

Introduction:

Data mining is an interdisciplinary subfield of computer science and measurements with a complete area to abstract information (with intelligent method) from a data set known as classification and convert the information into a logical structure for future known as prediction. There are many several classification techniques used in data mining such as KNN, Naïve Bayes, Decision Tree (ID3, CART), Association Rule and Clustering.

Among of them I choose Decision Tree (ID3) classification technique for classify and predicting University Admission Result (Pass or Fail) based on the GPA of JSC, SSC & HSC (customized data set).

Decision Tree:

Decision tree is the most prevailing and standard tool for classification and prediction. It is a classification technique because here class attribute always hold categorical information. It is also a supervised learning algorithm because it has a pre-defined target variable and they are mostly secondhand in non-linear decision making with simple linear decision surface. Decision Tree algorithm related with three key components such as Nodes, Links and Leaves. It ropes both numerical and categorical data to build the decision tree.

Reason for choosing Decision Tree:

Associating to other algorithms Decision Tree involves less effort, less time and less analysis for data preparation in pre-processing. It is secondhand for both binary and multiclass classification problem. Decision Tree defines the situations and activities that allow the forecaster to recognize the authentic decisions that must be made. However it doesn't need normalization and scaling of data as well. The main benefit is mislaid values in the data don't distress the process of building a decision tree to any considerable level. Also the notion behind Decision Tree is more aware to programmers and fairly easier to recognize than other similar algorithms. Moreover it is competent for huge data set with fewer complexity.

How does ID3 work for Decision Tree:

ID3 practices a greedy approach that's why it does not surety an optimal solution; it can catch wedged in local targets. ID3 can over-fit to the training data (to escape overfitting, minor decision trees should be ideal over larger ones). This algorithm typically harvests small trees, but it does not always produce the minimum possible tree. The algorithm's optimality can be upgraded by using backtracking during the search for the best decision tree at the rate of conceivably taking longer. ID3 can over-fit the training data. To evade overfitting, lesser decision trees should be preferred over larger ones.

Reason for choosing dataset:

We select a customized general dataset containing the attributes of Registration no, Name, Date of birth, JSC GPA, SSS GPA, HSC GPA and result of the university admission test (Pass or Fail). Because this dataset is a combination of numerical and categorical attributes which is perfect for

Decision Tree classification. This dataset also contains nominal attributes as well. The attributes type is written below,

- Reg no = nominal attributes (unique)
- Name = nominal attributes (unique)
- DoB = nominal attributes (unique)
- Gender = categorical attributes
- JSC = numerical attributes
- SSC = numerical attributes
- HSC = numerical attributes
- Result = categorical attributes (class attributes)

We can easily build the classification model by using these data set and apply that model for predicting the result in other data set. Moreover in our country's public university admission test, the authority of the university always counting a portion of SSC GPA and HSC GPA. Here we also include the GPA of JSC as well, so that we can easily classify how much GPA was required to pass university admission test and also predict the result (either Pass or Fail) in other data set as well. Testing data set also includes the same attributes with different types of value but we have to find the admission test result of the testing data set based on the model of the general dataset. Our selected customized general dataset is given below.

	A	B	C	D	E	F	G	H
1	Regno	Name	DoB	Gender	JSC	SSC	HSC	Result
2	1548	Nyme	9.11.1999	Male	4.75	5	4.5	Pass
3	1364	Himu	8.11.1998	Male	4.9	4.75	4.8	Pass
4	8614	Rifat	30.11.1998	Male	5	4.5	4.6	Fail
5	3648	Tonu	20.11.1999	Male	5	4.4	4.2	Pass
6	8426	Golam	1.12.1998	Male	4.8	5	4.4	Fail
7	1493	Akib	8.12.1999	Male	5	5	4.7	Pass
8	2064	Nila	20.5.1998	Female	4.75	5	5	Pass
9	1820	Woishe	22.5.1998	Female	4.8	4.5	4.6	Fail
10	6596	Ali	5.3.1999	Male	5	4.8	4.9	Pass
11	6931	Saim	3.11.1999	Male	4.3	4.5	4.8	Fail
12	1205	Sujana	15.7.1999	Female	5	5	4.8	Pass
13	5843	Joy	7.3.1998	Male	4.5	4.9	5	Pass
14	2064	Momo	15.9.1999	Female	4.8	4.7	4.6	Pass
15	1584	Rimi	25.10.1999	Female	5	4.6	4.7	Fail
16	6942	Nobel	26.7.1998	Male	5	4.6	4.4	Fail
17	2605	Bristy	6.9.1999	Female	5	5	4.4	Pass
18	6840	Methila	28.11.1999	Female	5	4.6	4.6	Fail
19	5642	Rafa	25.12.1998	Female	4.6	4.75	4.5	Pass
20	8563	Rimu	17.5.1998	Male	4.8	4.6	4.9	Pass
21	1504	Shifat	20.4.1999	Male	4.8	4.75	5	Pass
22	8426	Shifa	4.5.1998	Female	5	4.5	4.9	Pass
23	1568	Dihan	7.9.1998	Male	4.5	5	4.5	Fail
24	7523	Jannat	5.6.1998	Female	4.8	4.9	4.4	Fail
25	8436	Farhab	1.5.1999	Male	5	4.8	4.6	Pass
26	7296	Sozol	25.11.1998	Male	4.7	4.7	4.6	Fail
27	2650	Sonia	30.6.1999	Female	4.6	4.9	4.5	Fail
28	1583	Liza	4.8.1999	Female	4.7	5	4.8	Pass
29	4628	Nupur	5.9.1998	Female	5	4.8	4.5	Fail
30	9542	Abid	13.12.1998	Male	4.6	5	4.4	Fail
31	1266	Munna	19.10.1999	Male	4.8	4.7	4.7	Pass

Figure 01: General Dataset (training data set)

This is our general dataset. This is also known as training set. There are 8 attributes in this dataset. Among them Reg no is numerical unique attributes. Besides Name and DoB are nominal unique attributes. Gender is a nominal attributes carrying two category {Male, Female}. JSC, SSC, HSC are numerical attributes. And Result is the nominal and categorical class attribute {Pass, Fail} and also our target variable.

The predicting data set is given below.

	A	B	C	D	E	F	G	H
1	Reg no	Name	DoB	Gender	JSC	SSC	HSC	Result
2	526912	Rahim	9.11.1999	Male	4.9	4.7	4.8	?
3	843645	Karim	8.11.1998	Male	4.6	4.8	5	?
4	512648	Shimul	30.11.1998	Male	4.7	5	4.6	?
5	841266	Faria	20.11.1999	Female	4.6	4.6	5	?
6	542582	Sayem	1.12.1998	Male	5	4.8	4.6	?
7	845631	Rafat	8.12.1999	Male	4.5	4.6	4.65	?
8	452175	Maria	20.5.1998	Female	4.8	4.9	4.9	?
9	478963	Munni	22.5.1998	Female	4.6	5	4.8	?
10	154782	Nusrat	5.3.1999	Female	4.4	4.7	5	?
11	569210	Lamya	3.11.1999	Female	5	4.4	4.5	?
12	412560	Imran	15.7.1999	Male	4.6	4.8	5	?
13	266940	Afridi	7.3.1998	Male	4.7	4.6	4.9	?
14	126436	Borsha	15.9.1999	Female	4.9	5	4.7	?
15	158340	Nourin	25.10.1999	Female	4.8	4.9	5	?
16	230548	Rubel	26.7.1998	Male	4.65	4.5	4.6	?
17	356046	Sana	6.9.1999	Female	4.9	4.65	5	?
18	520047	Fariya	28.11.1999	Female	4.6	5	4.7	?
19	953485	Tussy	25.12.1998	Female	5	4.7	4.8	?
20	588445	Rajon	17.5.1998	Male	4.8	5	4.6	?
21	269425	Syed	20.4.1999	Male	5	4.8	4.9	?

Figure 02: Dataset for prediction (testing data set)

By classifying the first dataset, we are going to predict the Result of second dataset by creating the model of Decision Tree classification. The second dataset is also known as testing data set. These testing data set have the same attributes as the training dataset.

Procedure of Weka for classification and prediction:

- I. Open the Weka and click 'Explorer' application. Then click 'open file' and choose first dataset (Figure 01) to import for classification. Note that the extension of first data set must be .csv format.

