PROJECT REPORT

Course Name : Data Mining and Machine Learning

Course code : CSE-322

Submitted To :

Dewan Mamun Raza

Department of CSE

Submitted By :

MD Tasluf Morshed	191-15-12089
MD Assadujjaman Tilok	191-15-12594
Md. Riazul Islam Gisun	191-15-12772
Towhidul Islam Shyon	191-15-12728

Project Title:

SRTOCK PREDICTION FRAMEWORK USING CLASTERING ASGORITHM

Project Objective:

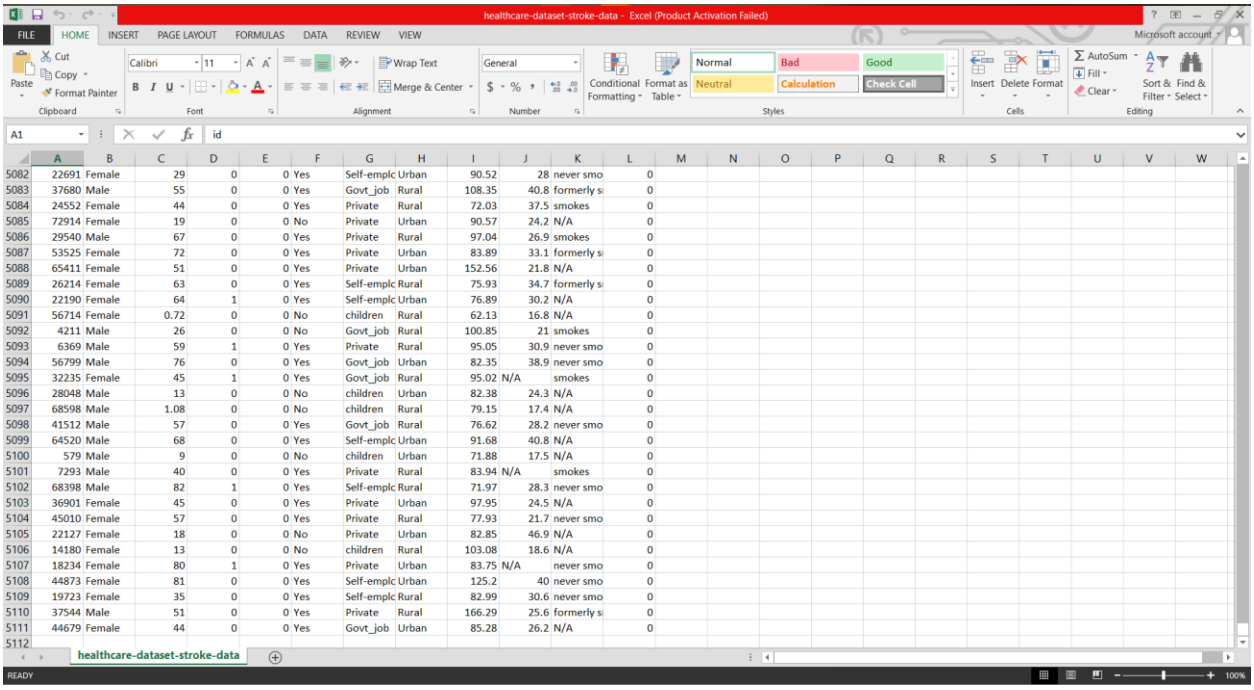
Stroke is considered to be a burning issue not only in South Asia but also all over the world. Every year, around 16 million people are affected for the first time by stroke, and 5.7 million of them die. Due to population aging, the number of individuals affected by a stroke is increasing over the years. It is generating a lot of controversies when the boom began from 1.1 to 13.7 million between 2000 to 2016. In this classification, I want to predict if someone has any risk of being affected by stroke tests by using Classification methods. I want to use Random Forest, Decision tree, and SVM algorithm and compare them to find which algorithm will be best for this. By using this classifier I think I may predict 3 to 12 months earlier if someone has any possibility for stroke.

