

CP 5634  
DATA MINING

DATA MINING REPORT ON CENSUS  
INCOME

**SUBMITTED BY:**

VIPULDEEP SINGH GULATI

STUDENT ID: 13636038

JC497971

## Contents

Abstract.....	3
Introduction .....	3
Census Income Data Mining .....	3
Introduction to Weka.....	3
Dataset Description: .....	7
Data Mining Steps.....	9
This is the look when the .csv extension file is opened in Weka.....	10
Data Cleaning .....	11
Data Integration .....	11
Data Transformation.....	11
Data Selection .....	12
Data Normalization .....	13
Data Discretization.....	14
Data Classification.....	15
Conclusion.....	17
References .....	18
Appendix .....	19

## Abstract

The mined dataset of Census Income is used to predict the number of people who has annual income above 50k. It is based on work class, occupation, native country and many more. Census Income dataset has categorical and integer attribute characteristics. The data set has a y attribute which describes the salaries of people earning above and below 50k. This Dataset consist of 14 attributes and 48842 instances. Census Income was downloaded from UCI Machine Learning Repository. Initially, Dataset was not Weka compatible. It was converted into CSV format. Attribute headings were also added to the dataset. Missing Values are removed and “?” in dataset represents other or unknown.

## Introduction

The aim is to use various data mining algorithms on Census Income in order to predict the number of people who are earning more than 50k annually. This can be done by understanding and using data mining tools. Data is usually hard to understand as it is bunch of numbers. This data is continuous, filled with errors, filled with gaps and in various formats (static, streamed, www etc). Also known as pattern mining turns data into knowledge. Data Mining is the non-trivial extraction of implicit, previously unknown, and potentially useful information from data (William J Frawley, n.d.).

Potentially large datasets take long time when working with Weka. Knowledge extraction from these datasets is based on the goal what a person needs to extracts. Weka is the software which is used to mine Census Income dataset.

## Census Income Data Mining

### Introduction to Weka

Weka is open source software which is written in java. Source code can be modified in this software. Weka is preferred because it is easy to use, it has a mouse point-and-click GUI

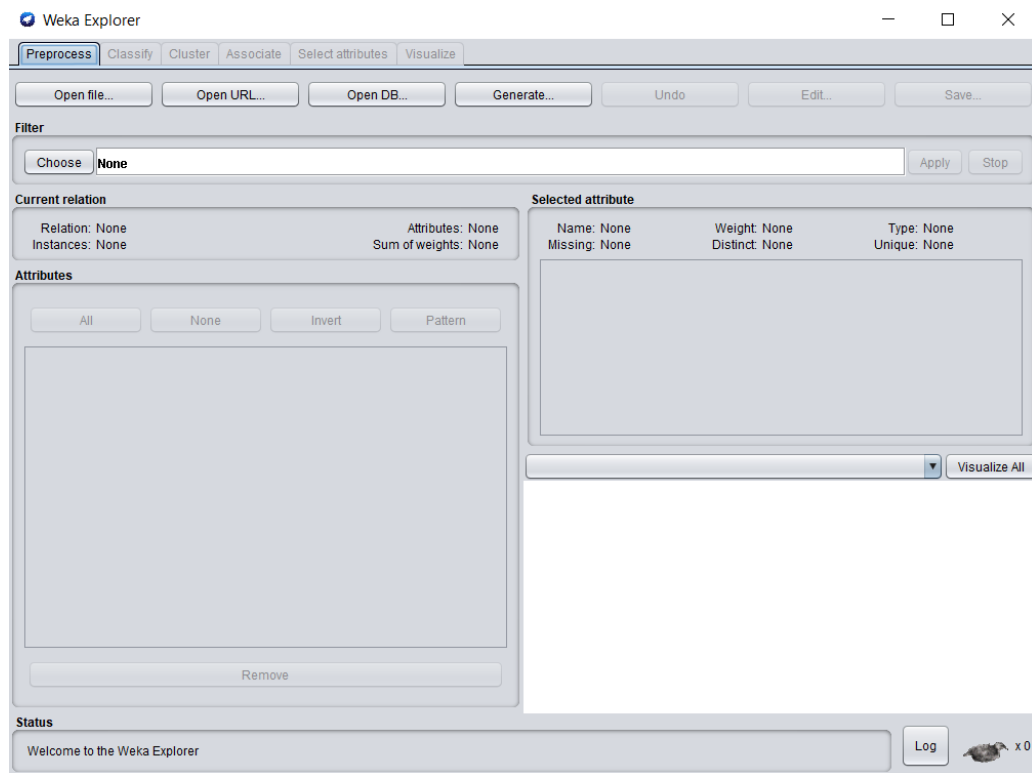
interface and can compare algorithms quickly. Weka has a list of applications which can be used to achieve different tasks. These applications are mentioned below:

- Explorer
- Experimenter
- Knowledge Flow
- Workbench
- Simple CLI

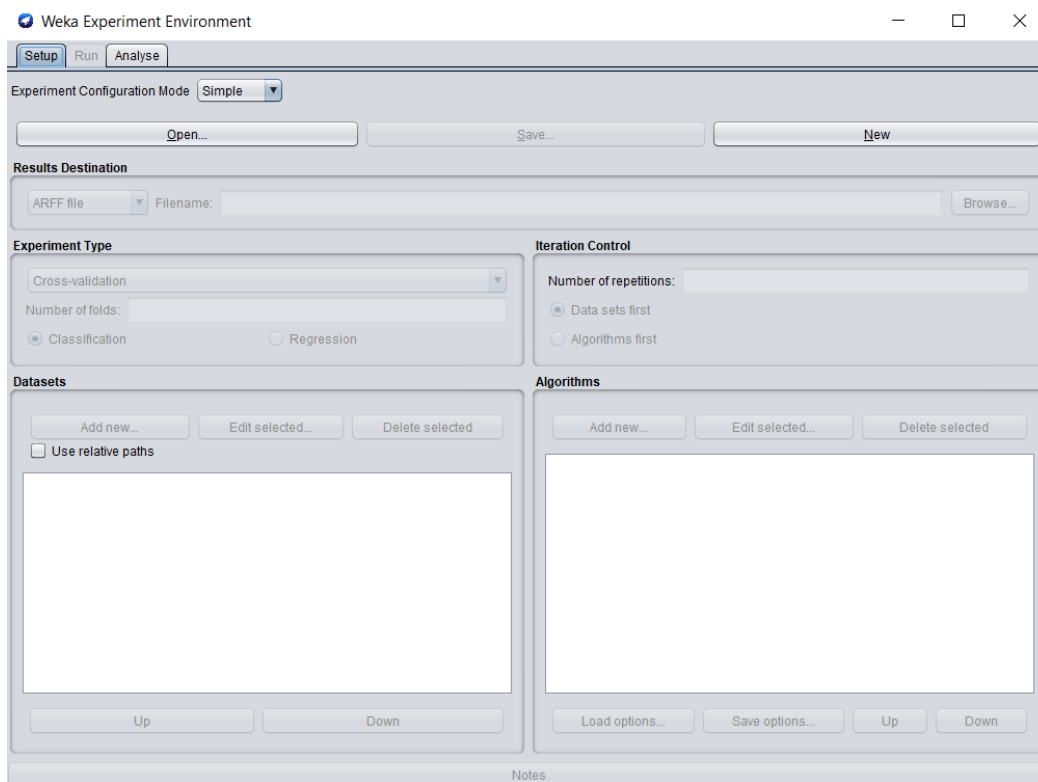
**Weka GUI Chooser:**



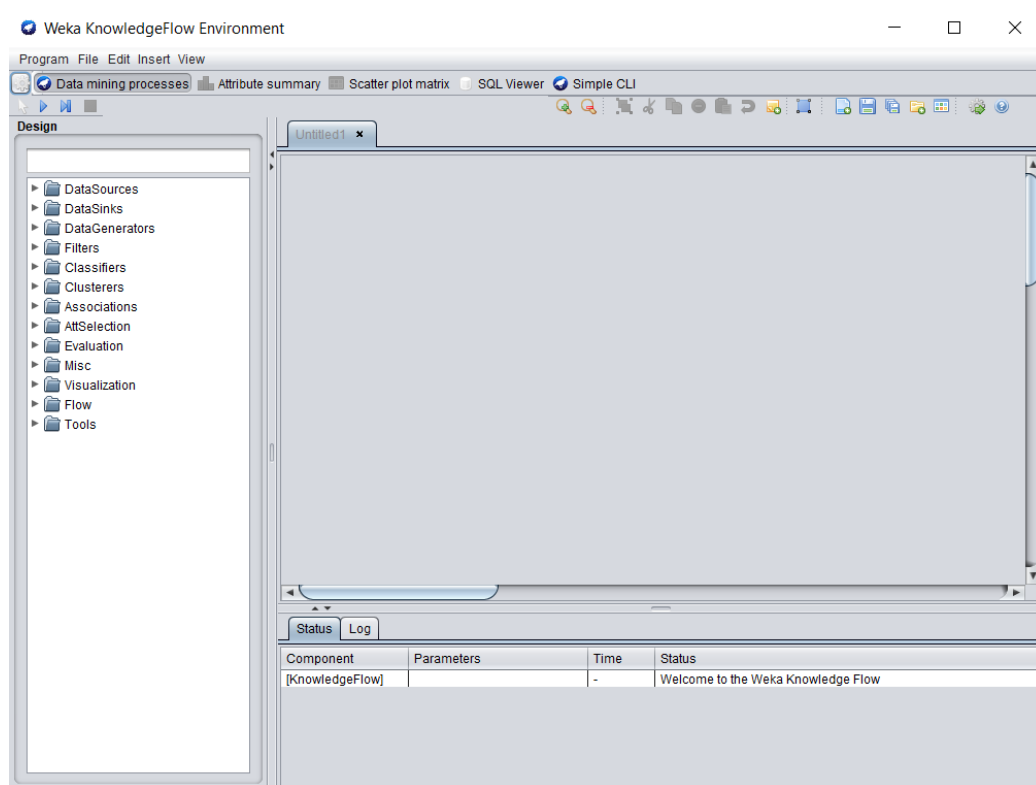
## Weka Explorer:



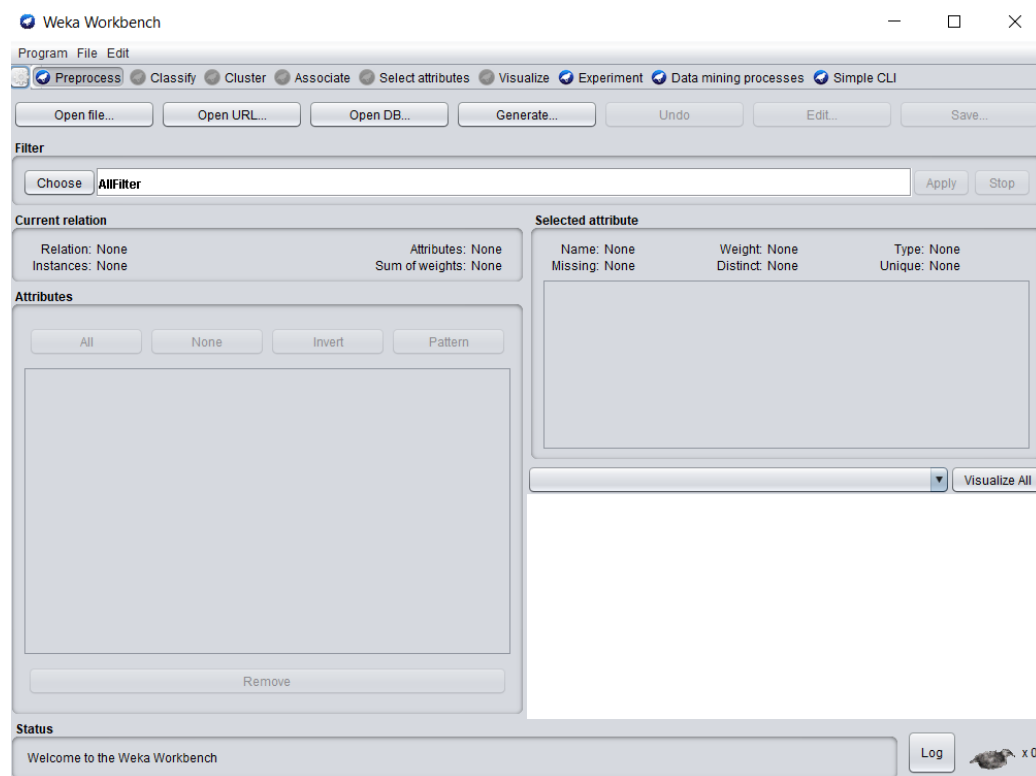
## Weka Experimenter:



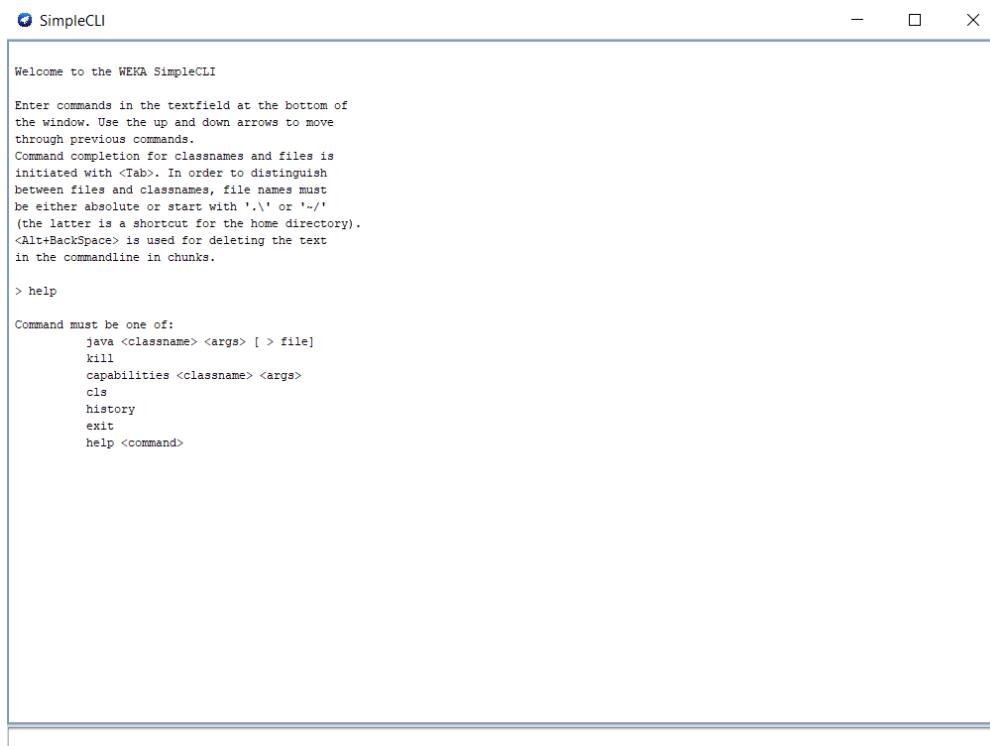
## Weka KnowledgeFlow:



## Weka Workbench:



## Weka Simple CLI:



```
SimpleCLI

Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with './' or './'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
  java <classname> <args> [ > file]
  kill
  capabilities <classname> <args>
  cls
  history
  exit
  help <command>
```

All the data mining tools and techniques are used in Weka Explorer.

## Dataset Description:

The dataset consists of data regarding the Persons attributes which lists down the information who has more income and who has less income. The aim is to determine who earns more than 50k annually. The last attribute represents the annually income as “<=50k” more than 50k or “>=50k” less than 50k. Initial information of the Census Income dataset is mentioned below:

ATTRIBUTE CHARACTERSTICS	Categorical, Integer
ASSOCIATED TASKS	Classification
NUMBER OF ATTRIBUTES	14
NUMBER OF INSTANCES	48842

## Census Income Attributes Description:

- **Age:** Age of the person.
  - Measurement Type: Continuous.

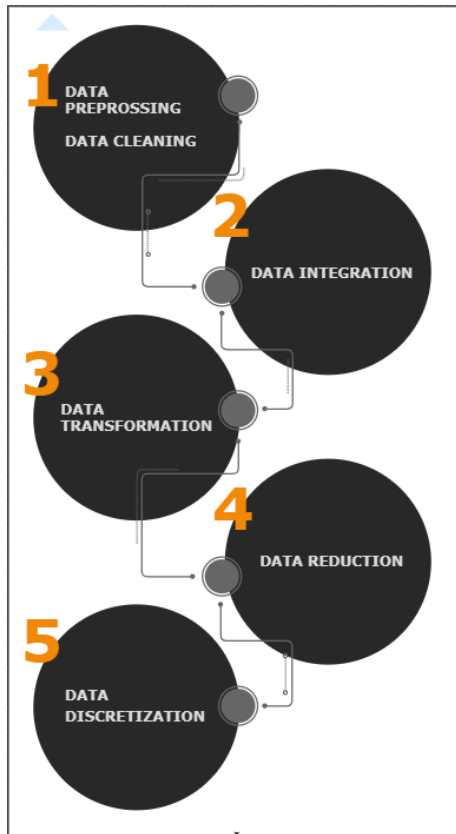
- Example: (39,50,38,53,27).
- **Workclass:** Sector of the organization.
  - **Measurement Type:** Numeric.
  - **Example:** (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked).
- **fnlwgt**
  - Measurement Type: Continuous.
  - Example: (39,50,38,53,27).
- **Education:** Level of how graduated a person is.
  - Measurement Type: Categorical.
  - Example: (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool).
- **Education-Num**
  - Measurement Type: Numeric.
  - Example: (13,9,7).
- **Marital Status**
  - Measurement Type: Categorical.
  - Example: (Never-married, Divorced, Married etc).
- **Occupation**
  - Measurement Type: Categorical.
  - Example: (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces).
- **Relationship**
  - Measurement Type: Categorical.
  - Example: (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried).
- **Race**
  - Measurement Type: Categorical.
  - Example: (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black).



- **Sex**
  - Measurement Type: Categorical.
  - Example: (Female, Male).
- **Capital-gain**
  - Measurement Type: Continuous.
  - Example: (2174, 0).
- **Capital-loss**
  - Measurement Type: Continuous.
  - Example: (3973,0,32).
- **hours-per-week**
  - Measurement Type: Continuous.
  - Example: (40,13,40).
- **Native-country**
  - Measurement Type: Categorical.
  - Example: (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad & Tobago, Peru, Hong, Holand-Netherlands).
- **Y:**
  - Is person earning more than 50k annually or less than 50k.
  - “<=50k”, “>=50k” (Becker, 1996).

## Data Mining Steps

**Data Pre-processing:** Data Pre-processing is important because data must be combined combined, transformed, reduced, discretized, visualized to extract maximum knowledge out of it. Data can be incomplete, noisy and inconsistent which means it can contain incomplete values, can contain errors and can contain discrepancies. The steps for data pre-processing are mentioned below:



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** [Apply] [Stop]

**Current relation**  
 Relation: adultdataa  
 Instances: 32561  
 Attributes: 15  
 Sum of weights: 32561

**Attributes**  
 [All] [None] [Invert] [Pattern]

No.	Name
1	<input checked="" type="checkbox"/> Age
2	<input type="checkbox"/> Workclass
3	<input type="checkbox"/> fnlwgt
4	<input type="checkbox"/> Education
5	<input type="checkbox"/> Education - Num
6	<input type="checkbox"/> Marital Status
7	<input type="checkbox"/> Occupation
8	<input type="checkbox"/> Relationship
9	<input type="checkbox"/> Race
10	<input type="checkbox"/> Sex
11	<input type="checkbox"/> Capital-gain
12	<input type="checkbox"/> Capital-loss
13	<input type="checkbox"/> Hours-per-week
14	<input type="checkbox"/> Native-country
15	<input type="checkbox"/> y

[Remove]

**Selected attribute**  
 Name: Age  
 Missing: 0 (0%)  
 Distinct: 73  
 Type: Numeric  
 Unique: 2 (0%)

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Class: y (Nom) [Visualize All]

Status: OK [Log] x 0

This is the look when the .csv extension file is opened in Weka.

## Data Cleaning

Data cleaning is referred to as filling up of missing values, null spaces and identifying or removing outliers in order to make it more relevant. Data cleaning and pre-processing the major task and the most difficult task out of all as it could end up consuming 60 percent of user's time. The dataset for this report was already clean so it required no cleaning. The "?" represented unknown.

## Data Integration

Data Integration combines multiple data sets of different formats into the same format before data transformation. For this assignment, only single format has been taken. So, Data Integration has been eliminated in data pre-processing.

## Data Transformation

After Data cleaning and data integration comes data transformation which leads to normalize data into same scale. Data before this is difficult to understand as it is only differentiated by comma. And for Weka, it requires CSV file extension to run.

The Text file is opened in MS Excel and converted to a clearer format by using Text-To-Columns data tool which delimits and separates the data on the basis of comma. Then the data is thoroughly checked and save with the Weka compatible format which is .csv. The difference before using Text-to-columns data tool and after its use is shown.

### Before using Text-to-columns data tool:

1	39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K	
2	50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K	
3	38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K	
4	53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K	
5	28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K	
6	37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K	
7	49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K	
8	52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K	
9	31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >50K	
10	42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >50K	
11	37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >50K	
12	30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K	
13	23, Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K	
14	32, Private, 205019, Assoc-acdm, 12, Never-married, Sales, Not-in-family, Black, Male, 0, 0, 50, United-States, <=50K	
15	40, Private, 121772, Assoc-voc, 11, Married-civ-spouse, Craft-repair, Husband, Asian-Pac-Islander, Male, 0, 0, 40, ?, >50K	
16	34, Private, 245487, 7th-8th, 4, Married-civ-spouse, Transport-moving, Husband, Amer-Indian-Eskimo, Male, 0, 0, 45, Mexico, <=50K	
17	25, Self-emp-not-inc, 176756, HS-grad, 9, Never-married, Farming-fishing, Own-child, White, Male, 0, 0, 35, United-States, <=50K	
18	32, Private, 186824, HS-grad, 9, Never-married, Machine-op-inspct, Unmarried, White, Male, 0, 0, 40, United-States, <=50K	
19	38, Private, 28887, 11th, 7, Married-civ-spouse, Sales, Husband, White, Male, 0, 0, 50, United-States, <=50K	
20	43, Self-emp-not-inc, 292175, Masters, 14, Divorced, Exec-managerial, Unmarried, White, Female, 0, 0, 45, United-States, >50K	
21	40, Private, 193524, Doctorate, 16, Married-civ-spouse, Prof-specialty, Husband, White, Male, 0, 0, 60, United-States, >50K	
22	54, Private, 302146, HS-grad, 9, Separated, Other-service, Unmarried, Black, Female, 0, 0, 20, United-States, <=50K	
23	35, Federal-gov, 76845, 9th, 5, Married-civ-spouse, Farming-fishing, Husband, Black, Male, 0, 0, 40, United-States, <=50K	
24	43, Private, 117037, 11th, 7, Married-civ-spouse, Transport-moving, Husband, White, Male, 0, 2042, 40, United-States, <=50K	
25	59, Private, 109015, HS-grad, 9, Divorced, Tech-support, Unmarried, White, Female, 0, 0, 40, United-States, <=50K	
26	56, Local-gov, 216851, Bachelors, 13, Married-civ-spouse, Tech-support, Husband, White, Male, 0, 0, 40, United-States, >50K	
27	19, Private, 168294, HS-grad, 9, Never-married, Craft-repair, Own-child, White, Male, 0, 0, 40, United-States, <=50K	
28	54, ?, 180211, Some-college, 10, Married-civ-spouse, ?, Husband, Asian-Pac-Islander, Male, 0, 0, 60, South, >50K	
29	39, Private, 367260, HS-grad, 9, Divorced, Exec-managerial, Not-in-family, White, Male, 0, 0, 80, United-States, <=50K	

## After using Text-to-columns data tool:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Workclass	fnlwgt	Education	ucation	N	Marital Status	Occupation	Relationship	Race	Sex	Capital-gain	Capital-loss	hours-per-week	native-country
2	39	State-gov	77516	Bachelors		13	Never-mar	Adm-cler	Not-in-far	White	Male	2174	0	40	United-St
3	50	Self-emp	83311	Bachelors		13	Married-c	Exec-man	Husband	White	Male	0	0	13	United-St
4	38	Private	215646	HS-grad		9	Divorced	Handlers	Not-in-far	White	Male	0	0	40	United-St
5	53	Private	234721	11th		7	Married-c	Handlers	Husband	Black	Male	0	0	40	United-St
6	28	Private	338409	Bachelors		13	Married-c	Prof-spec	Wife	Black	Female	0	0	40	Cuba
7	37	Private	284582	Masters		14	Married-c	Exec-man	Wife	White	Female	0	0	40	United-St
8	49	Private	160187	9th		5	Married-s	Other-ser	Not-in-far	Black	Female	0	0	16	Jamaica
9	52	Self-emp	209642	HS-grad		9	Married-c	Exec-man	Husband	White	Male	0	0	45	United-St
10	31	Private	45781	Masters		14	Never-mar	Prof-spec	Not-in-far	White	Female	14084	0	50	United-St
11	42	Private	159449	Bachelors		13	Married-c	Exec-man	Husband	White	Male	5178	0	40	United-St
12	37	Private	280464	Some-coll		10	Married-c	Exec-man	Husband	Black	Male	0	0	80	United-St
13	30	State-gov	141297	Bachelors		13	Married-c	Prof-spec	Husband	Asian-Pac	Male	0	0	40	India
14	23	Private	122272	Bachelors		13	Never-mar	Adm-cler	Own-child	White	Female	0	0	30	United-St
15	32	Private	205019	Assoc-acd		12	Never-mar	Sales	Not-in-far	Black	Male	0	0	50	United-St
16	40	Private	121772	Assoc-voc		11	Married-c	Craft-rep	Husband	Asian-Pac	Male	0	0	40	?
17	34	Private	245487	7th-8th		4	Married-c	Transport	Husband	Amer-Indi	Male	0	0	45	Mexico
18	25	Self-emp	176756	HS-grad		9	Never-mar	Farming-f	Own-child	White	Male	0	0	35	United-St
19	32	Private	186824	HS-grad		9	Never-mar	Machine	Unmarrie	White	Male	0	0	40	United-St
20	38	Private	28887	11th		7	Married-c	Sales	Husband	White	Male	0	0	50	United-St
21	43	Self-emp	292175	Masters		14	Divorced	Exec-man	Unmarrie	White	Female	0	0	45	United-St
22	40	Private	193524	Doctorate		16	Married-c	Prof-spec	Husband	White	Male	0	0	60	United-St
23	54	Private	302146	HS-grad		9	Separated	Other-ser	Unmarrie	Black	Female	0	0	20	United-St
24	35	Federal-g	76845	9th		5	Married-c	Farming-f	Husband	Black	Male	0	0	40	United-St
25	43	Private	117037	11th		7	Married-c	Transport	Husband	White	Male	0	2042	40	United-St
26	59	Private	109015	HS-grad		9	Divorced	Tech-supp	Unmarrie	White	Female	0	0	40	United-St
27	56	Local-gov	216851	Bachelors		13	Married-c	Tech-supp	Husband	White	Male	0	0	40	United-St
28	19	Private	168294	HS-grad		9	Never-mar	Craft-rep	Own-child	White	Male	0	0	40	United-St

The Text-to-columns data tool is mentioned under Data option.