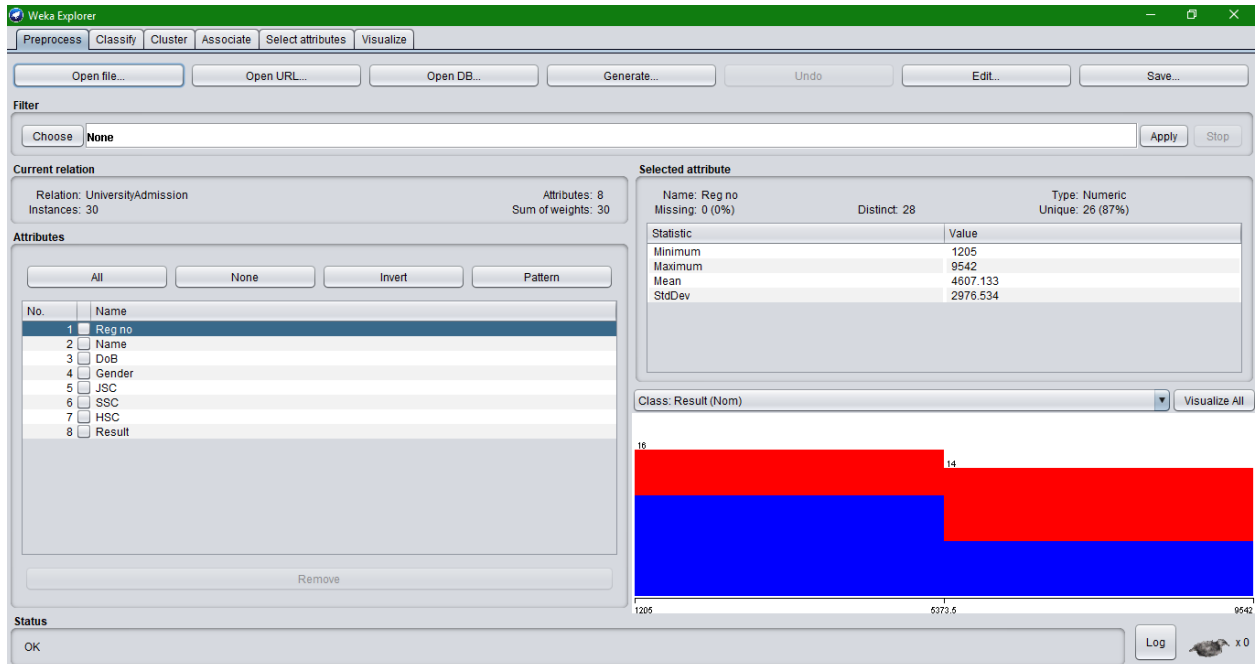


Figure 03: After importing the first dataset

- II. After importing the first dataset we see that total eight attributes are arranged serially in the left side. We can individually check each attributes status in the right side as well by selecting the attributes. We may also remove any attributes if they are not necessary for our classification by clicking them. Here Reg no, Name, DoB are unique attributes and they doesn't affect our classification. So, we may remove these attributes for classification.

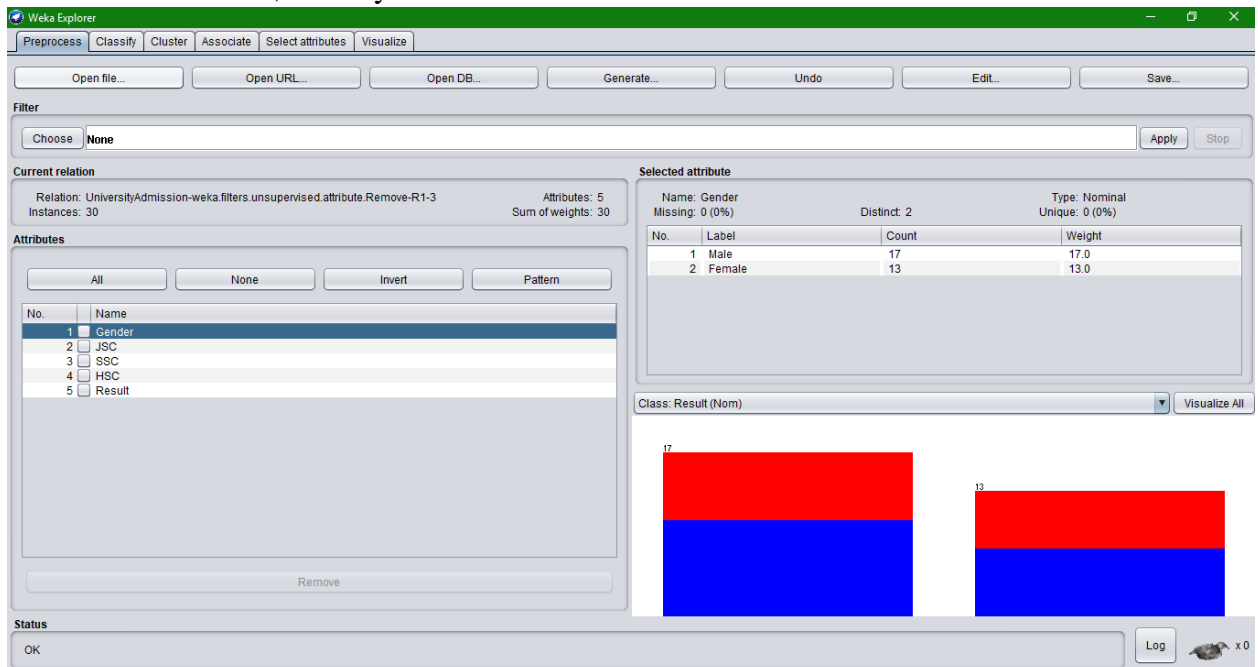


Figure 04: After removing the unique attributes for classification

- III. Go to classify section and choose J48 under trees section for decision tree. Here Decision tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3).

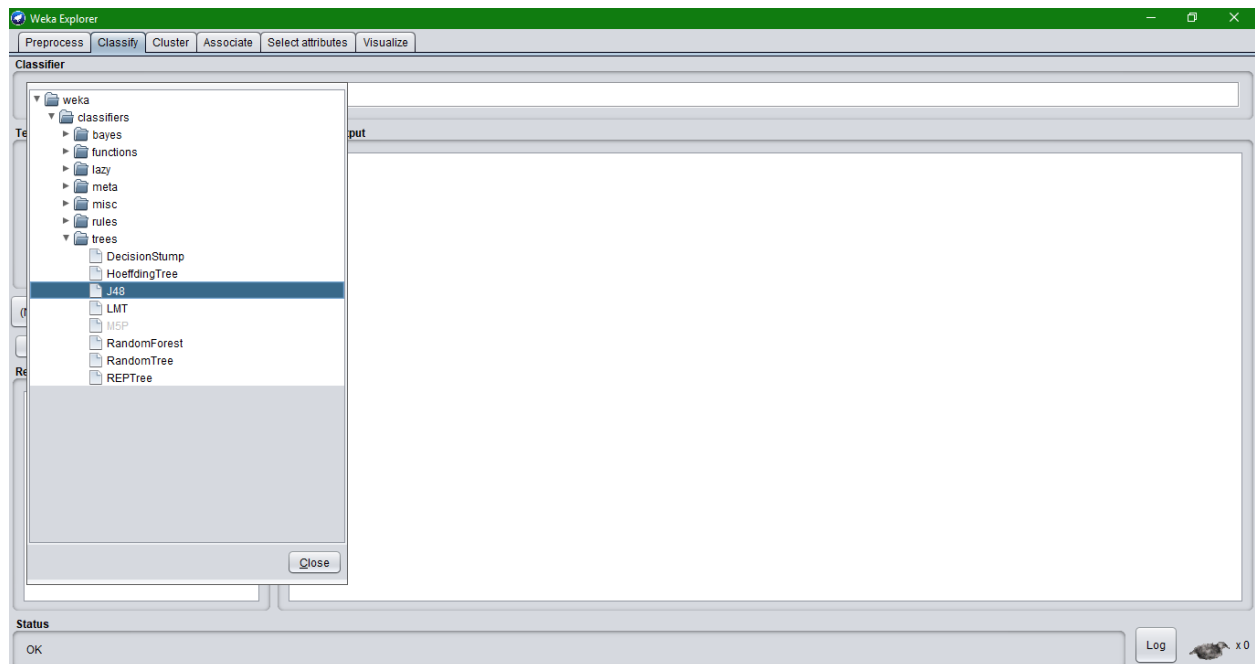


Figure 05: Selecting J48 (ID3) for Decision tree classification

- IV. Select the 'Use training set' for classification our general dataset and make sure to select the 'Result' attribute because we want to predict the Result by using this classification model.