Expected Outcomes:

Here buy designing this framework, we want to predict the class attribute, which can predict if someone has any risk of stroke. Therefore, people can take necessary steps as early as possible and prevent there life risk.

Procedure:

<u>Step 01:</u> we've to take to identify and select a dataset based on our work.



	A	B	C	D	E	F	G	H	I	J	K	L
5082	22691	Female	29	0	0	Yes	Self-emplc	Urban	90.52	28	never smo	0
5083	37680	Male	55	0	0	Yes	Govt_job	Rural	108.35	40.8	formerly si	0
5084	24552	Female	44	0	0	Yes	Private	Rural	72.03	37.5	smokes	0
5085	72914	Female	19	0	0	No	Private	Urban	90.57	24.2	N/A	0
5086	29540	Male	67	0	0	Yes	Private	Rural	97.04	26.9	smokes	0
5087	53525	Female	72	0	0	Yes	Private	Urban	83.89	33.1	formerly si	0
5088	65411	Female	51	0	0	Yes	Private	Urban	152.56	21.8	N/A	0
5089	26214	Female	63	0	0	Yes	Self-emplc	Rural	75.93	34.7	formerly si	0
5090	22190	Female	64	1	0	Yes	Self-emplc	Urban	76.89	30.2	N/A	0
5091	56714	Female	0.72	0	0	No	children	Rural	62.13	16.8	N/A	0
5092	4211	Male	26	0	0	No	Govt_job	Rural	100.85	21	smokes	0
5093	6369	Male	59	1	0	Yes	Private	Rural	95.05	30.9	never smo	0
5094	56799	Male	76	0	0	Yes	Govt_job	Urban	82.35	38.9	never smo	0
5095	32235	Female	45	1	0	Yes	Govt_job	Rural	95.02	N/A	smokes	0
5096	28048	Male	13	0	0	No	children	Urban	82.38	24.3	N/A	0
5097	68598	Male	1.08	0	0	No	children	Rural	79.15	17.4	N/A	0
5098	41512	Male	57	0	0	Yes	Govt_job	Rural	76.62	28.2	never smo	0
5099	64520	Male	68	0	0	Yes	Self-emplc	Urban	91.68	40.8	N/A	0
5100	579	Male	9	0	0	No	children	Urban	71.88	17.5	N/A	0
5101	7293	Male	40	0	0	Yes	Private	Rural	83.94	N/A	smokes	0
5102	68398	Male	82	1	0	Yes	Self-emplc	Rural	71.97	28.3	never smo	0
5103	36901	Female	45	0	0	Yes	Private	Urban	97.95	24.5	N/A	0
5104	45010	Female	57	0	0	Yes	Private	Rural	77.93	21.7	never smo	0
5105	22127	Female	18	0	0	No	Private	Urban	82.85	46.9	N/A	0
5106	14180	Female	13	0	0	No	children	Rural	103.08	18.6	N/A	0
5107	18234	Female	80	1	0	Yes	Private	Urban	83.75	N/A	never smo	0
5108	44873	Female	81	0	0	Yes	Self-emplc	Urban	125.2	40	never smo	0
5109	19723	Female	35	0	0	Yes	Self-emplc	Rural	82.99	30.6	never smo	0
5110	37544	Male	51	0	0	Yes	Private	Rural	166.29	25.6	formerly si	0
5111	44679	Female	44	0	0	Yes	Govt_job	Urban	85.28	26.2	N/A	0

<u>Step 02:</u>

We've to change ta formet of dataset from .csv to .arff

for that:

@relation stroke-data

@attribute id numeric

@attribute gender {Male,Female}

@attribute age numeric

@attribute hypertension numeric

@attribute heart_disease numeric

@attribute ever_married {Yes,No}

@attribute work_type {Private,Self-employed,Govt_job,children}

@attribute Residence_type {Urban,Rural)