## Data Selection

Data Selection is the selection of only that data which is necessary. When the Weka compatible file is run in Weka, some unnecessary data/attributes can be noticed that will not affect the whole model and is useless. This data needs to be removed before any further data pre-processing process. For Census Income Dataset, there were 15 attributes and not a single attribute was removed. This can be done by either selecting the attributes manually or by choosing filter RemoveUseless under unsupervised.

## After using RemoveUseless filter:

**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

**Filter**

Choose: **RemoveUseless - M 99.0** [Apply] [Stop]

**Current relation**

Relation: adultdata-weka.filters.unsupervised.attrib... Attributes: 15  
Instances: 32561 Sum of weights: 32561

**Attributes**

All | None | Invert | Pattern

No.	Name
1	Age
2	Workclass
3	fnlwgt
4	Education
5	Education - Num
6	Marital Status
7	Occupation
8	Relationship
9	Race
10	Sex
11	Capital-gain
12	Capital-loss
13	Hours-per-week
14	Native-country
15	y

Remove

**Selected attribute**

Name: Age  
Missing: 0 (0%)  
Distinct: 73  
Type: Numeric  
Unique: 2 (0%)

Statistic	Value
Minimum	17
Maximum	90
Mean	38.582
StdDev	13.64

Class: y (Nom) [Visualize All]

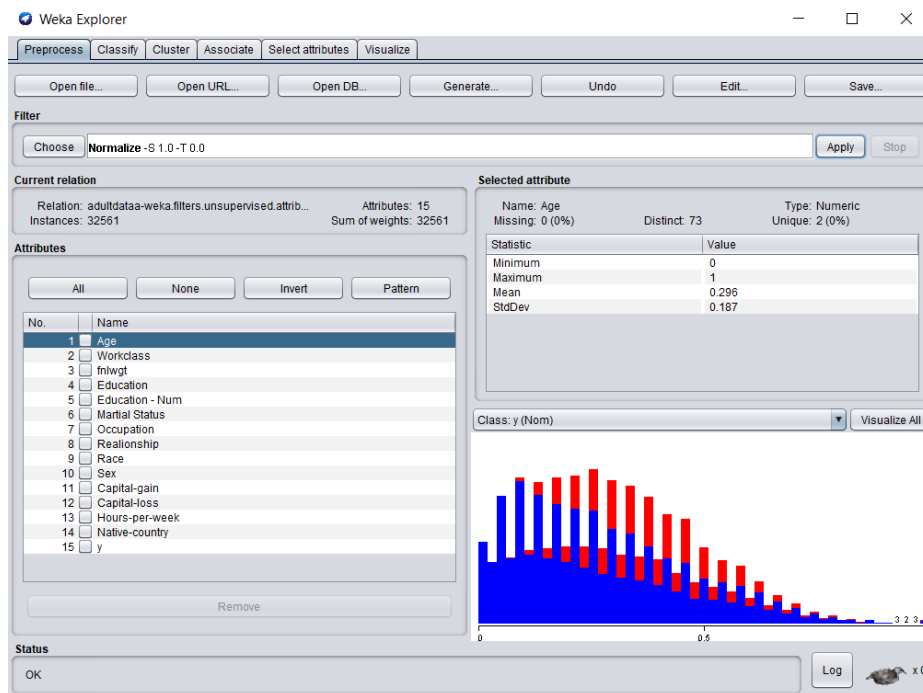
**Status**

OK [Log] x 0

## Data Normalization

Data Normalization is nothing but the process of transforming the number of variables of the data in a same range of 0 to 1. It is done in order to perform well and get better results.

### Data after normalization:



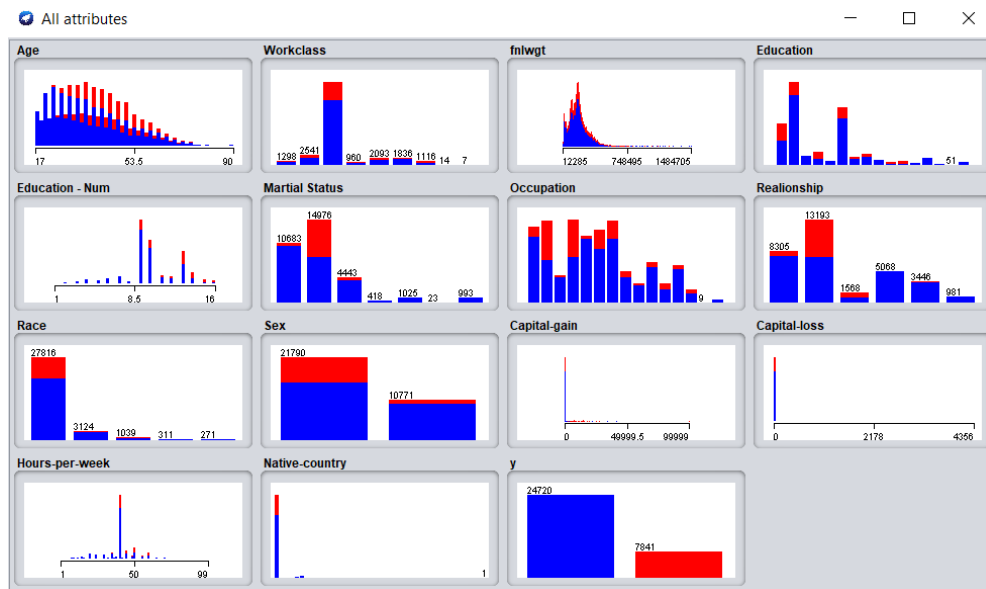
The data can be visualized separately with the help of visualize all option.



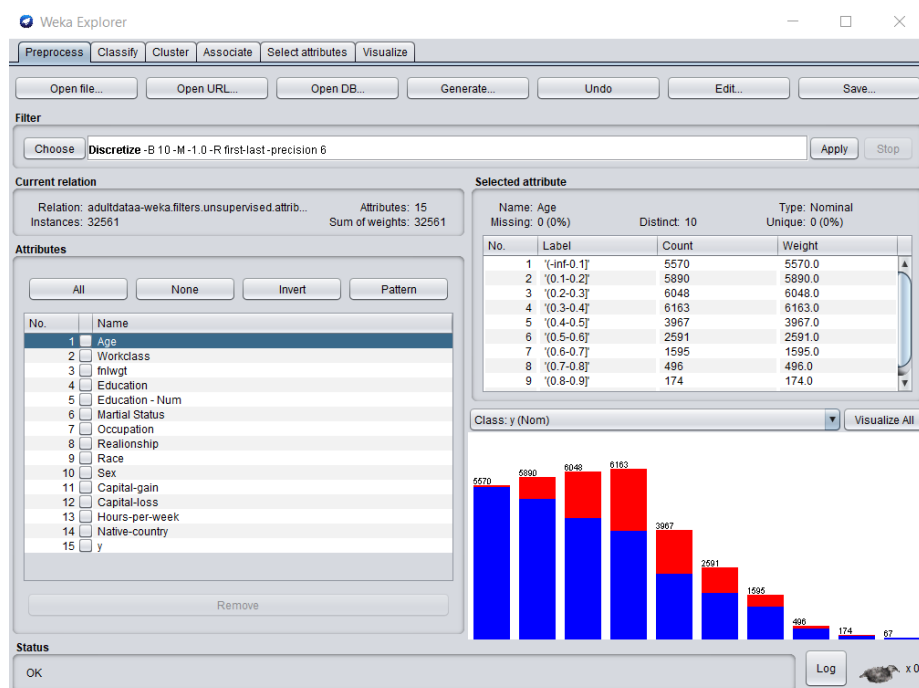
## Data Discretization

Data Discretization simplifies the data of the dataset. It divides the range of a continuous attribute into intervals. Some of the classification algorithms accept categorical attributes. Therefore, it reduces data size by discretization. Results are mentioned below:

### Visualization before discretization:



### After Discretization:



## Visualization of discretized data:



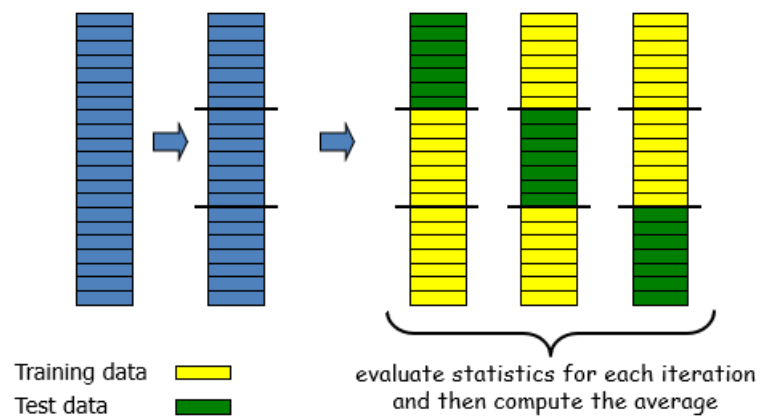
## Data Classification

Data Classification model creates a classification model to determine the class of provided data. This can be done by selecting options like percentage split or cross validation. Census Income dataset requires **True Positive** to be high which means it should predict that the classification tool should predict people who have annual income more than 50k. Also, the task is to avoid **False Negatives** which means failing to predict the number of people who has less income than 50k. Keeping this in mind, algorithms having the most **TP Rate** and **Precision** should be considered.

This can be done by either cross validation or percentage split

### Cross Validation

It divides the data into two halves not equal one. The training set and test data. The training set is use to train the model and test the data. Data is divided into k parts which is randomly set by the user.



### Percentage Split

It is a resampling method. It leaves out random N% of the original data. This can be explained by example. If we take 30 percent of data to be formed a test data then other 60 percent will be training data. And this will go again and again (Statistics - Resampling through Random Percentage Split, 2018).

Clustering for this dataset was not required as the goal which is needed to be achieved didn't required clustering.

Different types of classifications algorithms were used with 10 folds, 30 folds and 66.6 percent, percentage split. Even though there was no need to use percentage split but the result was slightly increased in some algorithms when tested. Thus, percentage split is also taken into consideration.

The algorithms are mentioned below in a well-structured manner and their build time is presented below.

Sr. No.	Algorithm	Folds	Time Taken to build model (sec)	Precision	TP Rate
1	OneR	10	0.09 seconds	0.829	0.782
		30	0.01 seconds	0.829	0.782
		Percentage split	0.02 seconds	0.831	0.787
2	BayesNet	10	0.17 seconds	<b>0.842</b>	0.819
		30	0.04 seconds	<b>0.843</b>	0.819
		Percentage split	0.03 seconds	<b>0.849</b>	<b>0.822</b>



3	ZeroR	10	0.01 seconds	?	0.759
		30	0.01 seconds	?	0.759
		Percentage split	0.01 seconds	?	0.765
4	Decision Table	10	2.98 seconds	0.826	0.835
		30	2.83 seconds	0.826	0.836
		Percentage split	0.01 seconds	0.833	0.842
5	Logistic	10	18.94 seconds	<b>0.846</b>	<b>0.852</b>
		30	20.23 seconds	<b>0.846</b>	<b>0.852</b>
		Percentage split	20.98 seconds	<b>0.846</b>	<b>0.852</b>
6	J48	10	0.43 seconds	0.836	0.843
		30	0.19 seconds	0.837	0.844
		Percentage split	0.19 seconds	<b>0.841</b>	<b>0.847</b>
7	NaiveBayes	10	0.04 seconds	<b>0.842</b>	0.819
		30	0.01 seconds	<b>0.842</b>	0.819
		Percentage split	0.01 seconds	<b>0.848</b>	0.822
8	NaiveBayesUpdatable	10	0.01 seconds	<b>0.842</b>	0.819
		30	0.01 seconds	<b>0.842</b>	0.819
		Percentage split	0.04 seconds	<b>0.848</b>	0.822
9	Bagging	Percentage split	0.03 seconds	0.834	0.839
10	ADABOOSTM1	Percentage split	0.01 seconds	0.795	0.804

These results are displayed with the preferred outcome (see Appendix for a table showing comparison of few classification functions).