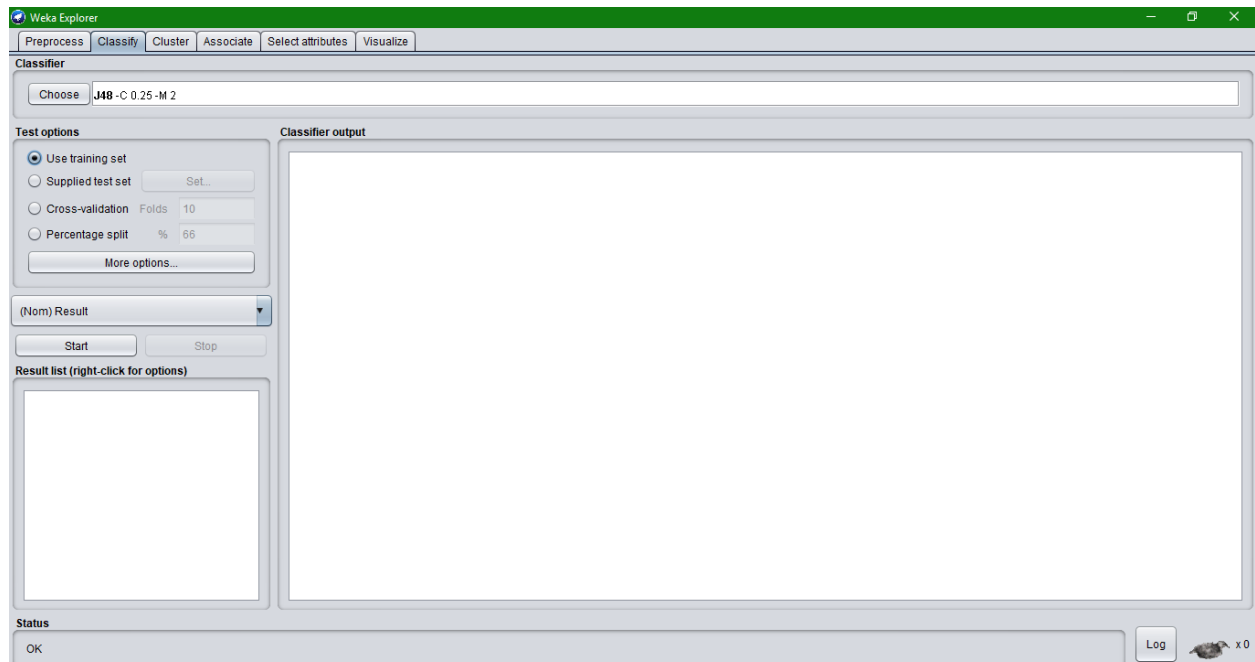


Figure 06: Selecting 'Use training set' and select 'Result' for classification based on the result

