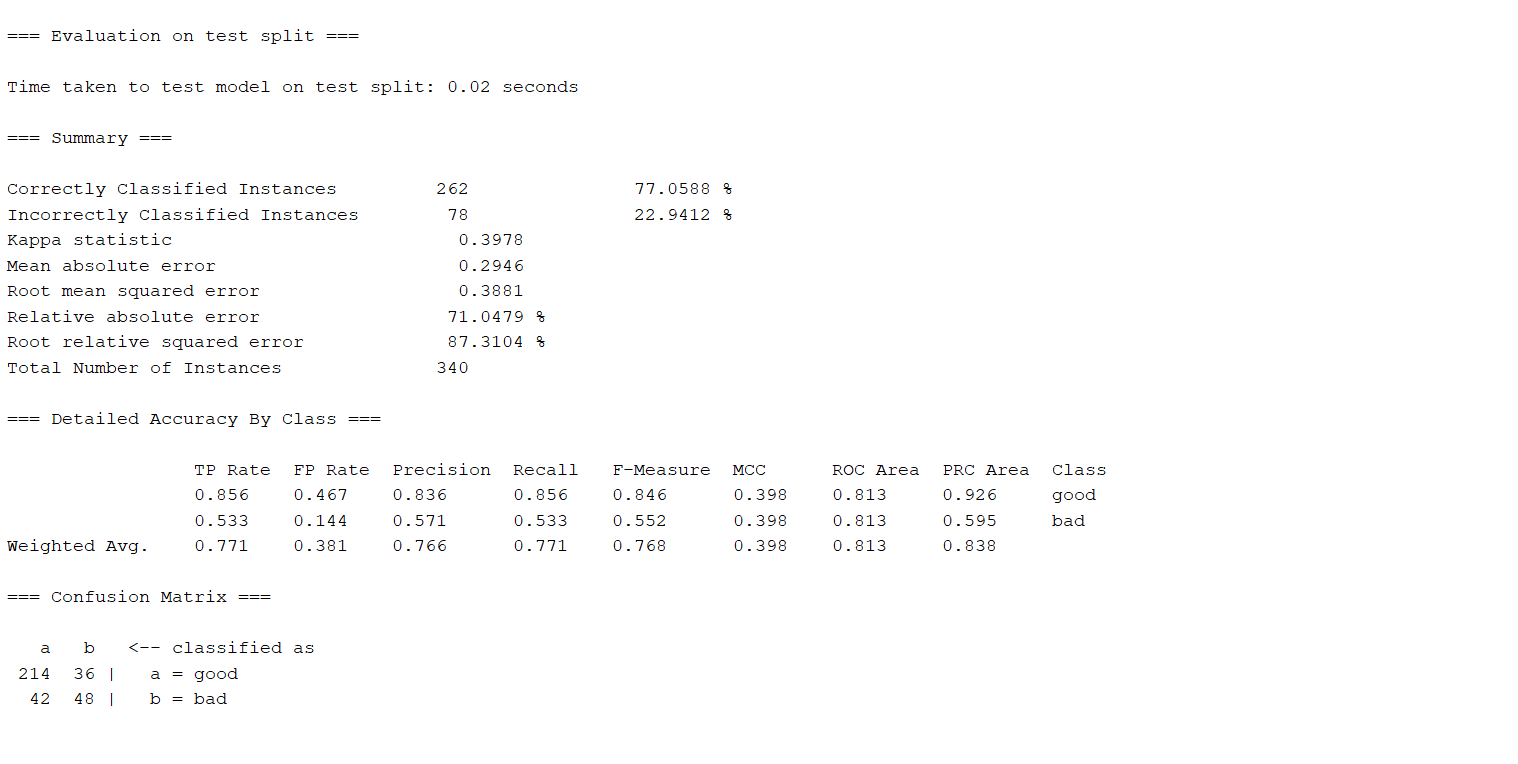
**A white screen with black text

Description automatically generated with medium confidence**

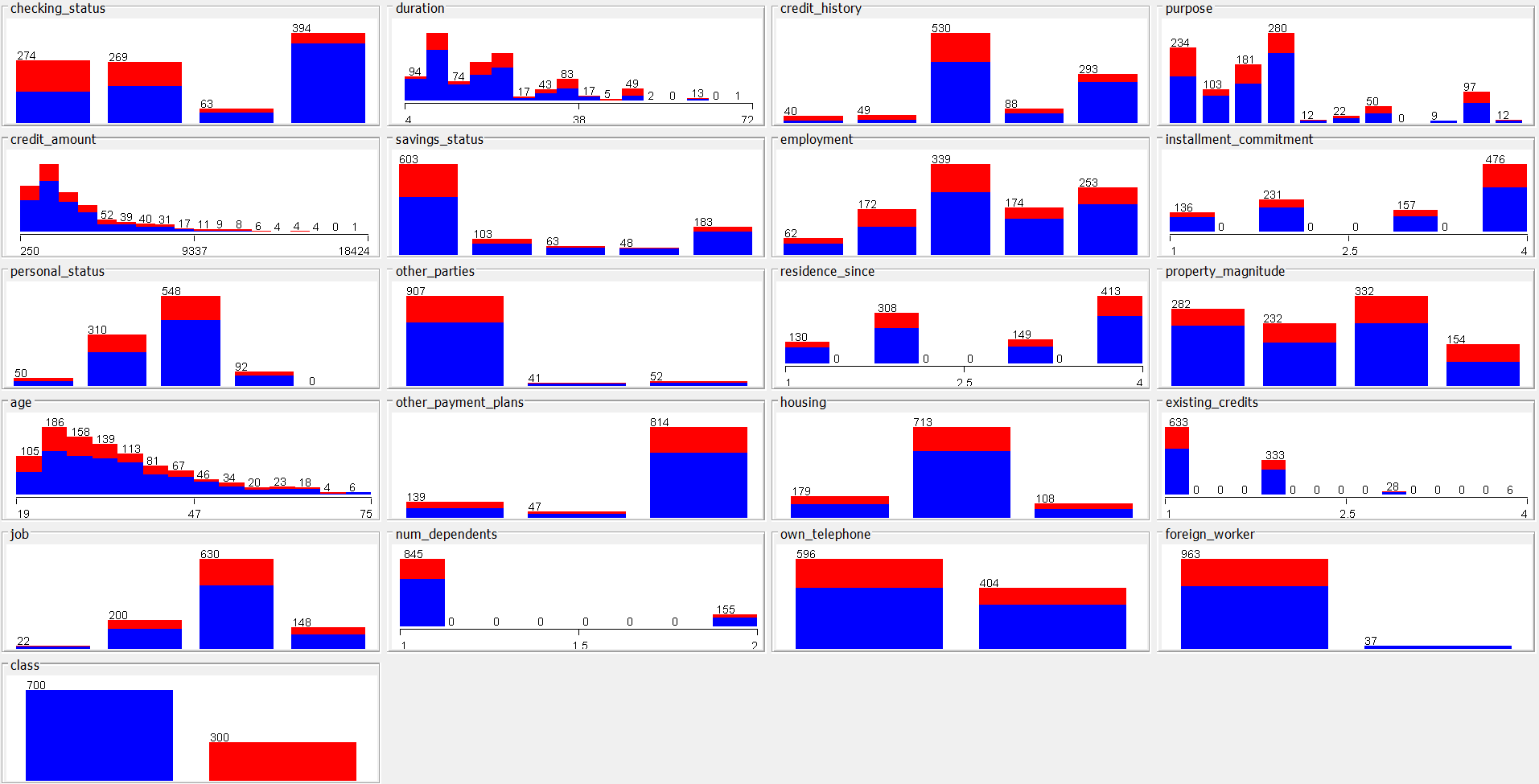
From the results, we can conclude that ZeroR is higher in terms of accuracy, but the J48 algorithm gives more precision and recall. However, since ZeroR is rather poor classifier (classifies based on majority class), we cannot claim it is better than J48, which is a learning algorithm.

**Apply SimpleLogistic:**

|  |  |  |  |
| --- | --- | --- | --- |
| **SNo** | **Applied Formula** | **Result** | **Reasoning** |