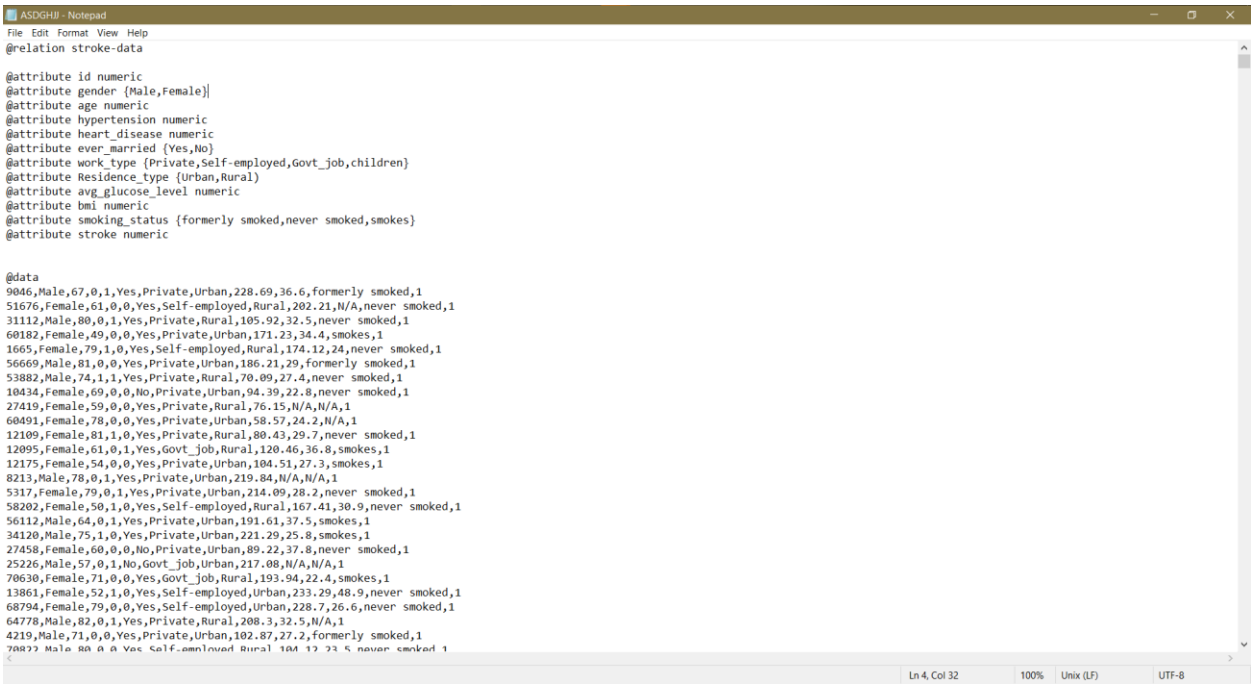
@attribute avg_glucose_level numeric

@attribute bmi numeric

@attribute smoking_status {formerly smoked,never smoked,smokes}

@attribute stroke numeric

@data

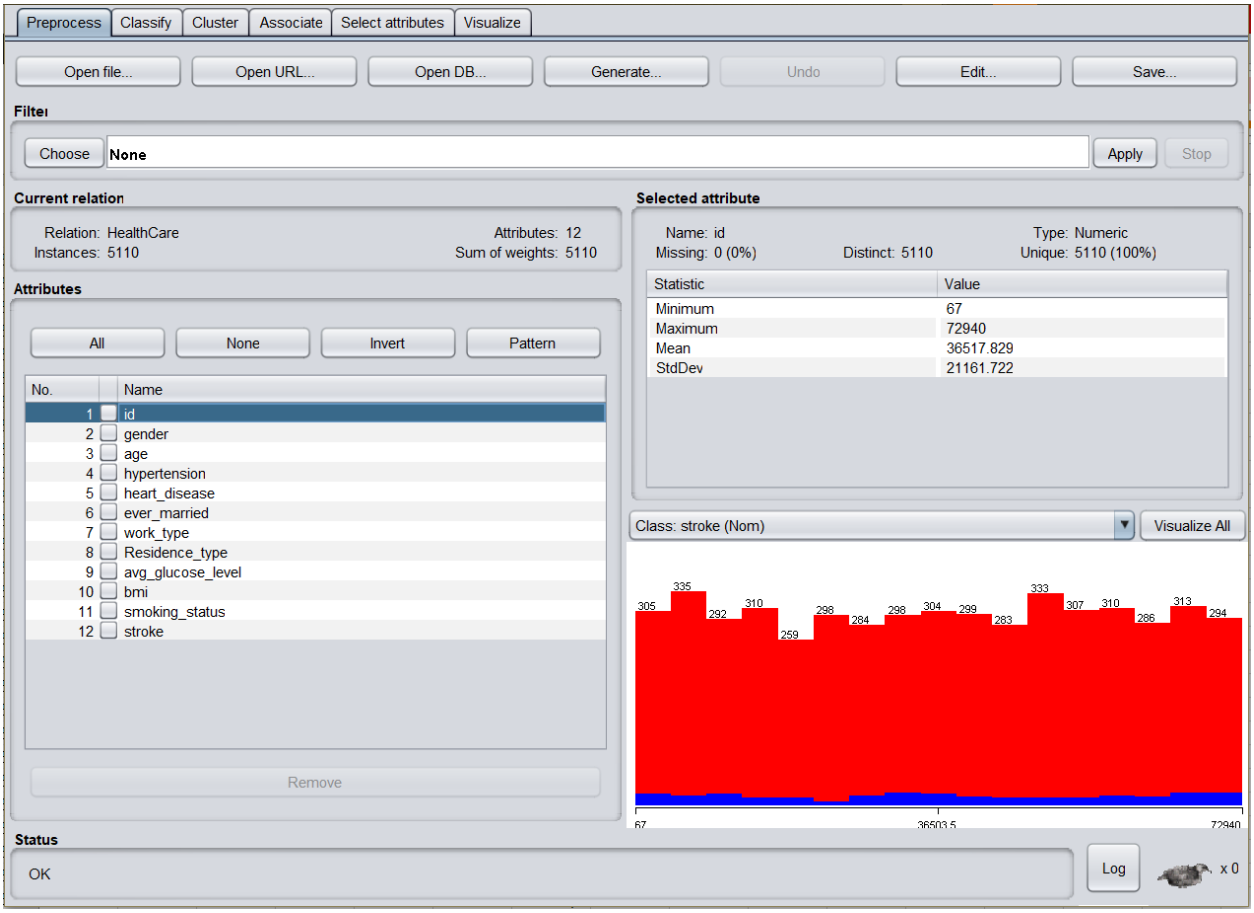


```
ASDGHJJ - Notepad
File  Edit  Format  View  Help
@relation stroke-data

@attribute id numeric
@attribute gender {Male,Female}
@attribute age numeric
@attribute hypertension numeric
@attribute heart_disease numeric
@attribute ever_married {Yes,No}
@attribute work_type {Private,Self-employed,Govt_job,children}
@attribute Residence_type {Urban,Rural)
@attribute avg_glucose_level numeric
@attribute bmi numeric
@attribute smoking_status {formerly smoked,never smoked,smokes}
@attribute stroke numeric


@data
9046,Male,67,0,1,Yes,Private,Urban,228.69,36.6,formerly smoked,1
51676,Female,61,0,0,Yes,Self-employed,Rural,202.21,N/A,never smoked,1
31112,Male,80,0,1,Yes,Private,Rural,105.92,32.5,never smoked,1
60182,Female,49,0,0,Yes,Private,Urban,171.23,34.4,smokes,1
1665,Female,79,1,0,Yes,Self-employed,Rural,174.12,24,never smoked,1
56669,Male,81,0,0,Yes,Private,Urban,186.21,29,formerly smoked,1
53882,Male,74,1,1,Yes,Private,Rural,70.09,27.4,never smoked,1