V. Now, tap the 'Start' button for classification.

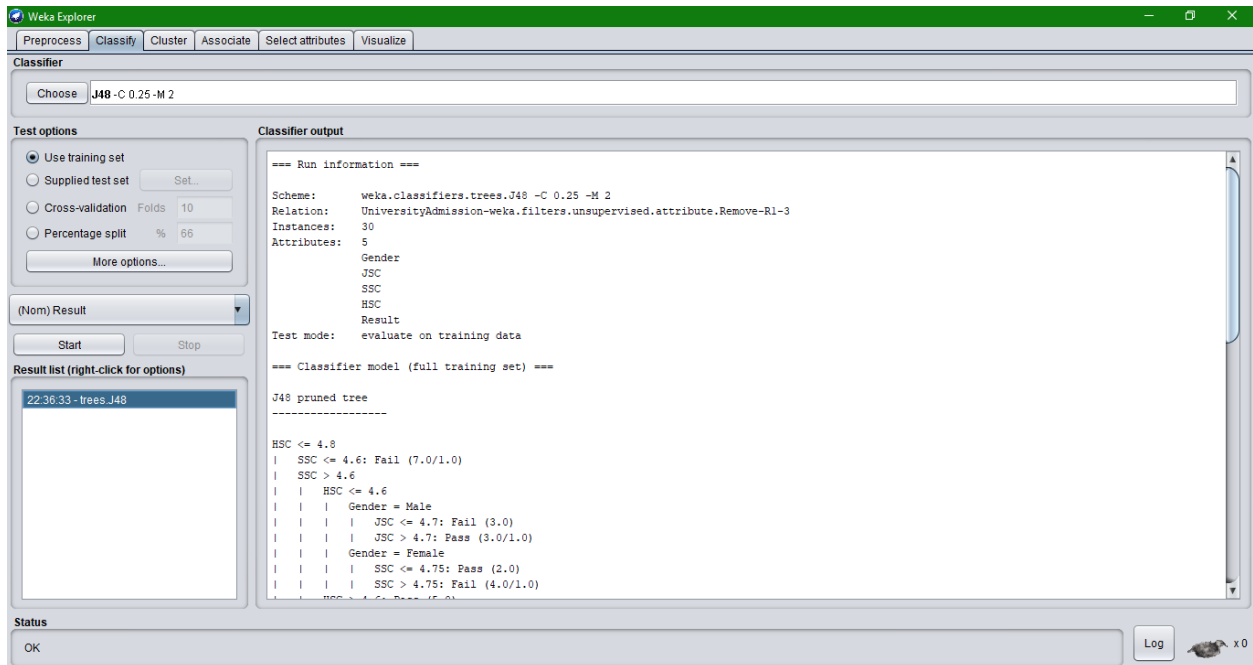


Figure 07: Classification result for Decision Tree of training set

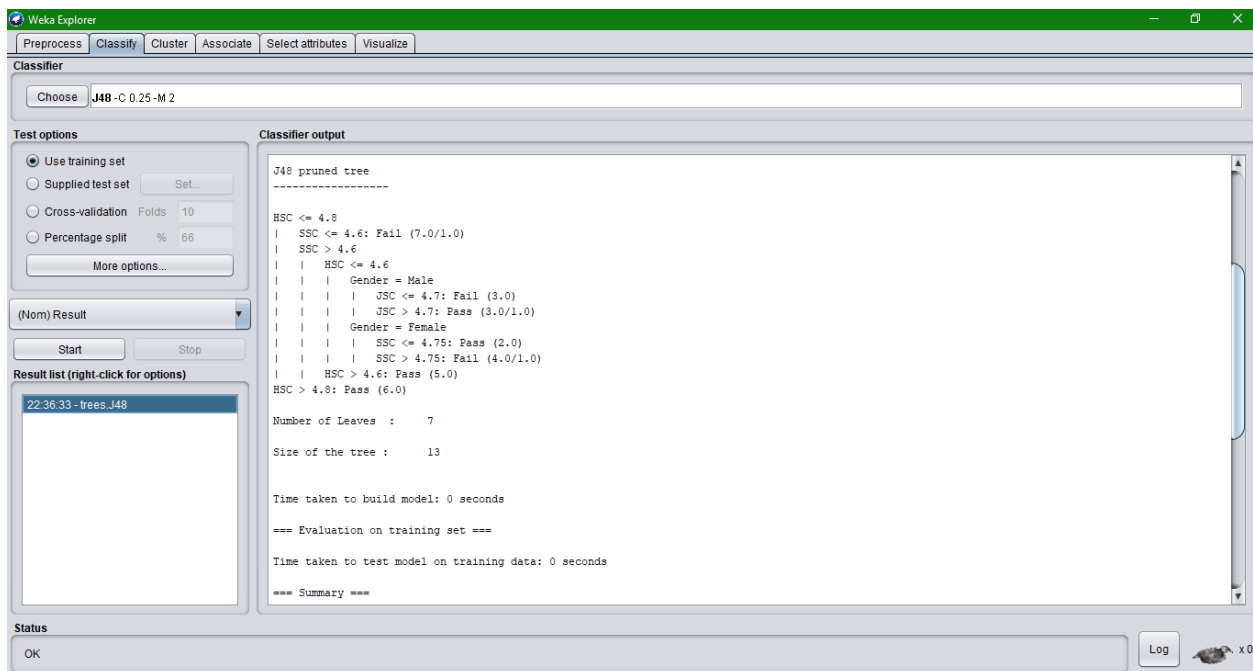


Figure 08: Classification result (tree) for Decision Tree of training set

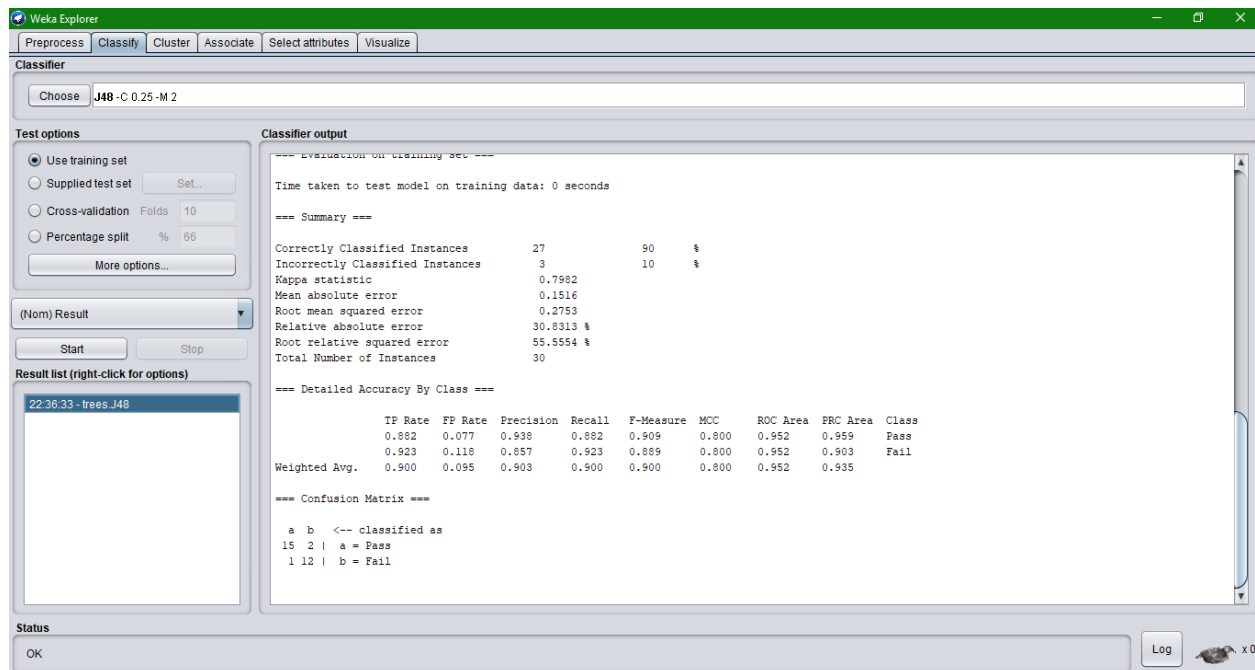


Figure 09: Summary of the classification result for Decision Tree of training set

This is the classification result of our training data-set.

In classification result we see that, there are 5 attributes Gender, JSC, SSC, HSC and Result.

Here,

Total number of instances = 30

Correctly classified instance = 90%

Incorrectly classified instance = 10%

Mean absolute error = 0.1516

Root mean squared error = 0.2753

Relative absolute error = 30.8313%

Root relative squared error = 55.5554%

Confusion Matrix = a b

15 2 | a = Pass

1 12 | b = Fail

At the end, we also see the confusion matrix. That's the classification of Decision Tree.

Let's see the Decision tree for this classification. Right click the 'trees.J48' in Result list and select the visualize tree. The Decision Tree is given below.