**AdaboostM1** (see Appendix) and **Bagging** (see Appendix) is usually used for radiance reduction and performance increment but for the above comparison it can be clearly stated that bagging and adaboostM1 decreased the percentage of **TP Rate** (see Appendix) and **Precision** (see Appendix). So, both of these Meta filters are eliminated.

## Conclusion

After running all different types of classification algorithm, it could be clearly seen that no algorithm is perfect. Data selection should be done carefully. Ultimate goal is to remove redundancies and have an algorithm that produces fasted results and with accuracy. This is

because if the algorithm ends up being not efficient then it high crash for the big test dataset. Initially, Logistic was selected as the best algorithm with precision of 84 percent and TP rate of 85 percent. Only reason to not select this algorithm is because it took approximately 19 seconds to build the model, so if a test data will be given to it; it will take a lot of time or it might crash too. Thus, NaiveBayes was selected as a good algorithm out of all the algorithms that have been tested on Census Income dataset with the precision of 84 percent, TP rate of 81 percent and taking minimum building time for the model of 0.1 seconds. The selection of best algorithm is based both on speed of how fast the model is and how précised the model is.

## References

- Becker, R. K. (1996, 05 01). *Census Income Data Set*. Retrieved from UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Census+Income>
- Statistics - Resampling through Random Percentage Split*. (2018, June 05). Retrieved from Gerardnico: [https://gerardnico.com/data\\_mining/validation\\_set](https://gerardnico.com/data_mining/validation_set)
- William J Frawley, G. P.-S. (n.d.). *Introduction to Data Mining*. Retrieved from Data Mining: [http://www.cs.ccsu.edu/~markov/ccsu\\_courses/datamining-1.html](http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-1.html)

# Appendix

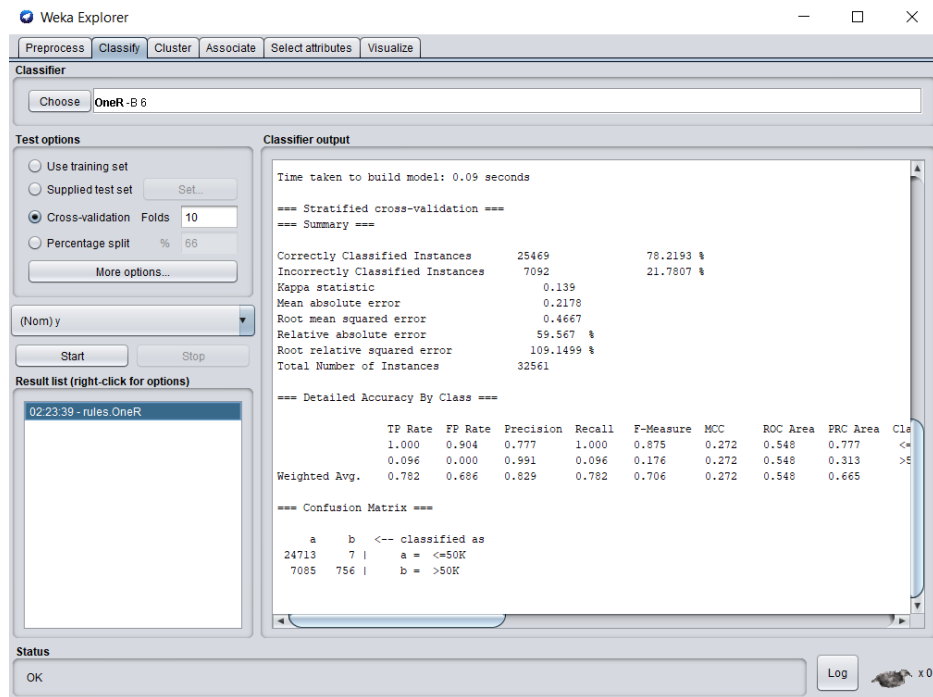


Figure 1: OneR- 10-Fold cross-validation.

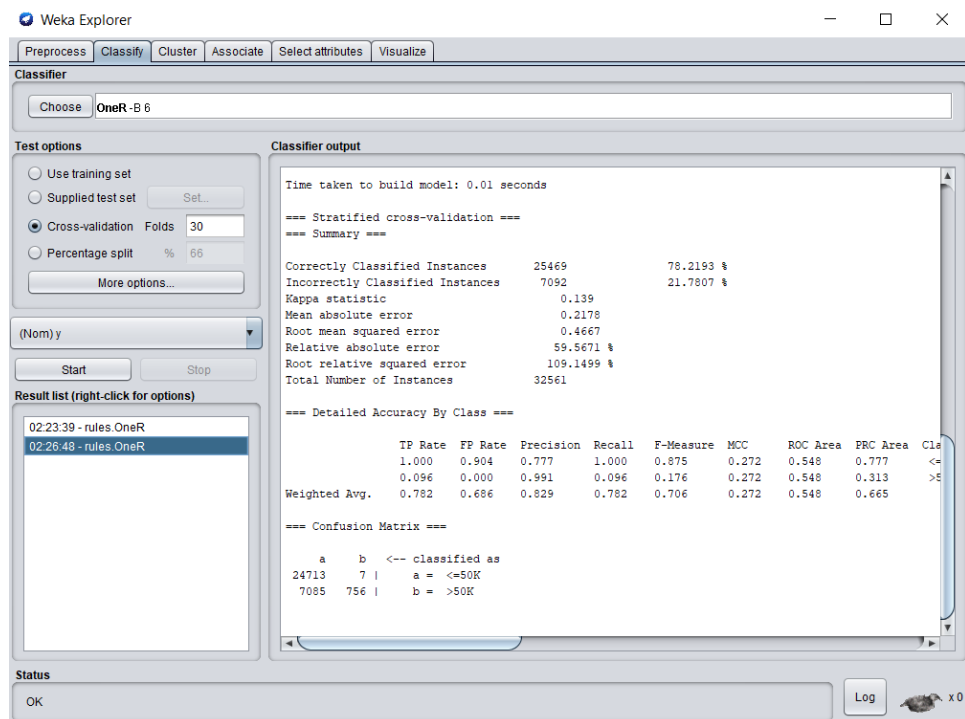


Figure 2: OneR- 30-Fold cross-validation.

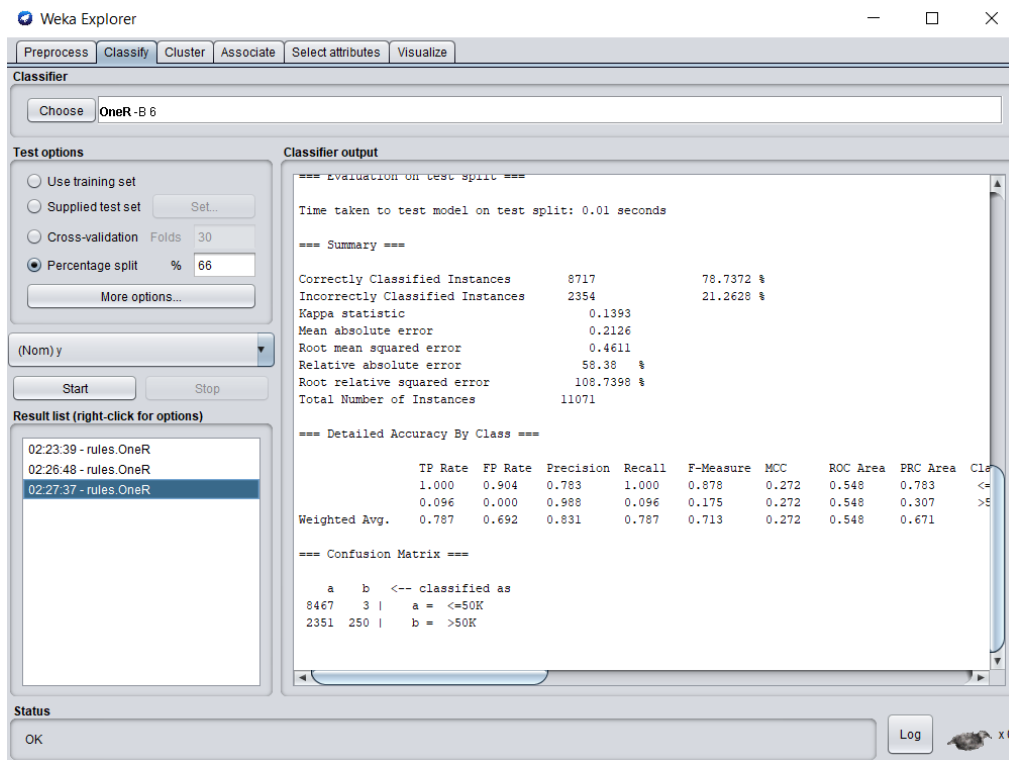


Figure 3: OneR- Percentage split.

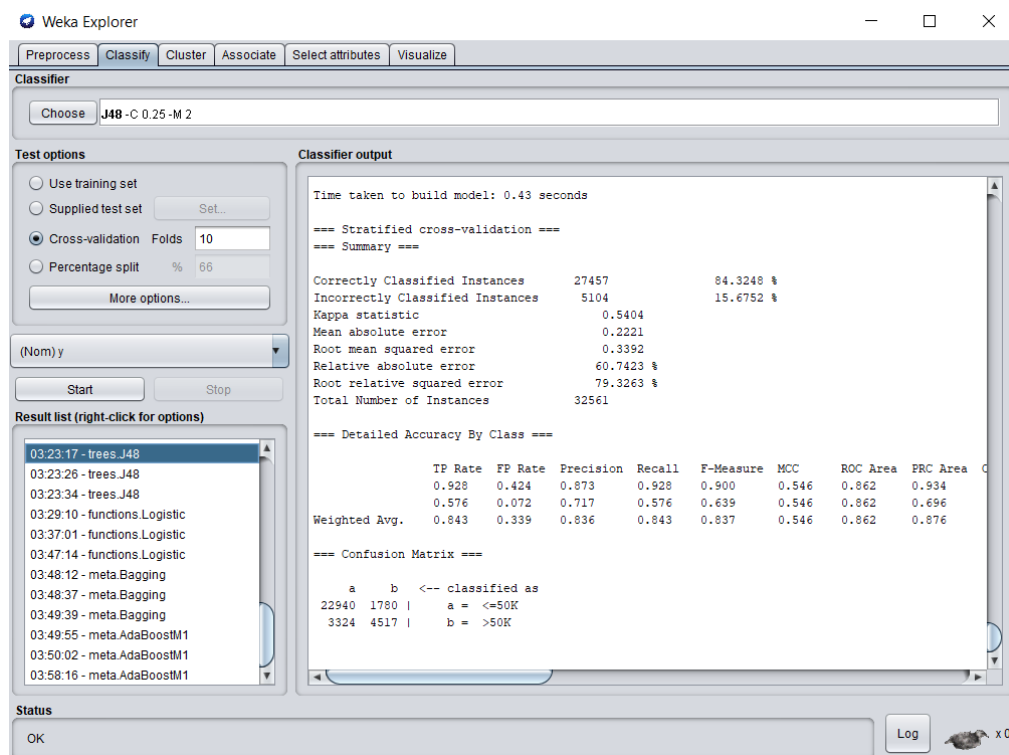


Figure 4: J48- 10-Fold cross-validation.

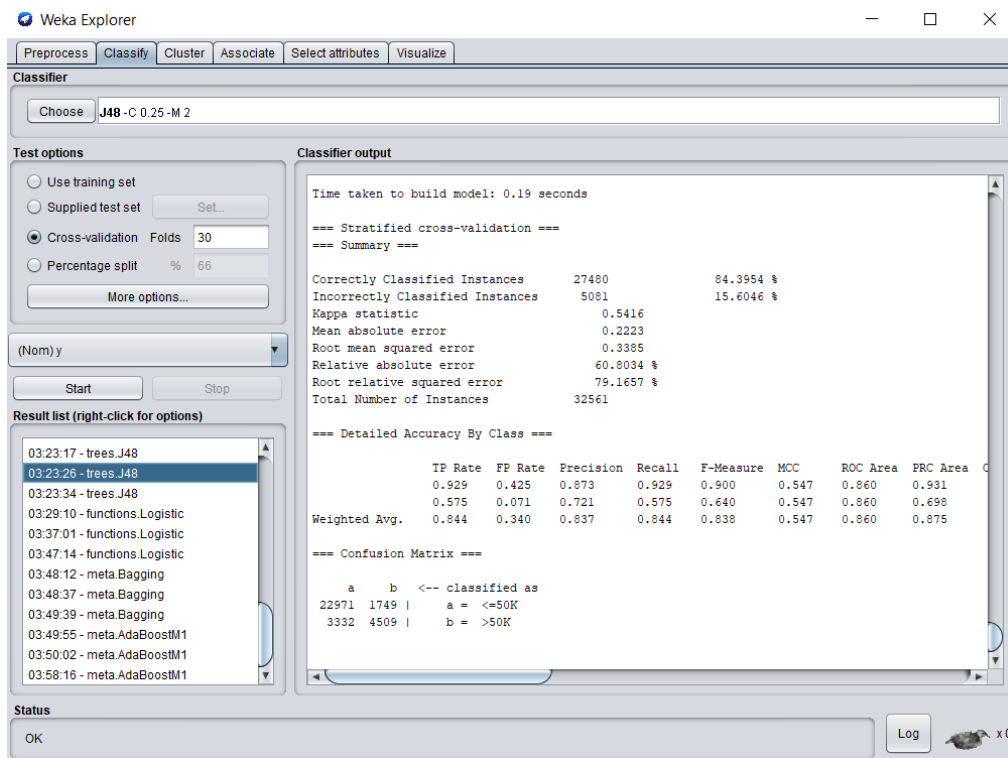


Figure 5: J48- 30-Fold cross-validation.