10434,Female,69,0,0,No,Private,Urban,94.39,22.8,never smoked,1
27419,Female,59,0,0,Yes,Private,Rural,76.15,N/A,N/A,1
60491,Female,78,0,0,Yes,Private,Urban,58.57,24.2,N/A,1
12109,Female,81,1,0,Yes,Private,Rural,80.43,29.7,never smoked,1
12095,Female,61,0,1,Yes,Govt_job,Rural,120.46,36.8,smokes,1
12175,Female,54,0,0,Yes,Private,Urban,104.51,27.3,smokes,1
8213,Male,78,0,1,Yes,Private,Urban,219.84,N/A,N/A,1
5317,Female,79,0,1,Yes,Private,Urban,214.09,28.2,never smoked,1
58202,Female,50,1,0,Yes,Self-employed,Rural,167.41,30.9,never smoked,1
56112,Male,64,0,1,Yes,Private,Urban,191.61,37.5,smokes,1
34120,Male,75,1,0,Yes,Private,Urban,221.29,25.8,smokes,1
27458,Female,60,0,0,No,Private,Urban,89.22,37.8,never smoked,1
25226,Male,57,0,1,No,Govt_job,Urban,217.08,N/A,N/A,1
70630,Female,71,0,0,Yes,Govt_job,Rural,193.94,22.4,smokes,1
13861,Female,52,1,0,Yes,Self-employed,Urban,233.29,48.9,never smoked,1
68794,Female,79,0,0,Yes,Self-employed,Urban,228.7,26.6,never smoked,1
64778,Male,82,0,1,Yes,Private,Rural,208.3,32.5,N/A,1
4219,Male,71,0,0,Yes,Private,Urban,102.87,27.2,formerly smoked,1
70832,Male,80,0,0,Yes,Self-employed,Rural,104.12,23.5,never smoked,1
Ln 4, Col 32   100%   Unix (LF)   UTF-8
```