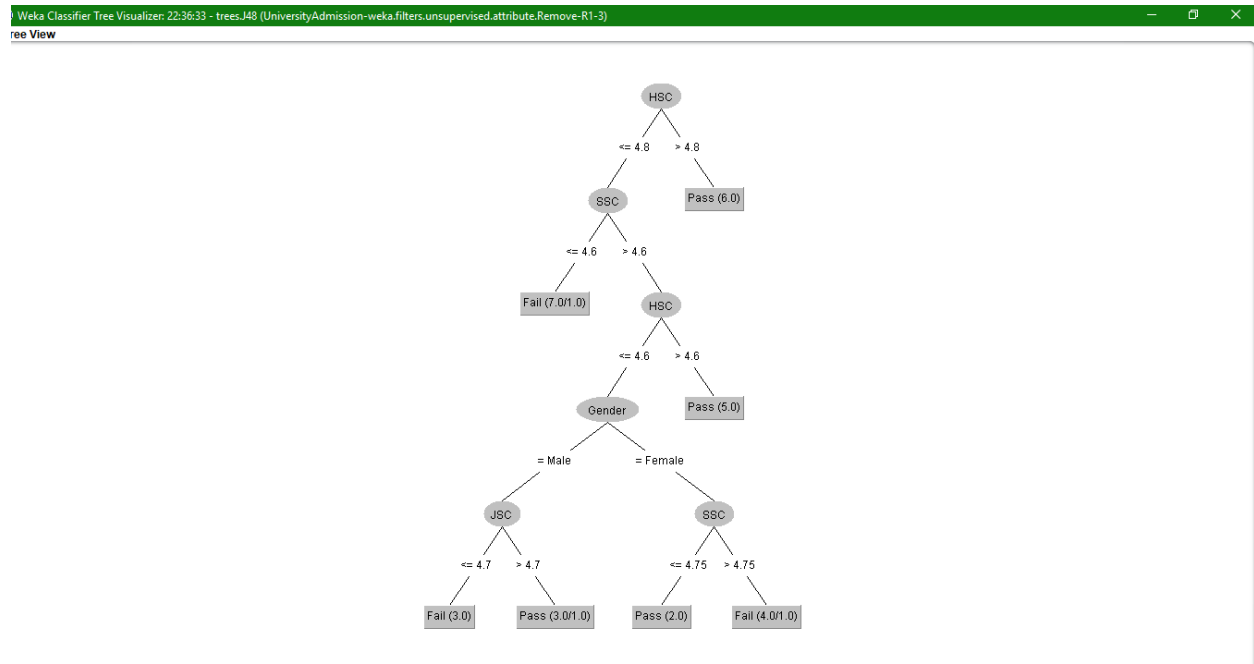


Figure 10: The Decision Tree for the classification of training set

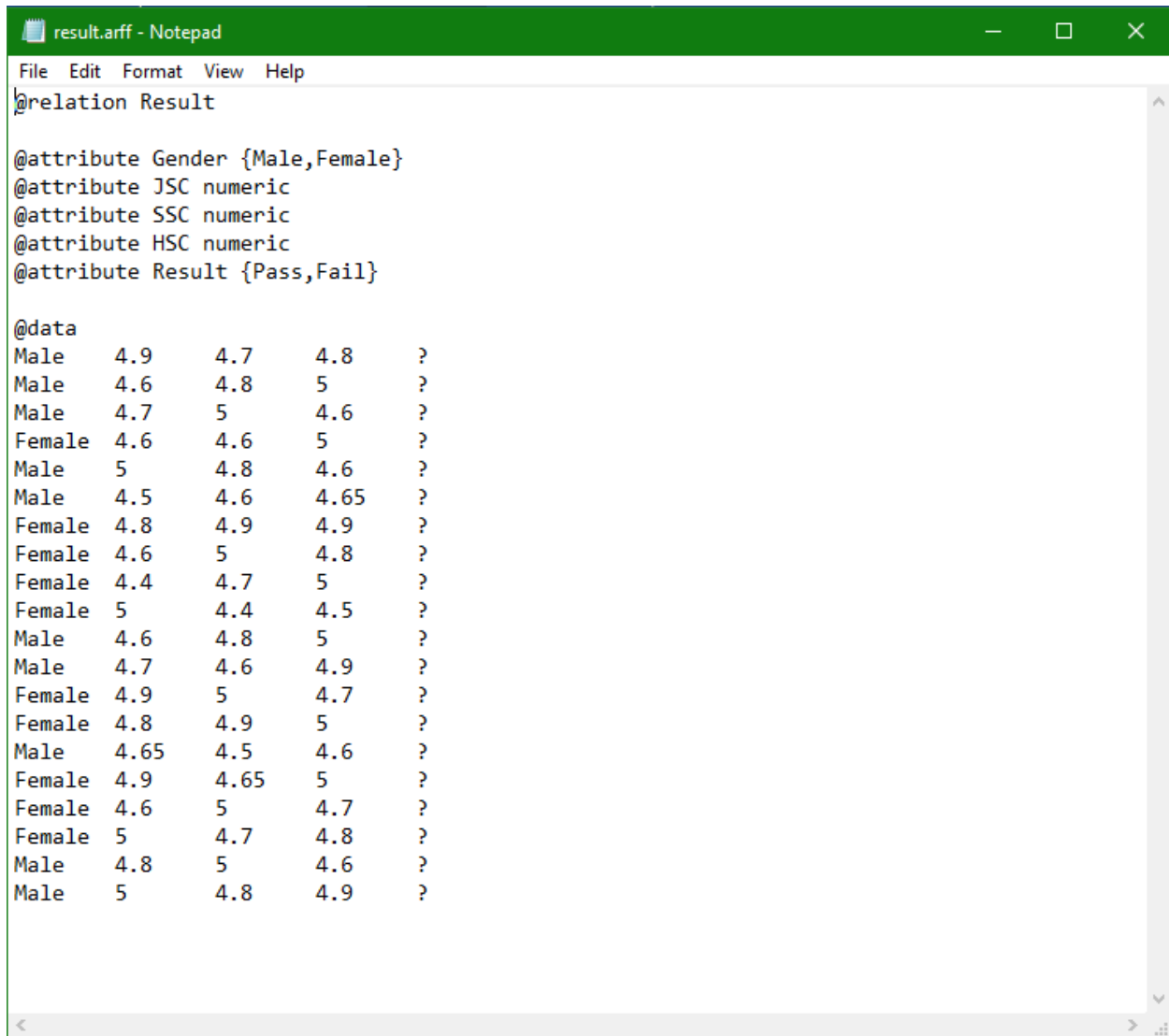
Explaining the Decision Tree:

In the Decision tree, total number of candidates (instance) is 30. If HSC GPA is greater than 4.8, then 6 candidates are pass in the admission test. If HSC GPA is less or equal to 4.8, then you have to check candidates' SSC GPA. If SSC GPA is less or equal to 4.6, then 7 candidates are fail in the admission test. If SSC GPA is greater than 4.6, then you have to check candidates' HSC GPA again. If HSC GPA is greater than 4.6, then 5 candidates are pass in the admission test. If HSC GPA is less or equal to 4.6, then you have to check the gender of the candidates. If the candidates are Male, then you have to check these candidates JSC GPA. If JSC GPA of Male candidates are greater than 4.7, then 3 candidates are pass in the admission test. If JSC GPA of Male candidates are less or equal to 4.7, then 3 candidates are fail in the admission test as well. If the candidates are Female, then you have to check these candidates SSC GPA. If SSC GPA of Female candidates are greater than 4.75, then 4 candidates are fail in the admission test. If SSC GPA of Female candidates are less or equal to 4.75, then 2 candidates are pass in the admission test. That's the well explanation of the above tree.

Now time for predicting the result of second dataset (Figure 02) using this classification model. The procedures are given below.

- i. You have to copy the necessary attributes (Gender, JSC, SSC, HSC) columns and then paste in the any text editor app like Notepad and save the file extension as .arff for making ARFF file. ARFF file was developed by the Machine Learning Project for use with the Weka machine learning software. Without ARFF file we can't test or predict any dataset by using the classification model. An ARFF file contains two sections – header and data. The header describes the attributes types. The data section contains a comma or space separate list of data. As we only paste the data in text editor app, so we have to write the header portion for describe

these attributes. For categorical attributes, we have to mention the categories by using second parenthesis separated by comma. ARFF file containing both sections is given below.



```
@relation Result

@attribute Gender {Male,Female}
@attribute JSC numeric
@attribute SSC numeric
@attribute HSC numeric
@attribute Result {Pass,Fail}

@data
Male 4.9 4.7 4.8 ?
Male 4.6 4.8 5 ?
Male 4.7 5 4.6 ?
Female 4.6 4.6 5 ?
Male 5 4.8 4.6 ?
Male 4.5 4.6 4.65 ?
Female 4.8 4.9 4.9 ?
Female 4.6 5 4.8 ?
Female 4.4 4.7 5 ?
Female 5 4.4 4.5 ?
Male 4.6 4.8 5 ?
Male 4.7 4.6 4.9 ?
Female 4.9 5 4.7 ?
Female 4.8 4.9 5 ?
Male 4.65 4.5 4.6 ?
Female 4.9 4.65 5 ?
Female 4.6 5 4.7 ?
Female 5 4.7 4.8 ?
Male 4.8 5 4.6 ?
Male 5 4.8 4.9 ?
```

Figure 11: Creating .arff file for predicting result

- ii. Select the 'Supplied test set' in Weka then click 'Set' and choose that .arff file for predicting. Then close the file choosing tab and make sure to select the Result attributes because we want to predict that result.

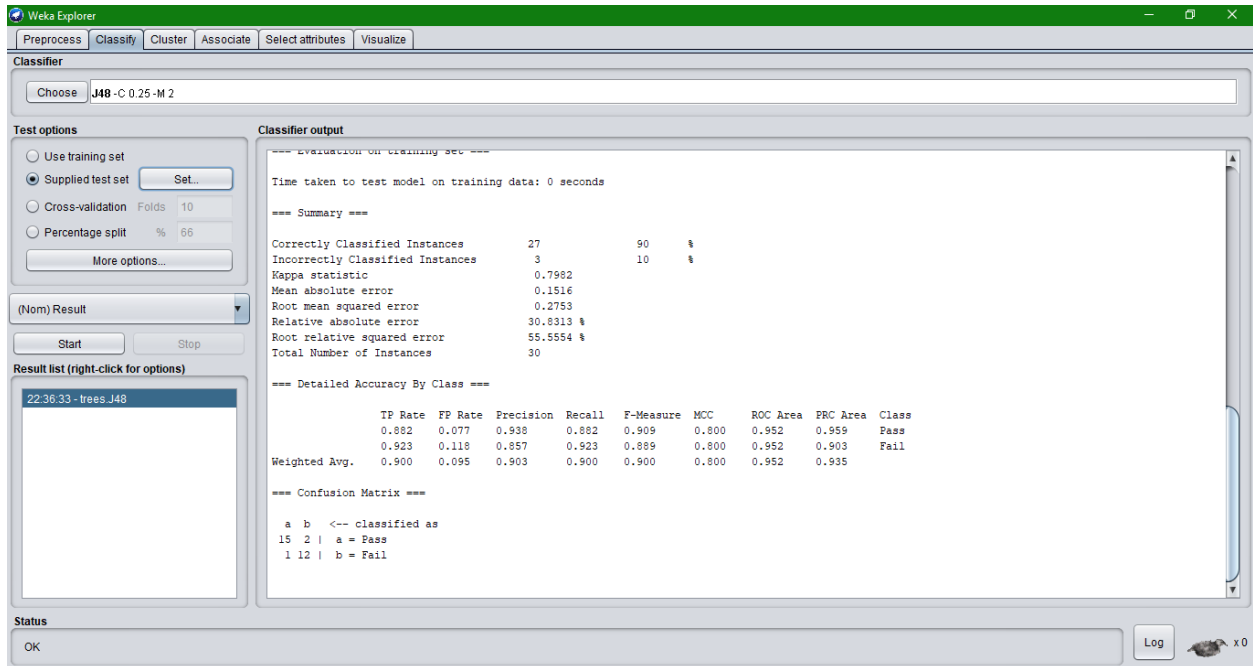


Figure 12: Inserting .arff file for predicting the result

- iii. Click the ‘start’ button for getting the predicting result. You have to ensure that ‘Result’ bar is selected because we want to predict the result. Then right click on the second row of Result list and select ‘Visualize classifiers error’. The procedure for these instruction is given below.