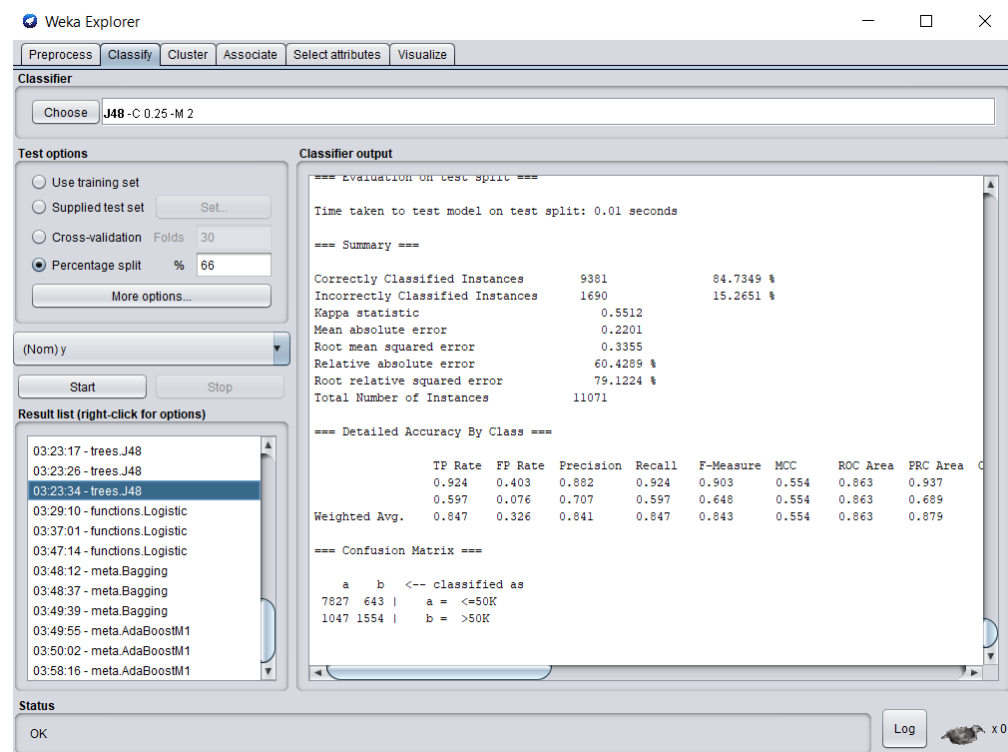


Figure 6: J48- Percentage split.

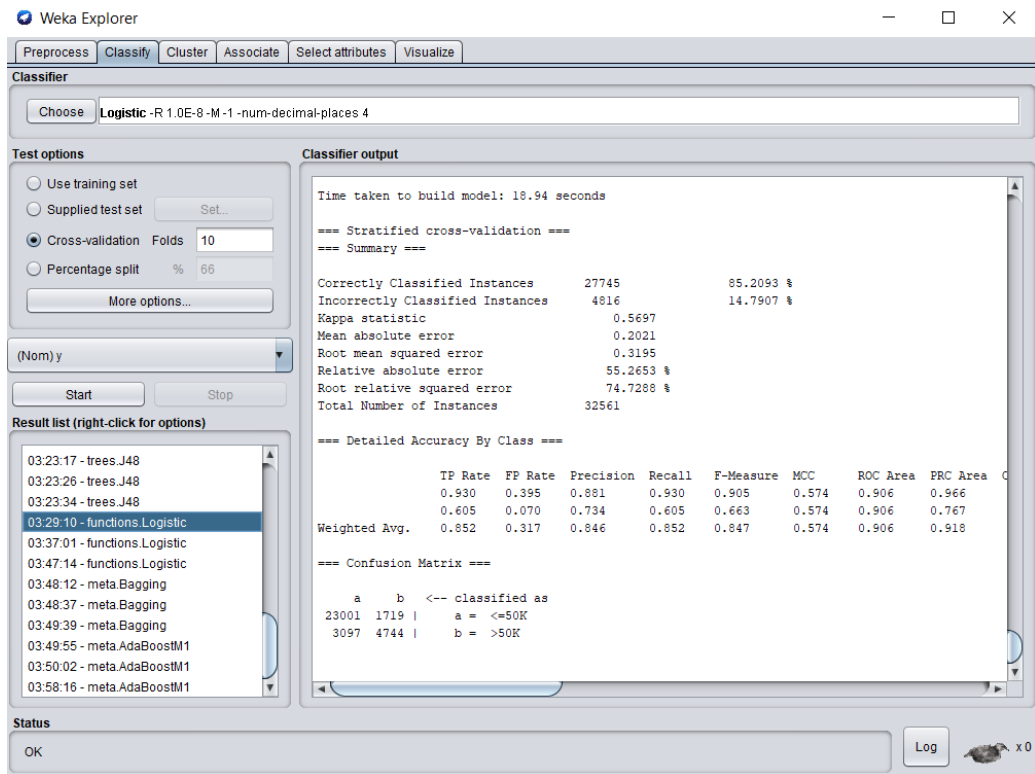


Figure 7: Logistic- 10-Fold cross-validation.

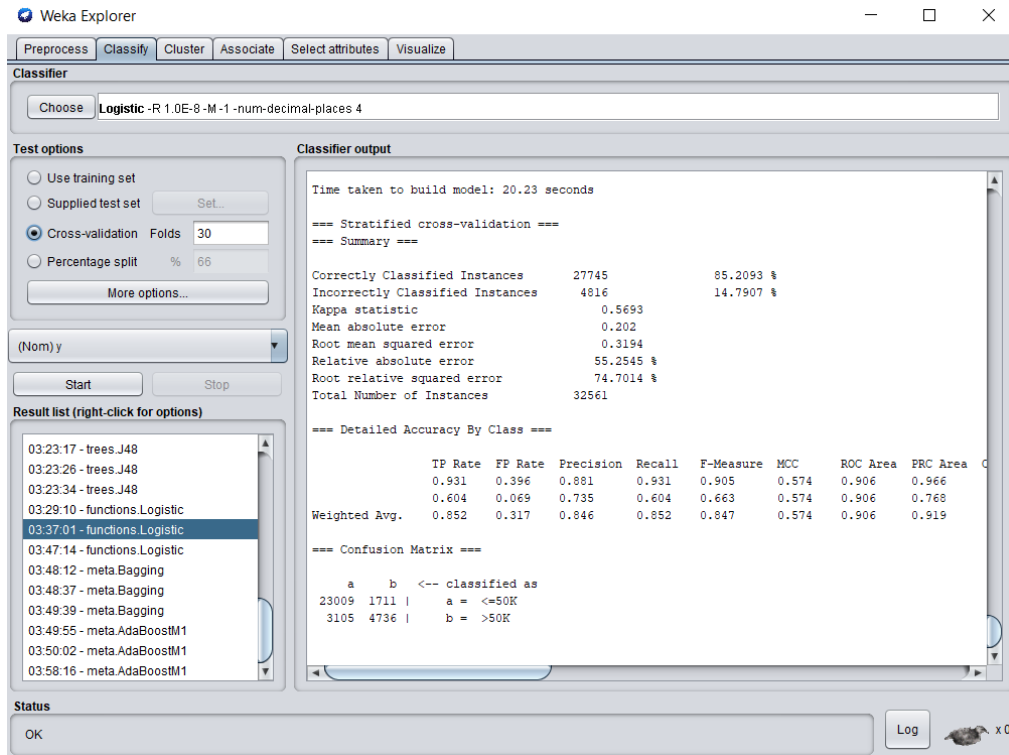


Figure 8: Logistic- 30-Fold cross-validation.

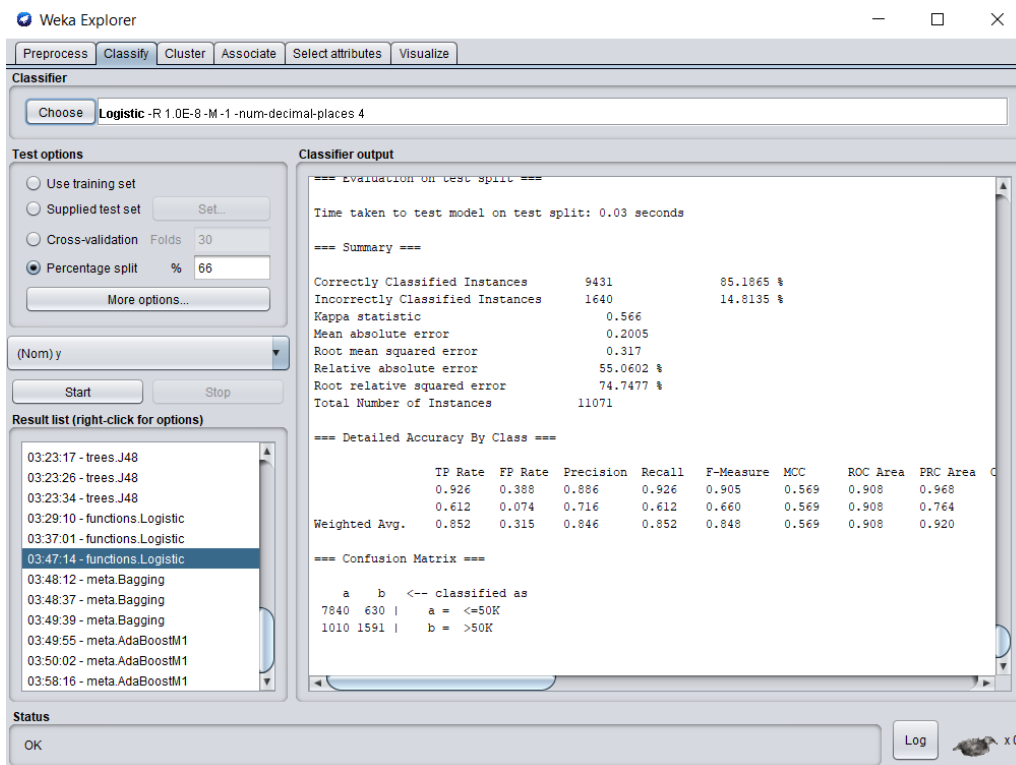


Figure 9: Logistic- Percentage split.

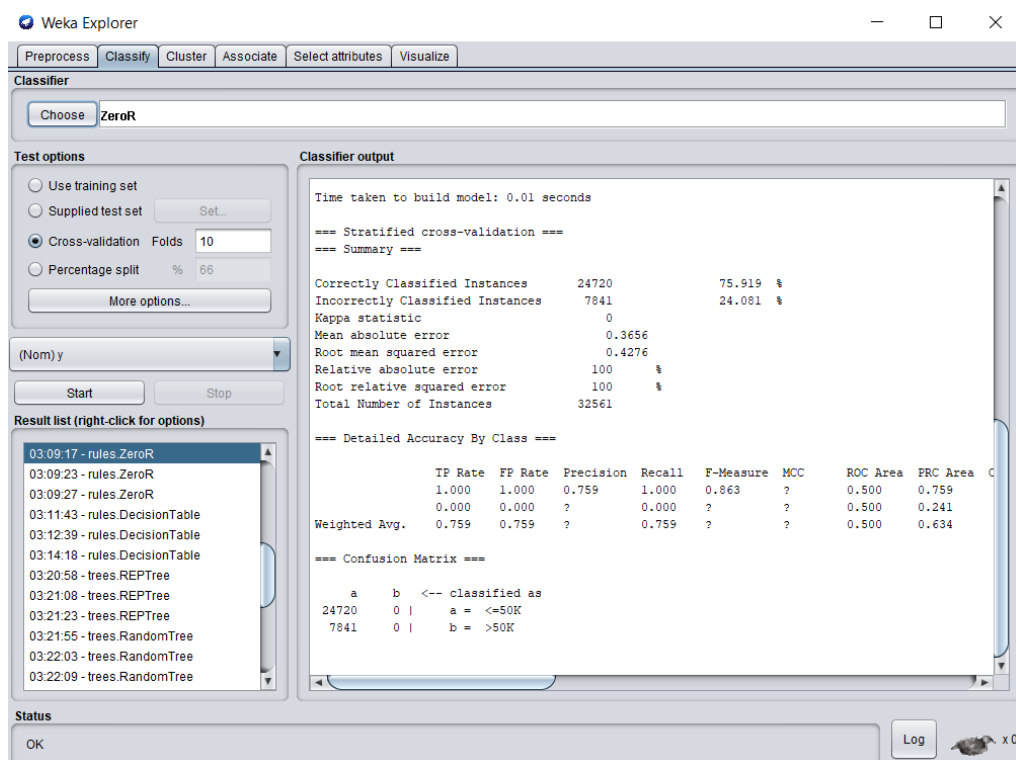


Figure 10: ZeroR- 10-Fold cross-validation.