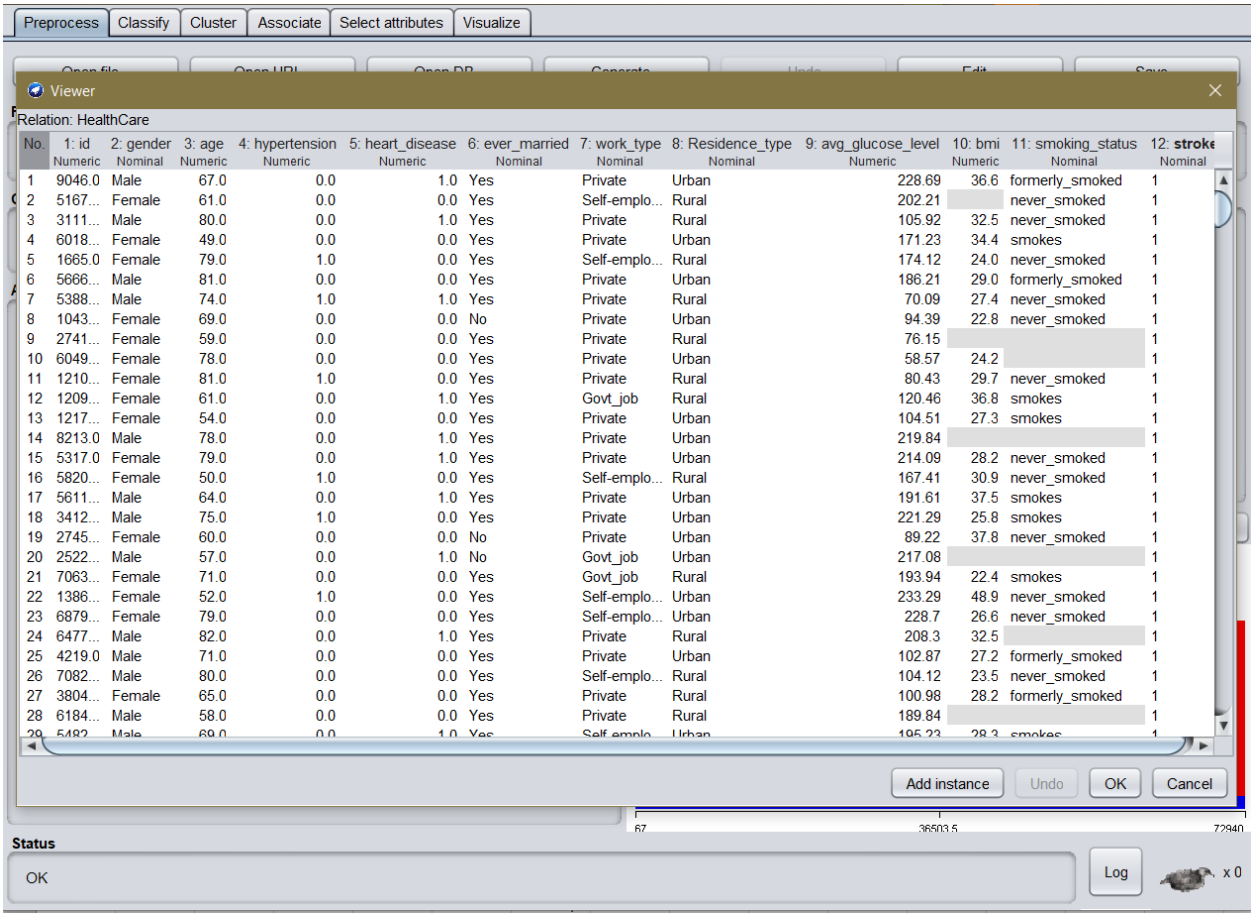
Go to save as then gave a file name (strock-data) then save the file as .arff format