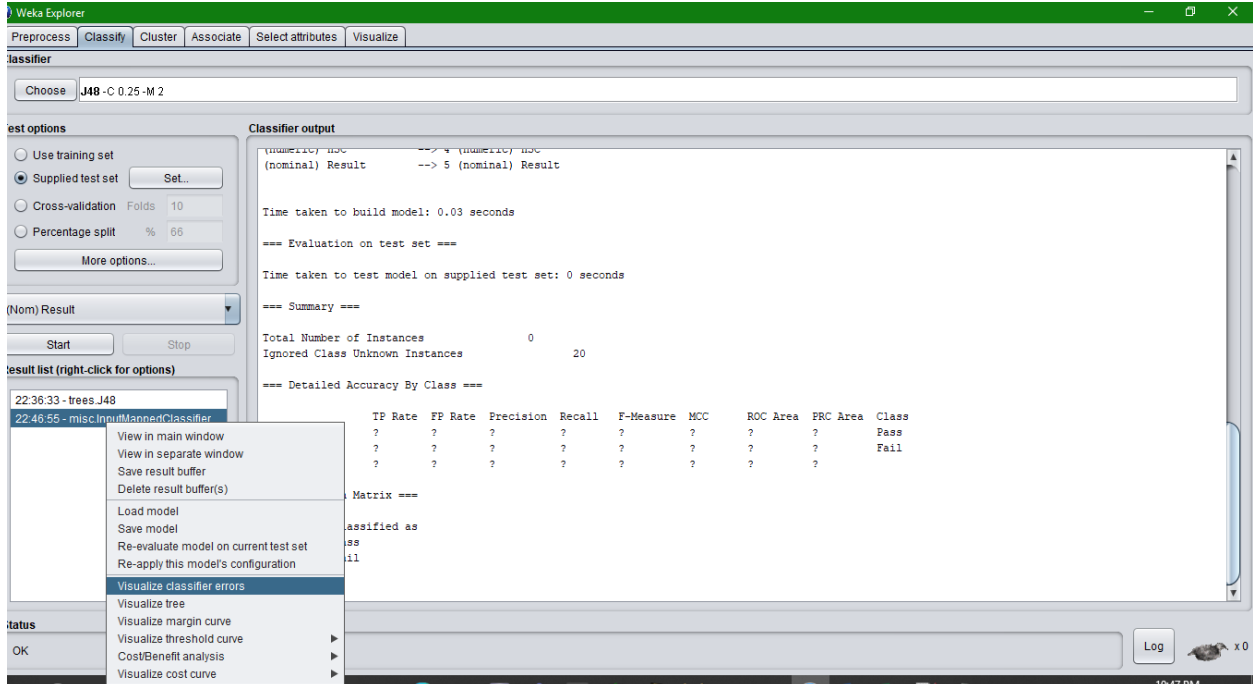


Figure 13: Procedure for predicting result in new dataset

- iv. Click the ‘save’ button for storing the output result in your device.

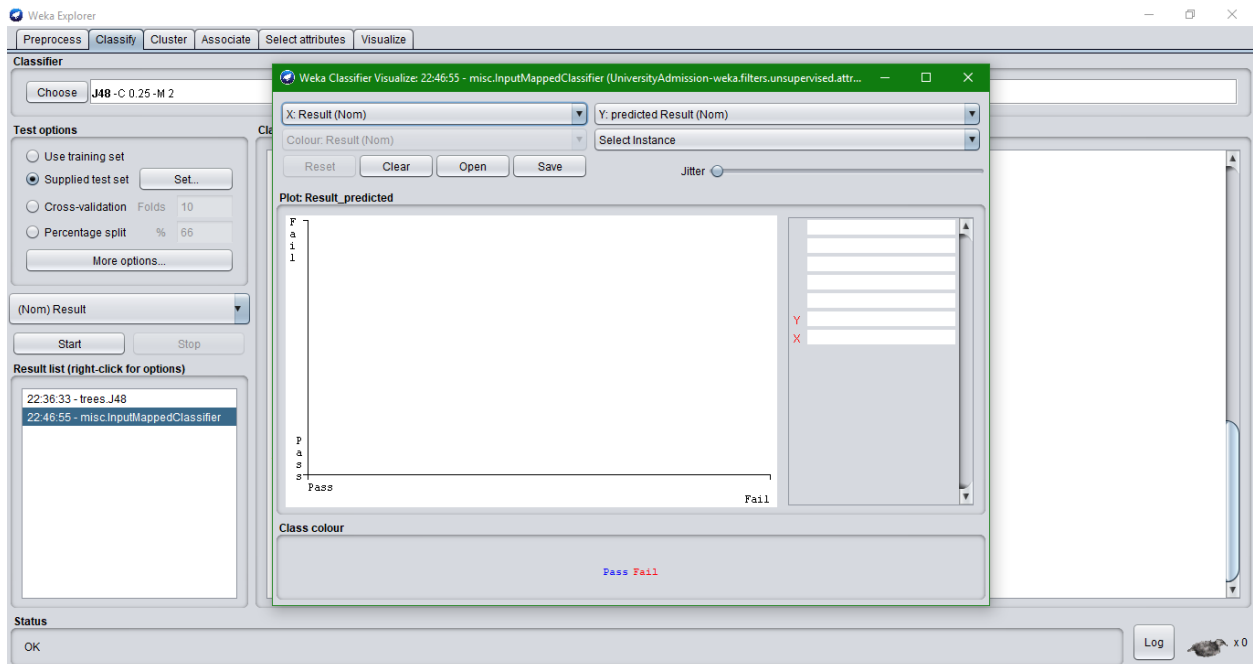


Figure 14: Procedure for storing the result of testing dataset

- v. Open that file in Notepad++ or any other text editor app to see the predicting result.

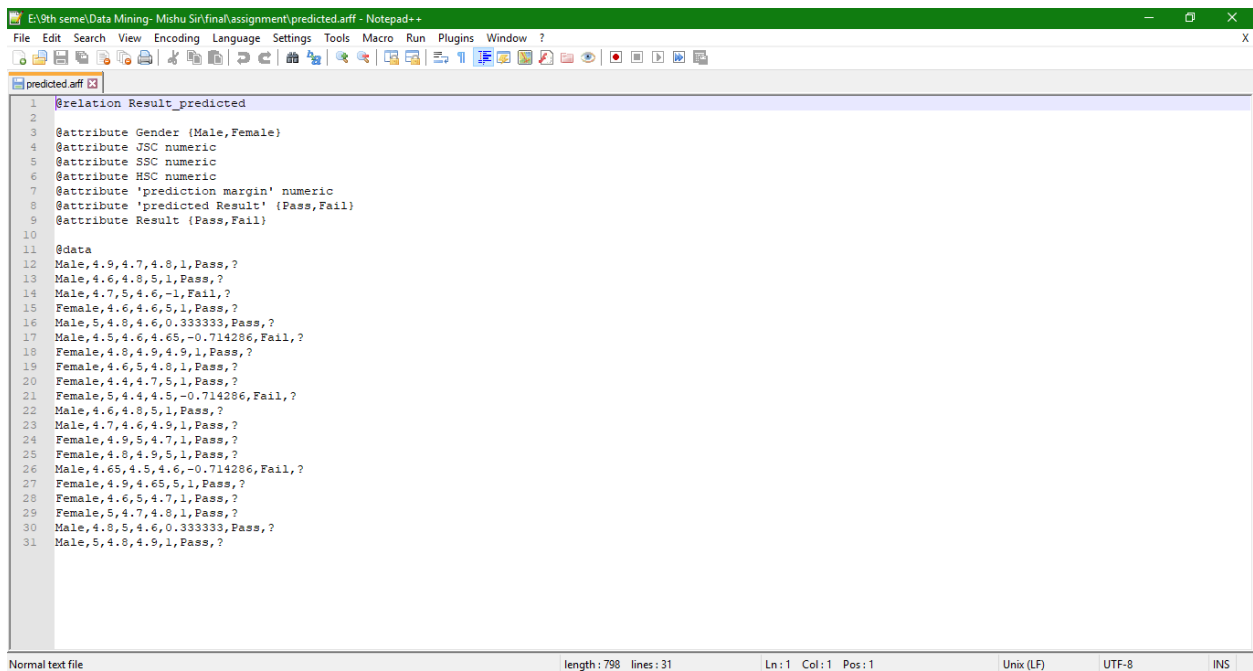


Figure 15: Predicting the Result (Pass or Fail) of new dataset.