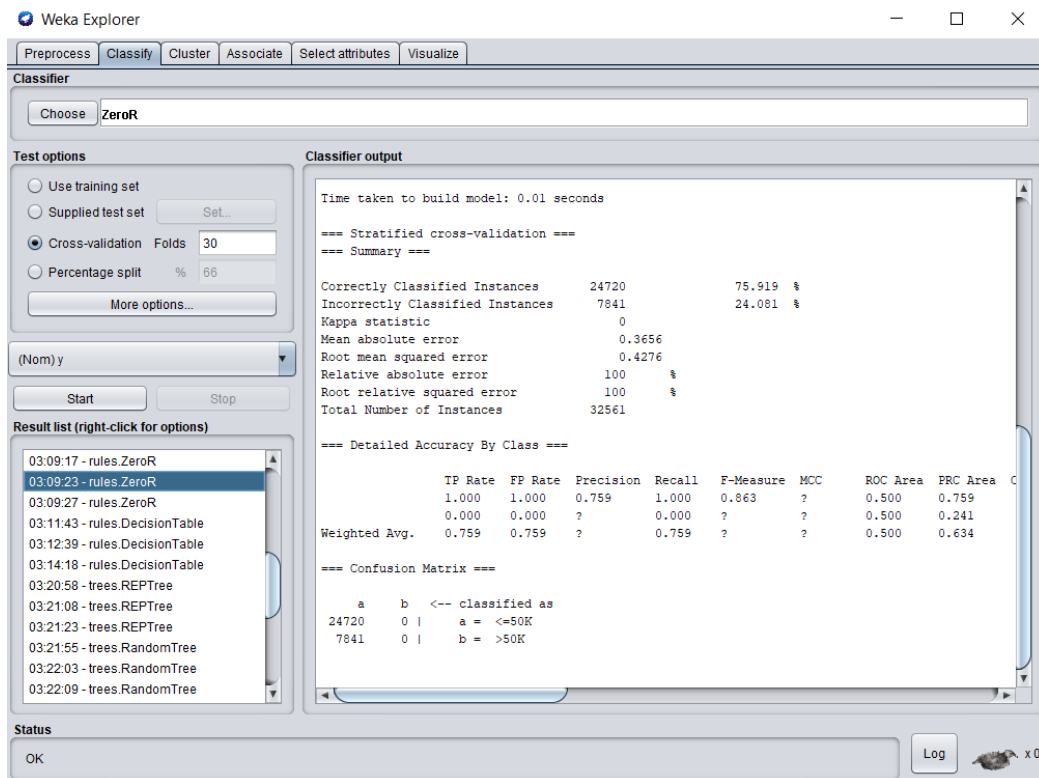


Figure 11: ZeroR- 30-Fold cross-validation.

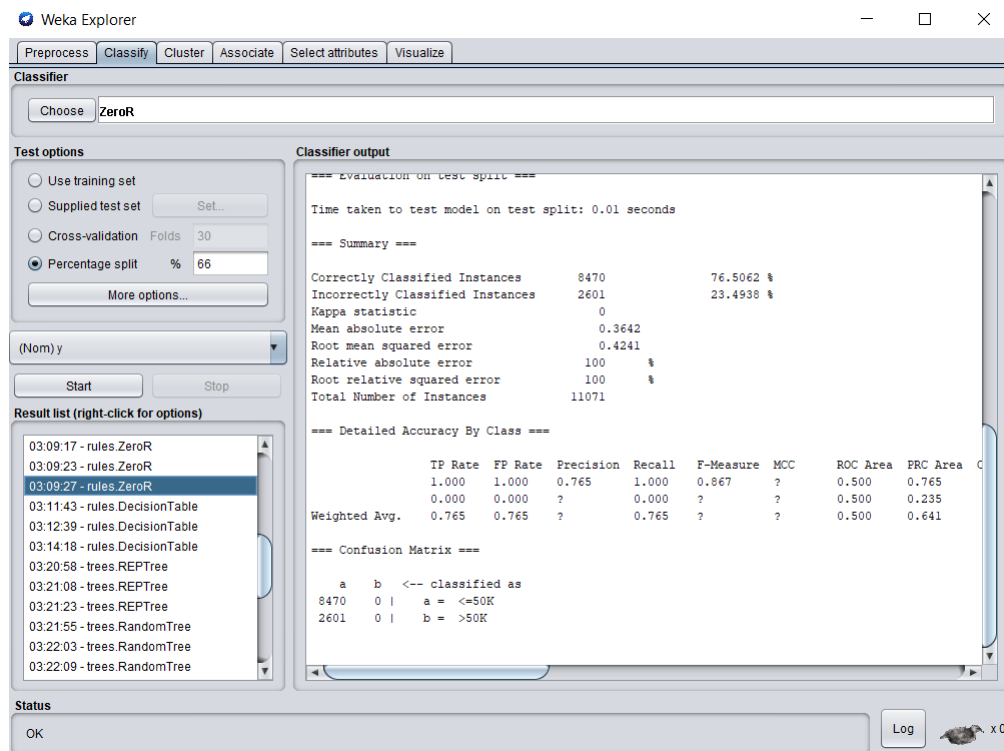


Figure 12: ZeroR- Percentage split.



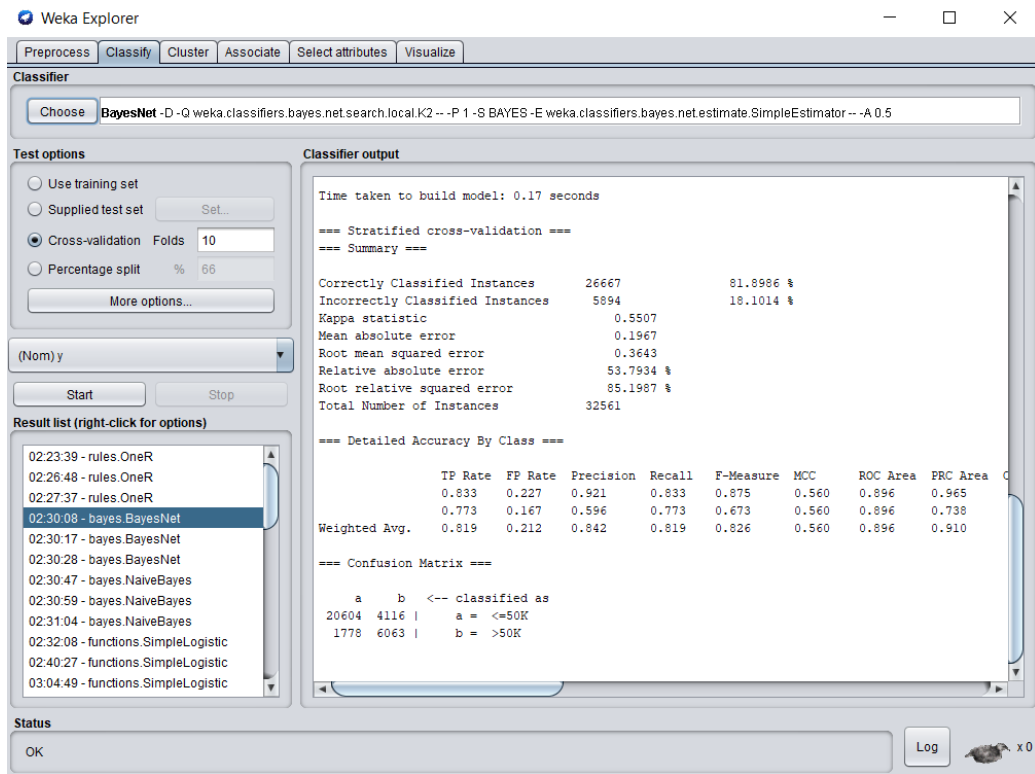


Figure 13: BayesNet- 10-Fold cross-validation.

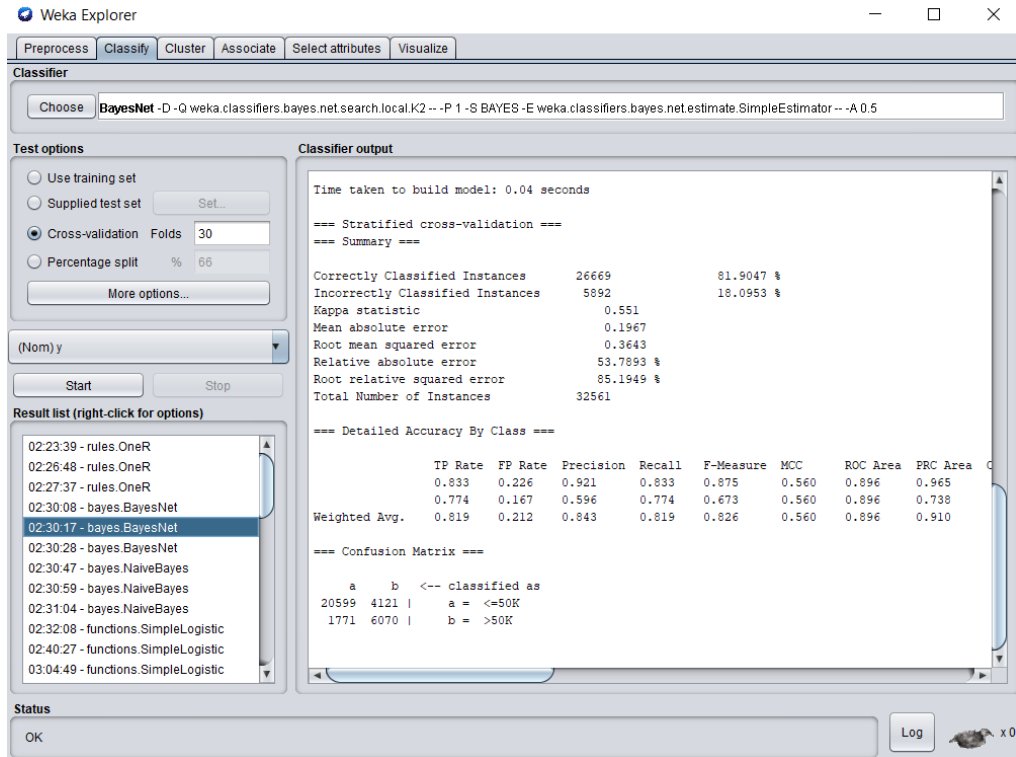


Figure 14: BayesNet- 30-Fold cross-validation.

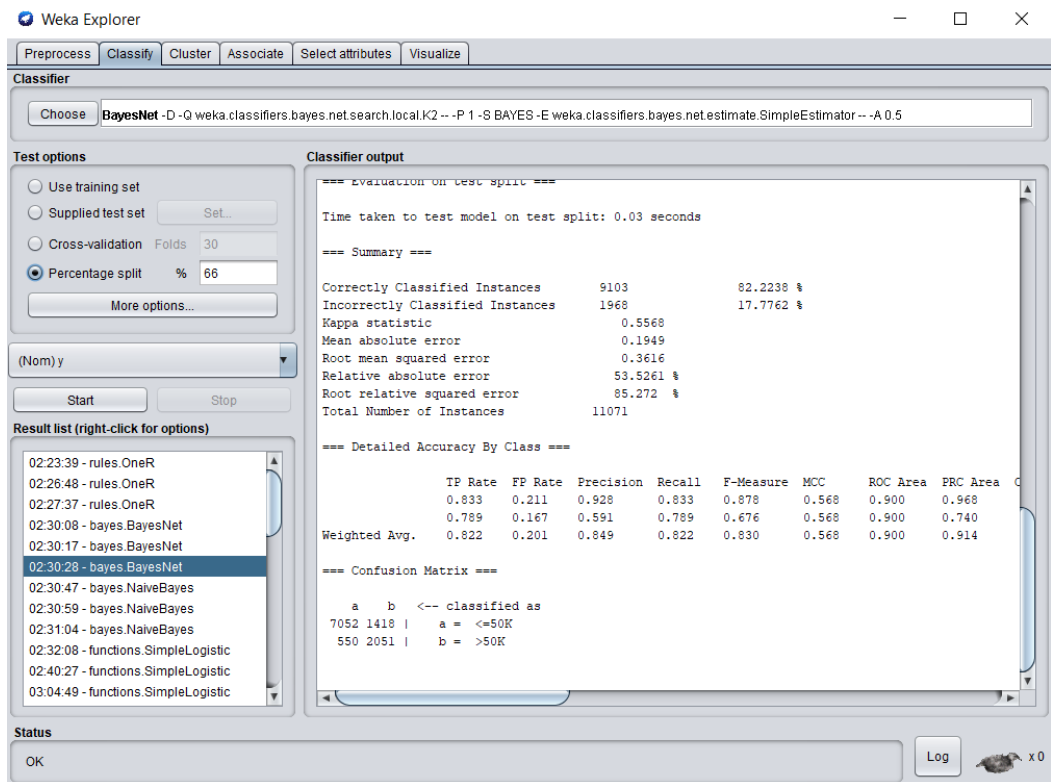


Figure 15: BayesNet- Percentage split.