Step 03:

We've to insert our project into Weka. Here the first task bas shoes our attribute names. The top right corner shows type and details of the data set. And bottom one gives the graphical interface of the attributes.



Our data chart is loaded where some values are missing and we can access it by pressing edit button.

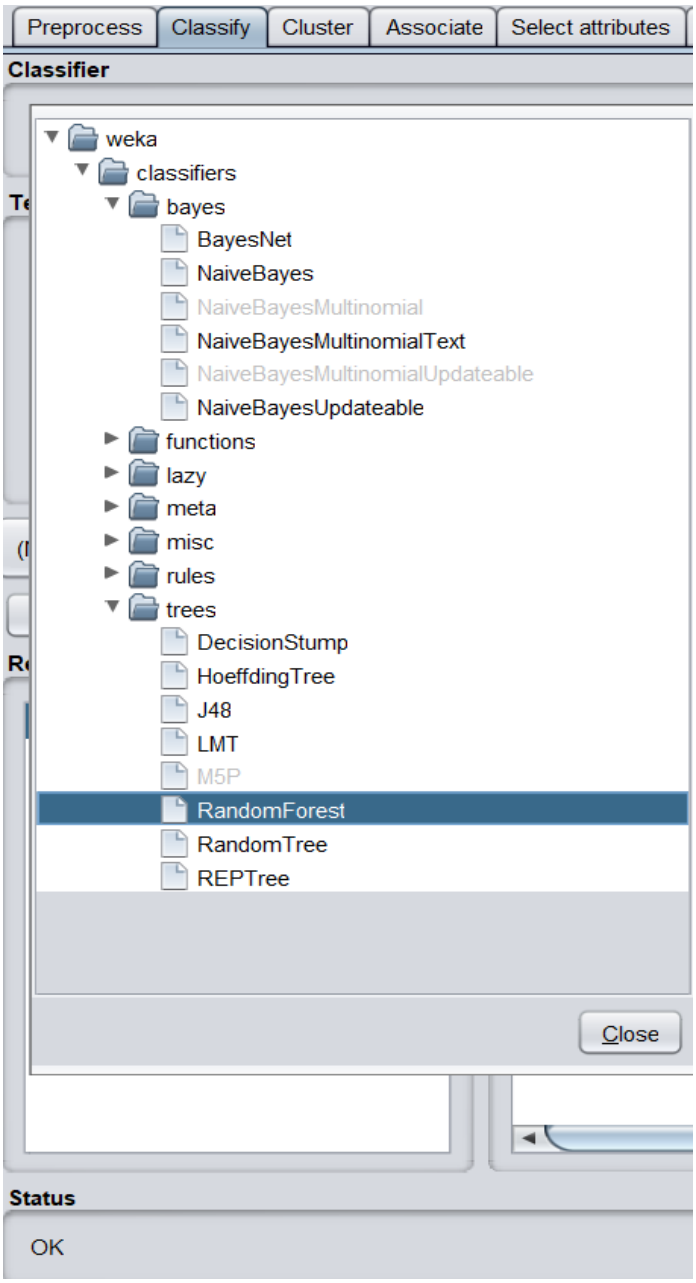


No.	1: id	2: gender	3: age	4: hypertension	5: heart_disease	6: ever_married	7: work_type	8: Residence_type	9: avg_glucose_level	10: bmi	11: smoking_status	12: stroke
	Numeric	Nominal	Numeric	Numeric	Numeric	Nominal	Nominal	Nominal	Numeric	Numeric	Nominal	Nominal
1	9046.0	Male	67.0	0.0	1.0	Yes	Private	Urban	228.69	36.6	formerly_smoked	1
2	5167...	Female	61.0	0.0	0.0	Yes	Self-emplo...	Rural	202.21		never_smoked	1
3	3111...	Male	80.0	0.0	1.0	Yes	Private	Rural	105.92	32.5	never_smoked	1
4	6018...	Female	49.0	0.0	0.0	Yes	Private	Urban	171.23	34.4	smokes	1
5	1665.0	Female	79.0	1.0	0.0	Yes	Self-emplo...	Rural	174.12	24.0	never_smoked	1
6	5666...	Male	81.0	0.0	0.0	Yes	Private	Urban	186.21	29.0	formerly_smoked	1
7	5388...	Male	74.0	1.0	1.0	Yes	Private	Rural	70.09	27.4	never_smoked	1
8	1043...	Female	69.0	0.0	0.0	No	Private	Urban	94.39	22.8	never_smoked	1
9	2741...	Female	59.0	0.0	0.0	Yes	Private	Rural	76.15			1
10	6049...	Female	78.0	0.0	0.0	Yes	Private	Urban	58.57	24.2		1