| 3 | SimpleLogistic -I 0 -M 500 -H 50 -W 0.0 | Screenshot | Classifies data based one dependent variable. |



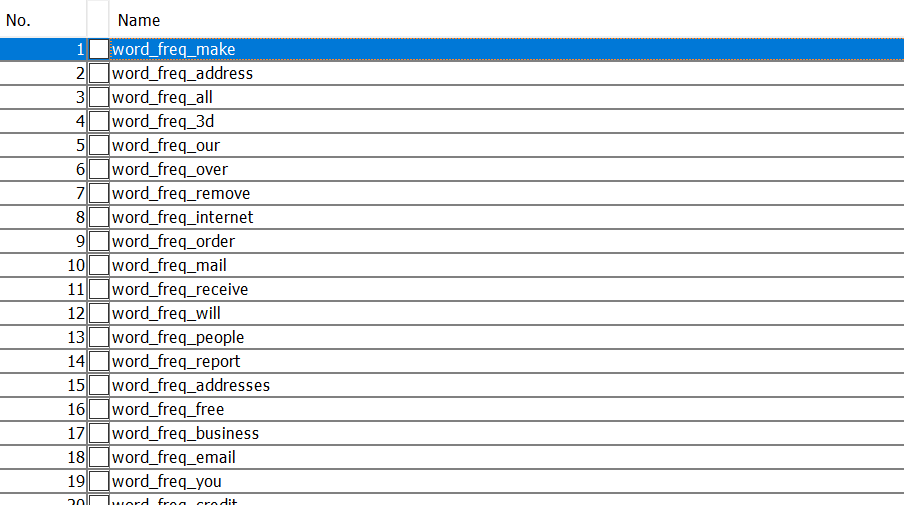
7. The distribution being complete (no missing values ‘none’ in the classes, like existing\_credits and num\_dependents) can prove to be more beneficial during the classification process, than the randomly distributed variables. Further, imbalanced classes like foreign\_workers, and other\_parties provide less support for the model than the ones with a skewed or normal distribution.