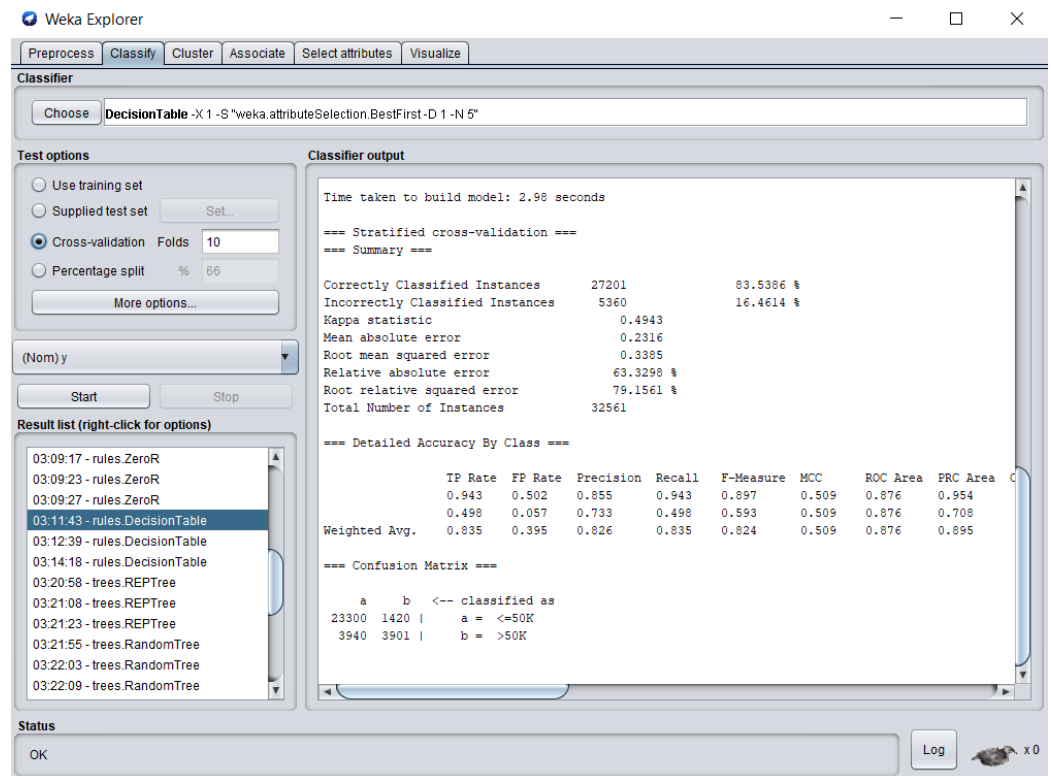


Figure 16: Decision Table- 10-Fold cross-validation.

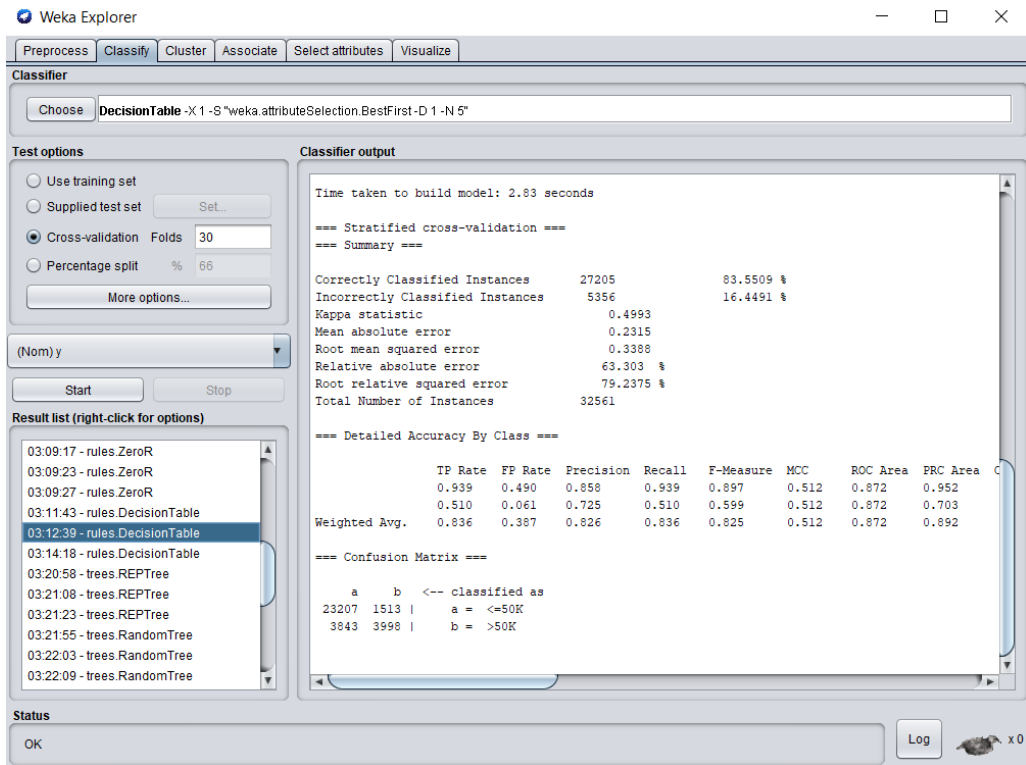


Figure 17: Decision Table- 30-Fold cross-validation.

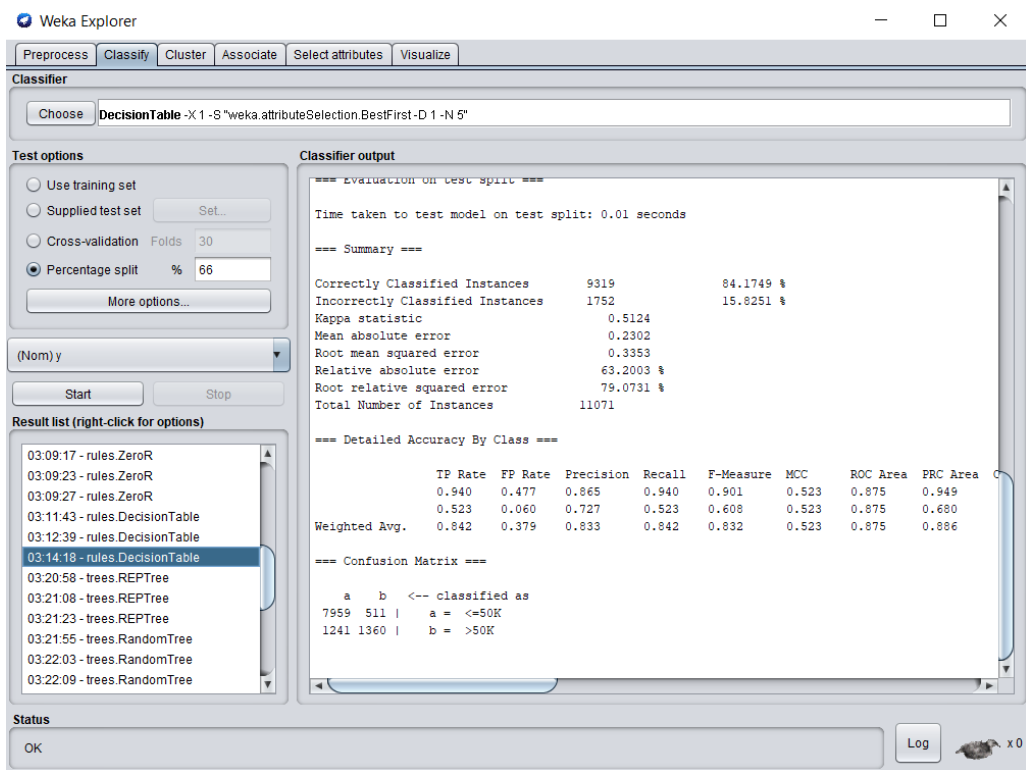


Figure 18: Decision Table- Percentage split.

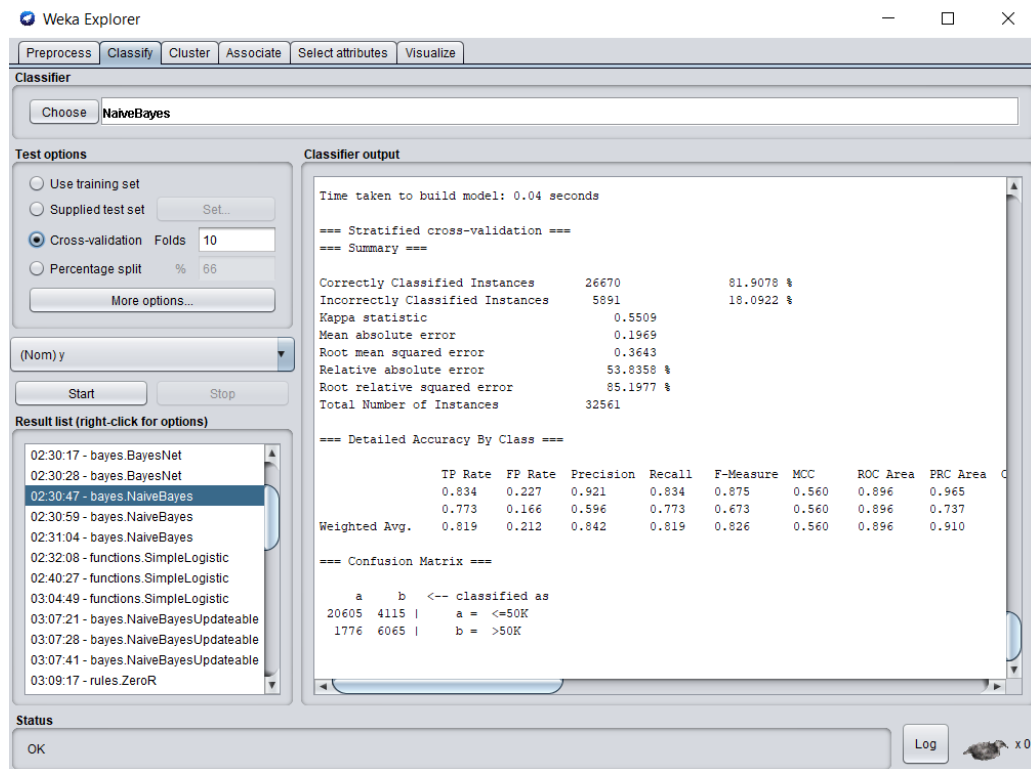


Figure 19: NaiveBayes- 10-Fold cross-validation.

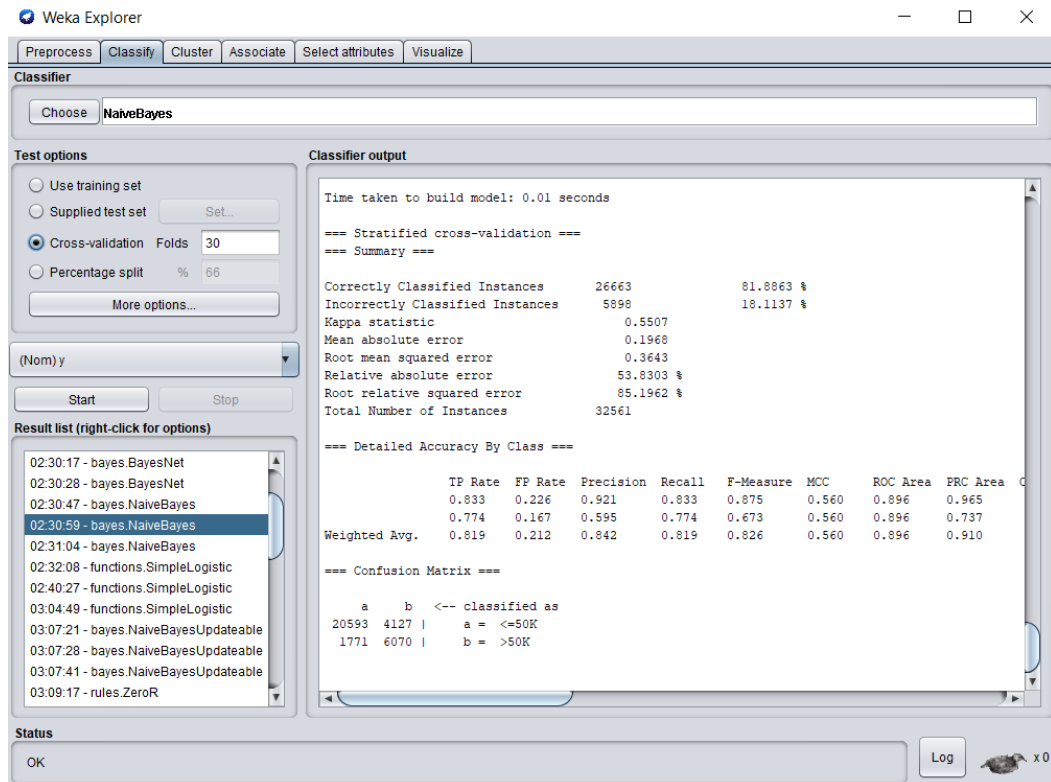


Figure 20: NaiveBayes- 30-Fold cross-validation.