11	1210...	Female	81.0	1.0	0.0	Yes	Private	Rural	80.43	29.7	never_smoked	1
12	1209...	Female	61.0	0.0	1.0	Yes	Govt_job	Rural	120.46	36.8	smokes	1
13	1217...	Female	54.0	0.0	0.0	Yes	Private	Urban	104.51	27.3	smokes	1
14	8213.0	Male	78.0	0.0	1.0	Yes	Private	Urban	219.84			1
15	5317.0	Female	79.0	0.0	1.0	Yes	Private	Urban	214.09	28.2	never_smoked	1
16	5820...	Female	50.0	1.0	0.0	Yes	Self-emplo...	Rural	167.41	30.9	never_smoked	1
17	5611...	Male	64.0	0.0	1.0	Yes	Private	Urban	191.61	37.5	smokes	1
18	3412...	Male	75.0	1.0	0.0	Yes	Private	Urban	221.29	25.8	smokes	1
19	2745...	Female	60.0	0.0	0.0	No	Private	Urban	89.22	37.8	never_smoked	1
20	2522...	Male	57.0	0.0	1.0	No	Govt_job	Urban	217.08			1
21	7063...	Female	71.0	0.0	0.0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
22	1386...	Female	52.0	1.0	0.0	Yes	Self-emplo...	Urban	233.29	48.9	never_smoked	1
23	6879...	Female	79.0	0.0	0.0	Yes	Self-emplo...	Urban	228.7	26.6	never_smoked	1
24	6477...	Male	82.0	0.0	1.0	Yes	Private	Rural	208.3	32.5		1
25	4219.0	Male	71.0	0.0	0.0	Yes	Private	Urban	102.87	27.2	formerly_smoked	1
26	7082...	Male	80.0	0.0	0.0	Yes	Self-emplo...	Rural	104.12	23.5	never_smoked	1
27	3804...	Female	65.0	0.0	0.0	Yes	Private	Rural	100.98	28.2	formerly_smoked	1
28	6184...	Male	58.0	0.0	0.0	Yes	Private	Rural	189.84			1