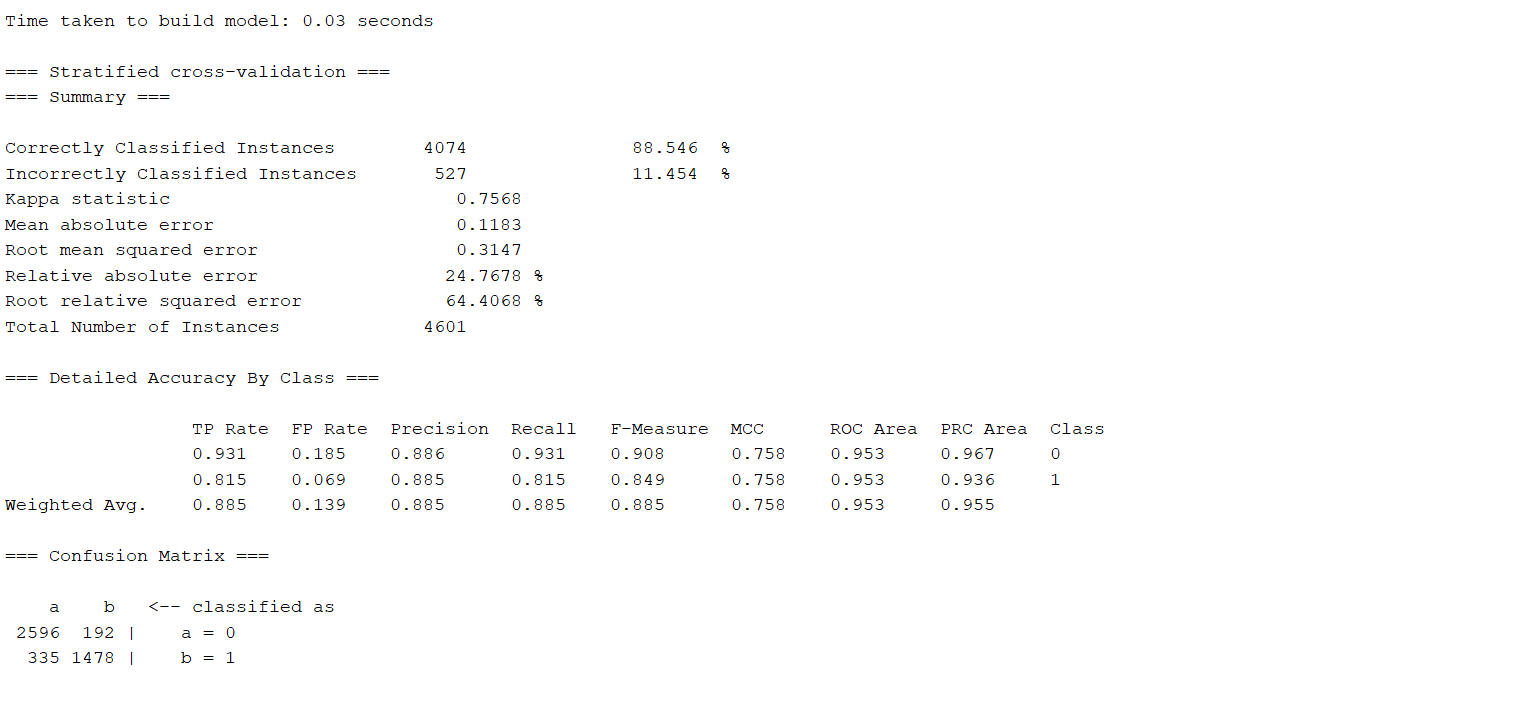
8. Different data types, distributions and class balances need different treatment in finding the right algorithm. A machine learning model's performance is directly affected by data quality as well. Thus, looking at the data can help know gaps, weaknesses, and strengths of your dataset. Lastly, looking at your dataset can help intuitively know what attributes are useful for classification.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dataset used: spambase.arff**



|  |  |  |  |
| --- | --- | --- | --- |
| **SNo** | **Applied Formula** | **Result** | **Reasoning** |
| 1 | NumericToBinary -R first-last | Converted numeric data to Boolean classes | Eases processing |
| 2 | NaiveBayes | Screenshot | --- |



1. **Can you come up with a reason for the good performance? What would be the main practical problems we would face if we were not to make this assumption for this particular dataset?**
2. **How long did your classifier take to train and classify? Given this, how scalable do you think the Naïve Bayes classifier is to large datasets? Can you come up with a good reason for this?**

The classifier barely took 0.03 seconds to complete, which means that it is highly scalable on larger language models.