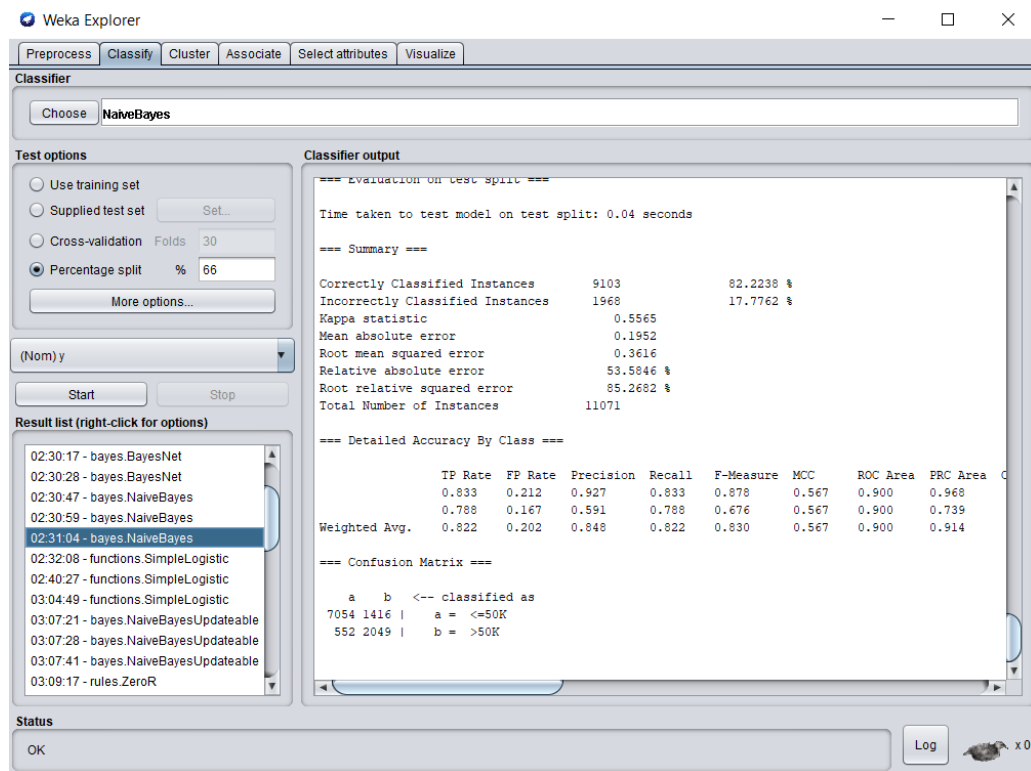


Figure 21: NaiveBayes- Percentage split.

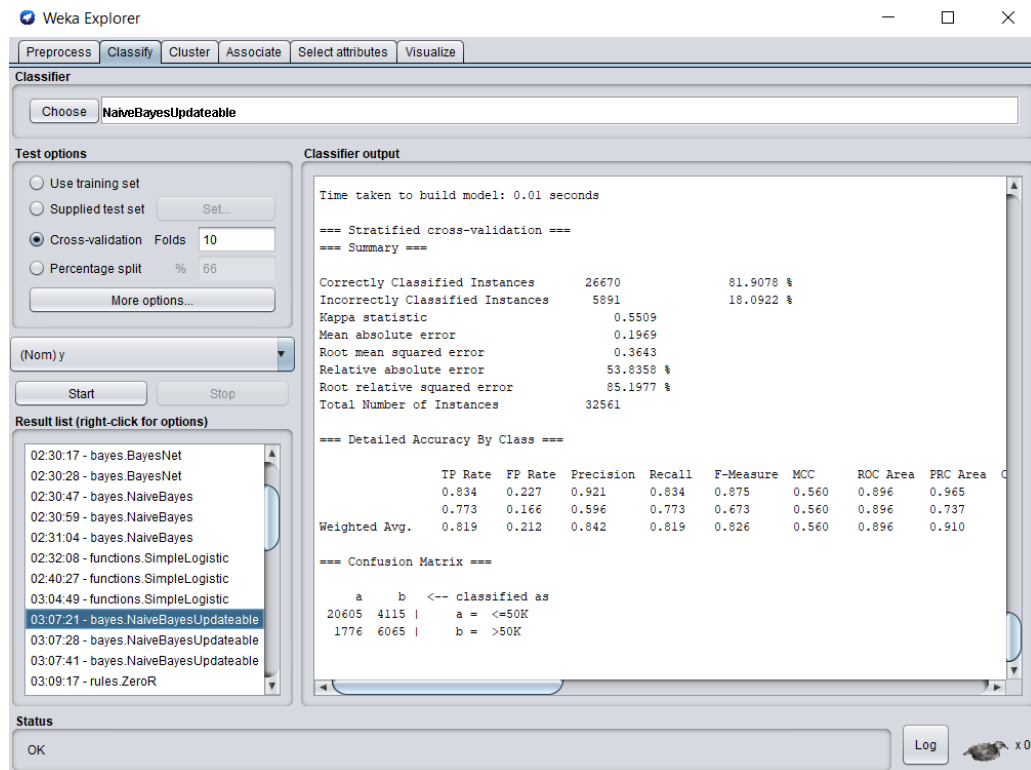


Figure 22: NaiveBayesUpdateable- 10-Fold cross-validation.

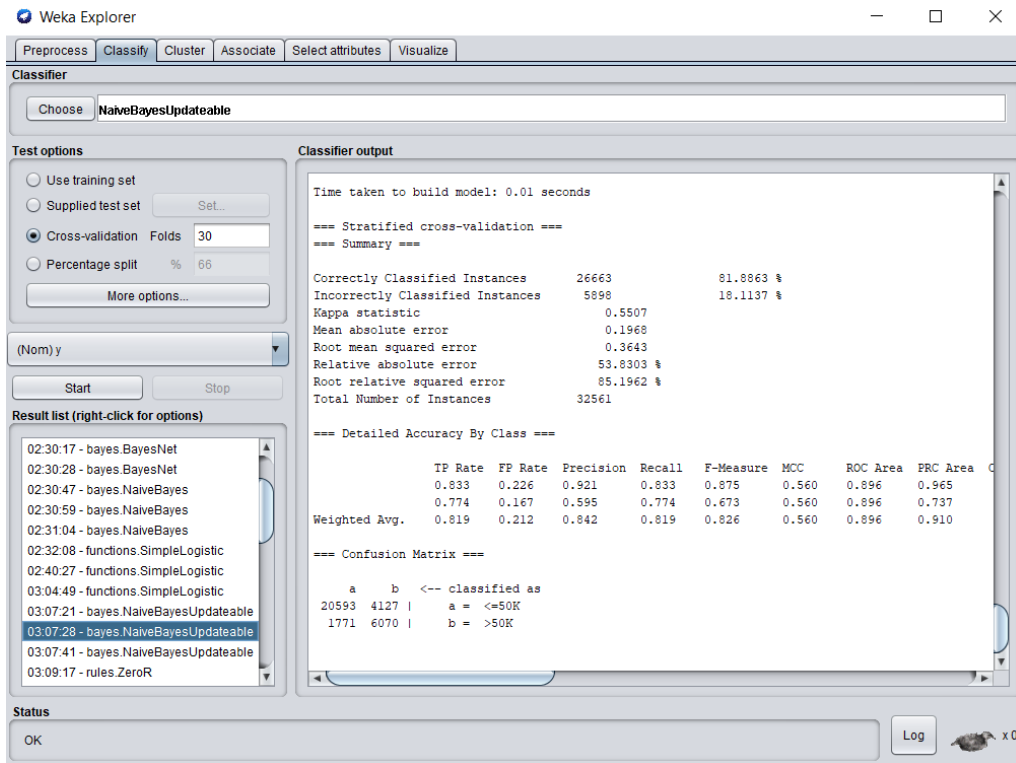


Figure 23: NaiveBayesUpdatable- 30-Fold cross-validation.

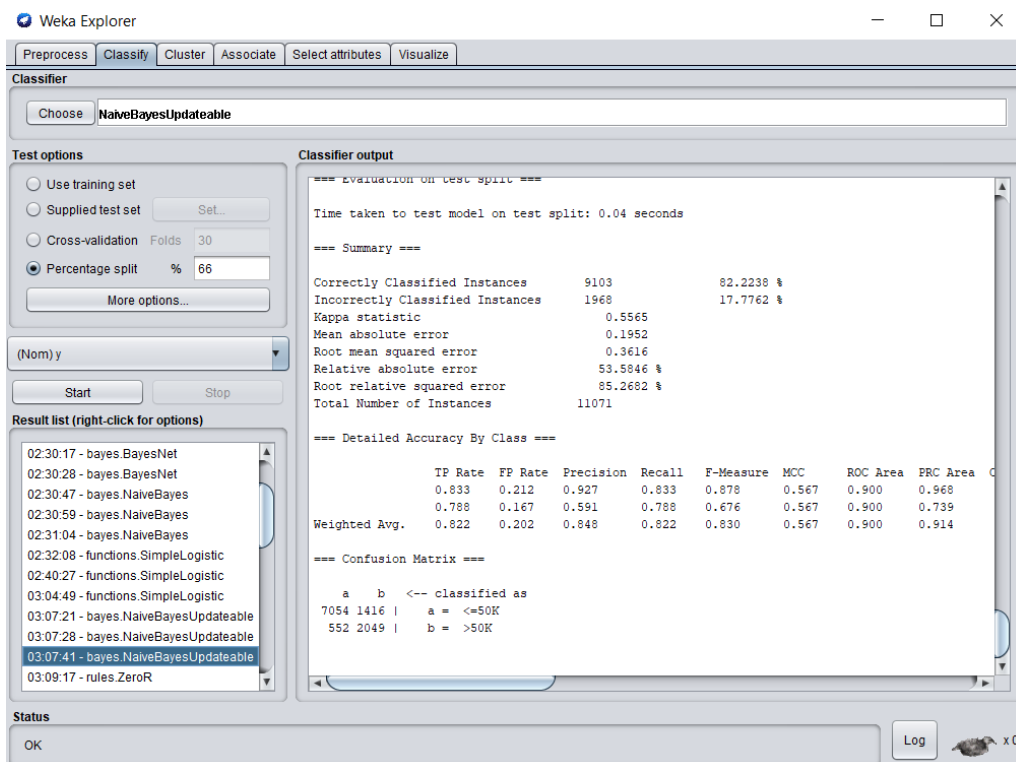


Figure 24: NaiveBayesUpdatable- Percentage split.

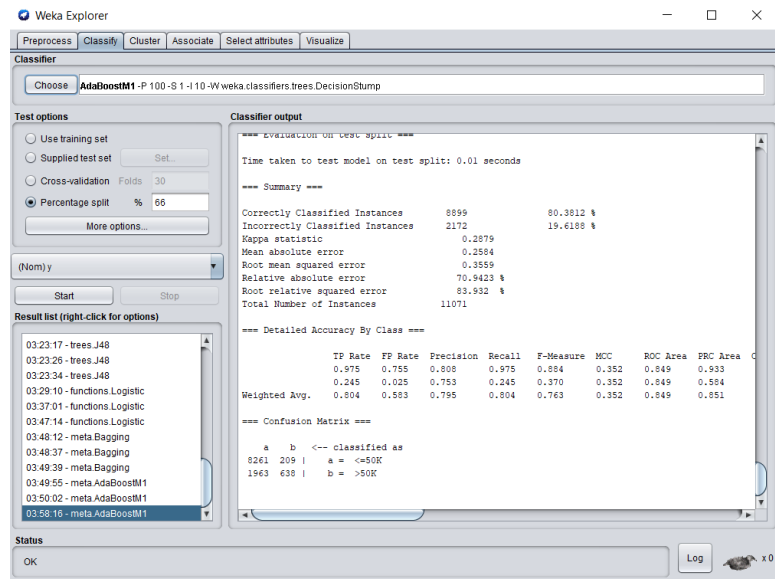


Figure 25: AdaBoostM1- Percentage split.

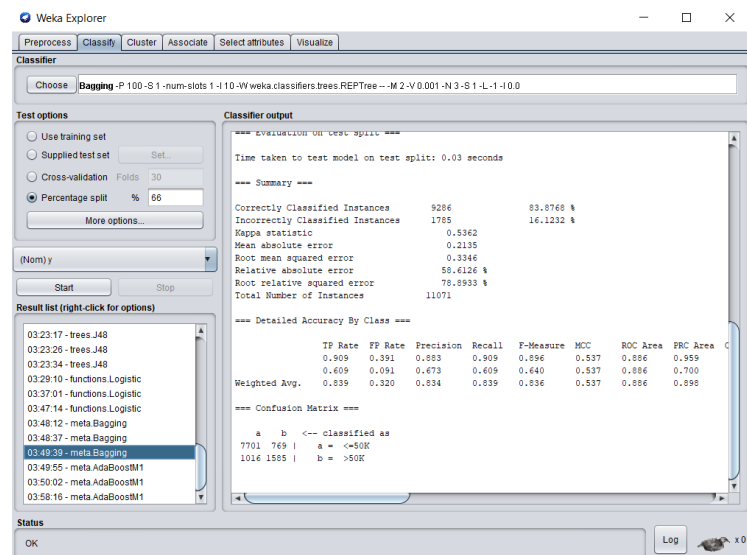


Figure 26: Bagging- Percentage split.

*Boosting* is a Classification algorithm which lies under meta section is used to decrease bias in datasets (Page 17).

*Bagging* is a classification algorithm which lies under meta section is used to decrease variance in datasets (Page 17).

*Precision* can be well defined as proportion of cases classified as positive are indeed positive (Page 17).

*True Positive* can be defined as something that correctly indicates passing (Page 17).