In Figure 15, we see that an ARFF file is automatic generated from Weka containing the predicting result. This file contains two section as well, they are header and data. Header has the attributes list of numerical and categorical data. Here we also see an extra attribute in header which is

‘prediction margin’. This attribute is generated automatically while testing this dataset. In data section, every data containing either P (Pass) or F (Fail) at the end. These value of the attribute is predicting from the classification model of the first data set.

That’s how we are testing a new dataset and predicting the value of class attributes based on the classification model which we made earlier.

Comparison between training set and testing set:

Now put the predicting data in the testing set for comparison between training set and testing set.

	A	B	C	D	E	F	G	H
1	Reg no	Name	DoB	Gender	JSC	SSC	HSC	Result
2	526912	Rahim	9.11.1999	Male	4.9	4.7	4.8	Pass
3	843645	Karim	8.11.1998	Male	4.6	4.8	5	Pass
4	512648	Shimul	30.11.1998	Male	4.7	5	4.6	Fail
5	841266	Faria	20.11.1999	Female	4.6	4.6	5	Pass
6	542582	Sayem	1.12.1998	Male	5	4.8	4.6	Pass
7	845631	Rafat	8.12.1999	Male	4.5	4.6	4.65	Fail
8	452175	Maria	20.5.1998	Female	4.8	4.9	4.9	Pass
9	478963	Munni	22.5.1998	Female	4.6	5	4.8	Pass
10	154782	Nusrat	5.3.1999	Female	4.4	4.7	5	Pass
11	569210	Lamya	3.11.1999	Female	5	4.4	4.5	Fail
12	412560	Imran	15.7.1999	Male	4.6	4.8	5	Pass
13	266940	Afridi	7.3.1998	Male	4.7	4.6	4.9	Pass
14	126436	Borsha	15.9.1999	Female	4.9	5	4.7	Pass
15	158340	Nourin	25.10.1999	Female	4.8	4.9	5	Pass
16	230548	Rubel	26.7.1998	Male	4.65	4.5	4.6	Fail
17	356046	Sana	6.9.1999	Female	4.9	4.65	5	Pass
18	520047	Fariya	28.11.1999	Female	4.6	5	4.7	Pass
19	953485	Tussy	25.12.1998	Female	5	4.7	4.8	Pass
20	588445	Rajon	17.5.1998	Male	4.8	5	4.6	Pass
21	269425	Syed	20.4.1999	Male	5	4.8	4.9	Pass

Figure 16: Putting the predicting result in the tasting data set

Now, follow the same procedure for classification this testing set which is given earlier. The classification result of the testing set is given below.

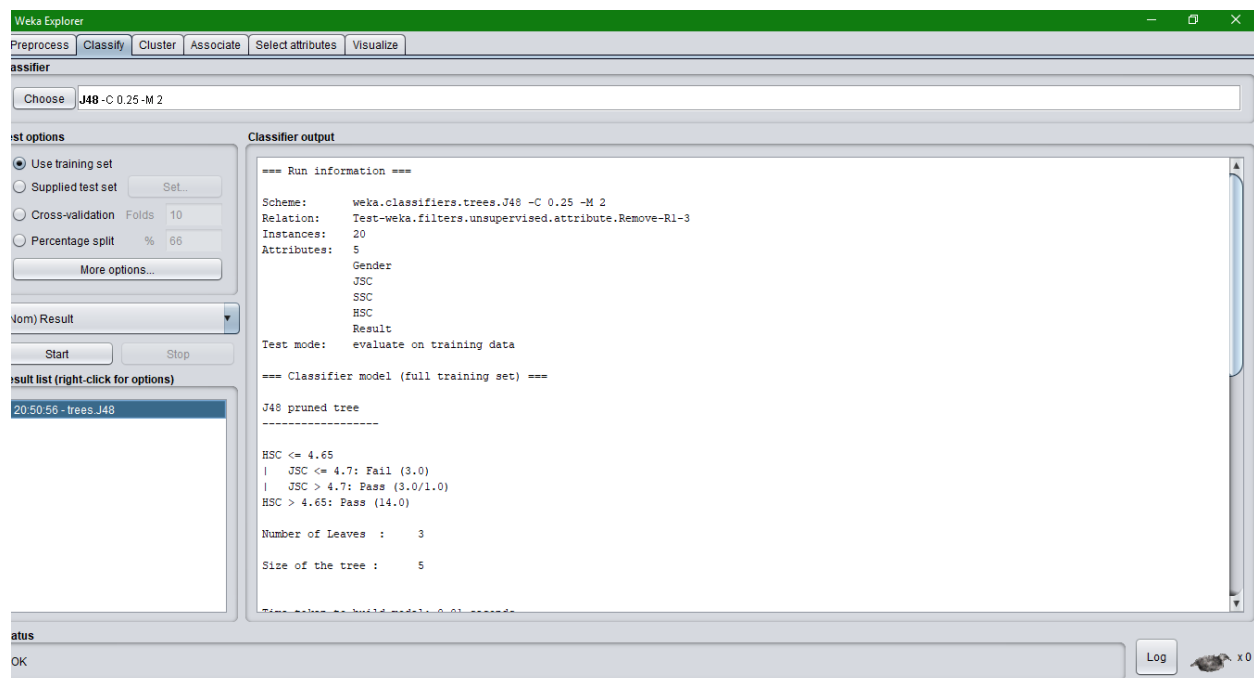


Figure 17: Classification result for Decision Tree of testing set

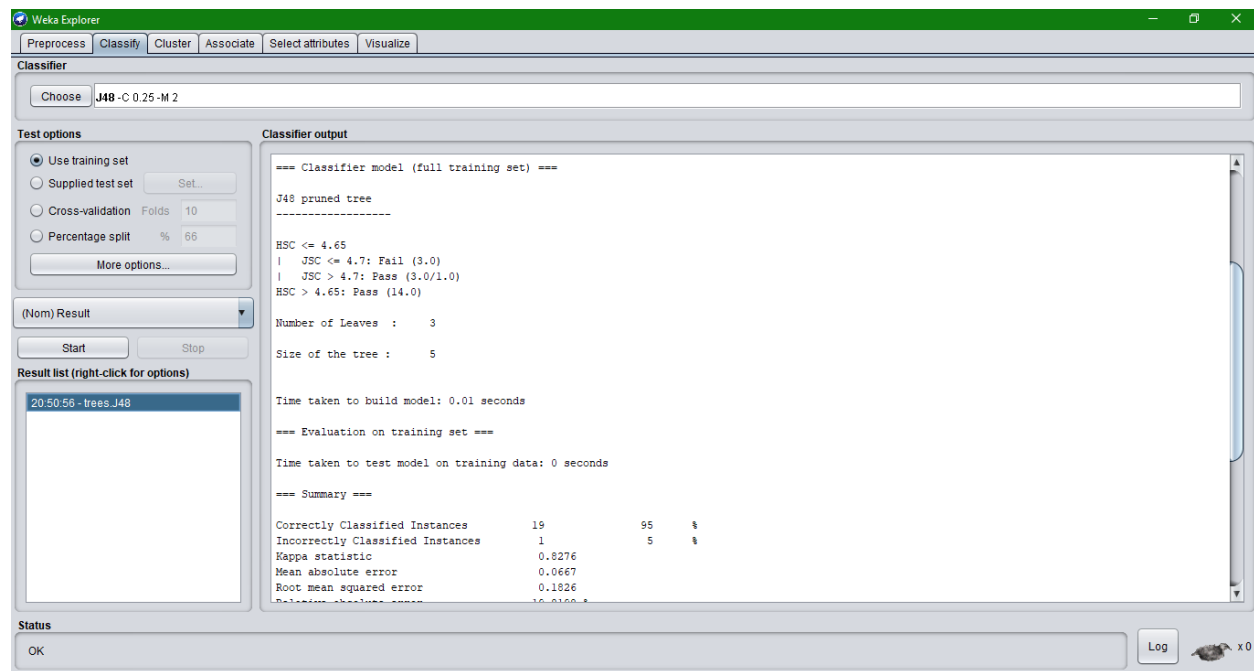


Figure 18: Classification result (tree) for Decision Tree of testing set

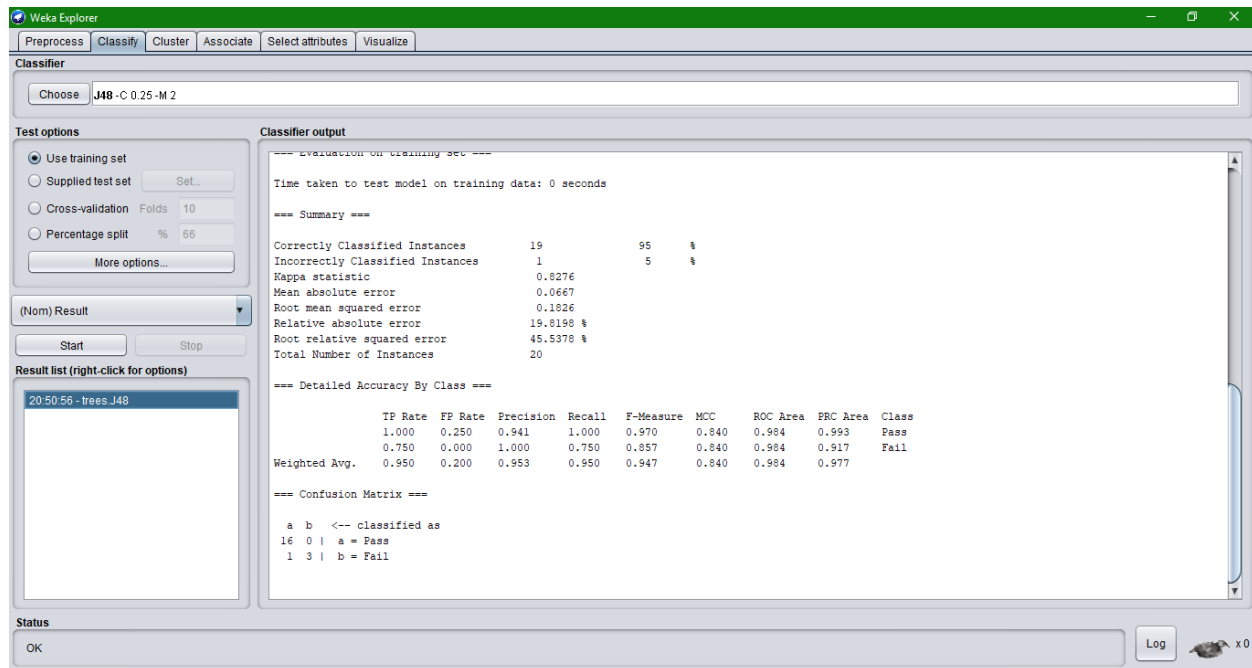


Figure 19: Summary of the classification result for Decision Tree of testing set

This is the classification result of our testing data-set.

In classification result we see that, there are 5 attributes Gender, JSC, SSC, HSC and Result as well.

Here,

Total number of instances = 20

Correctly classified instance = 95%

Incorrectly classified instance = 5%

Mean absolute error = 0.0667

Root mean squared error = 0.1826

Relative absolute error = 19.8198%

Root relative squared error = 45.5378%

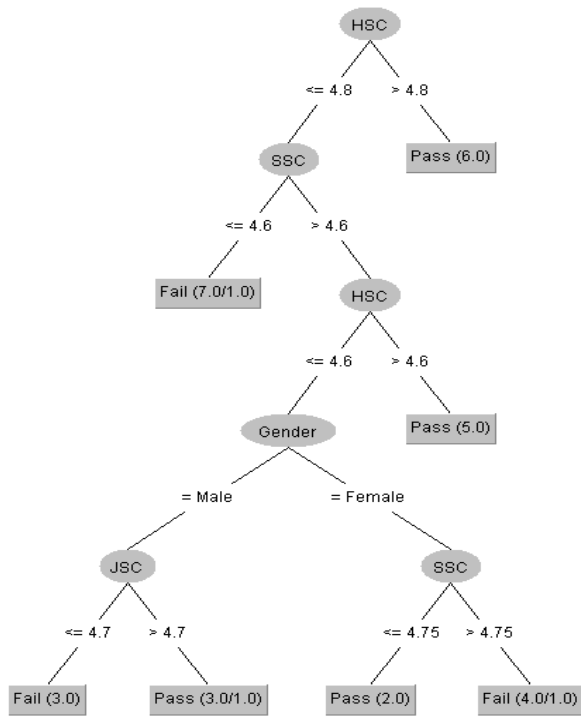
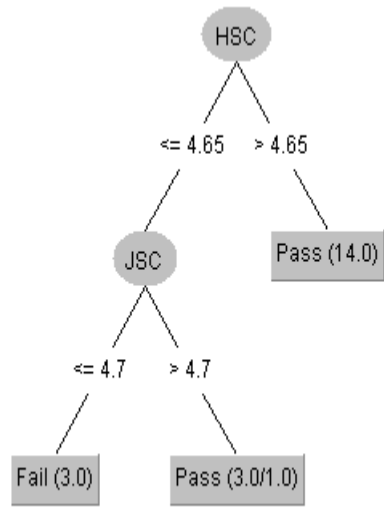
Confusion Matrix = a b

16 0 | a = Pass

1 3 | b = Fail

Comparison table based on the different criteria is given below.

Table 01: Comparison between training set and testing set

	Training data set	Testing data set
Total number of instances	30	20
Correctly classified instance	90%	95%
Incorrectly classified instance	10%	5%
Mean absolute error	0.1516	0.0667
Root mean squared error	0.2753	0.1826
Relative absolute error	30.8313%	19.8198%
Root relative squared error	55.5554%	45.5378%
Tree	 <pre> graph TD HSC1((HSC)) -- "<= 4.8" --> SSC1((SSC)) HSC1 -- "> 4.8" --> Pass60[Pass (6.0)] SSC1 -- "<= 4.6" --> Fail70[Fail (7.0/1.0)] SSC1 -- "> 4.6" --> HSC2((HSC)) HSC2 -- "<= 4.6" --> Gender((Gender)) HSC2 -- "> 4.6" --> Pass50[Pass (5.0)] Gender -- "= Male" --> JSC1((JSC)) Gender -- "= Female" --> SSC2((SSC)) JSC1 -- "<= 4.7" --> Fail30[Fail (3.0)] JSC1 -- "> 4.7" --> Pass30[Pass (3.0/1.0)] SSC2 -- "<= 4.75" --> Pass20[Pass (2.0)] SSC2 -- "> 4.75" --> Fail40[Fail (4.0/1.0)] </pre>	 <pre> graph TD HSC3((HSC)) -- "<= 4.65" --> JSC3((JSC)) HSC3 -- "> 4.65" --> Pass140[Pass (14.0)] JSC3 -- "<= 4.7" --> Fail300[Fail (3.0)] JSC3 -- "> 4.7" --> Pass300[Pass (3.0/1.0)] </pre>

In the comparison table, we see that the correctly classified instance is increasing in the testing data set which is 95%. Also the mean absolute error, root mean squared error, relative absolute error, root related squared error are decreasing in the testing data set. So, we can easily tell that our classification model of Decision Tree (training data set) is suitable for any kinds of testing data set. Now, we can test any other dataset by using our classification model because this classification model helps to increase the accuracy of other testing data set.

Conclusion:

As one of the most important and supervised algorithms, Decision Tree plays a vital role in decision analysis in real life. As a predictive model, it is used in many areas for its split approach which helps in identifying solutions based on different conditions by either classification or regression method.

Reference(s):

- [1] [Decision Tree \(saedsayad.com\)](https://saedsayad.com/decision-tree/)
- [2] [Decision Tree Classification. A Decision Tree is a simple... | by Afroz Chakure | The Startup | Medium](#)
- [3] [Classification Algorithms - Decision Tree - Tutorialspoint](#)
- [4] [Machine Learning Decision Tree Classification Algorithm - Javatpoint](#)
- [5] [Weka Decision Tree | Build Decision Tree Using Weka \(analyticsvidhya.com\)](#)
- [6] [Classification via Decision Trees in WEKA \(depaul.edu\)](#)
- [7] [Classification via Decision Trees in WEKA \(depaul.edu\)](#)
- [8] [Prediction of Student Results #Data Mining - YouTube](#)