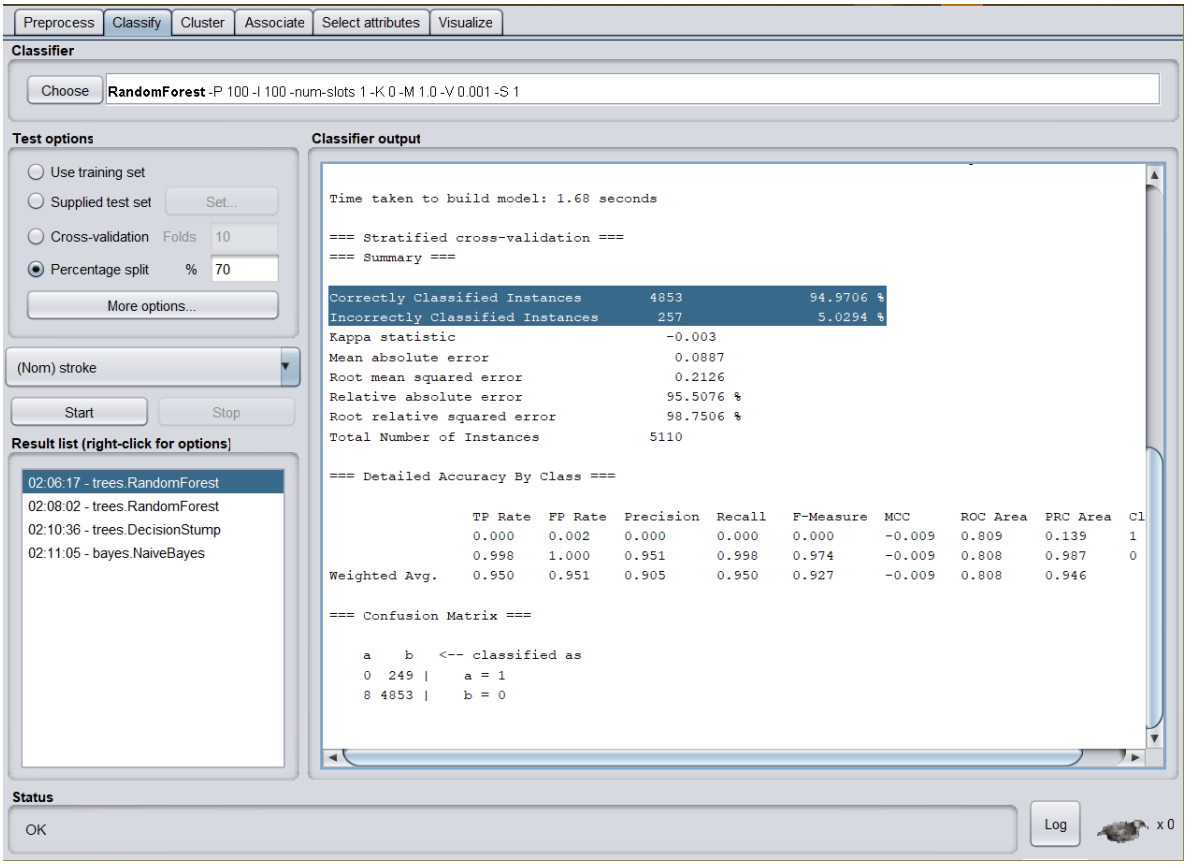
Step 04:

Now we'll classify our dataset by using classifying algorithm. First We'll go to the classify tab then chose a tree base algorithm.

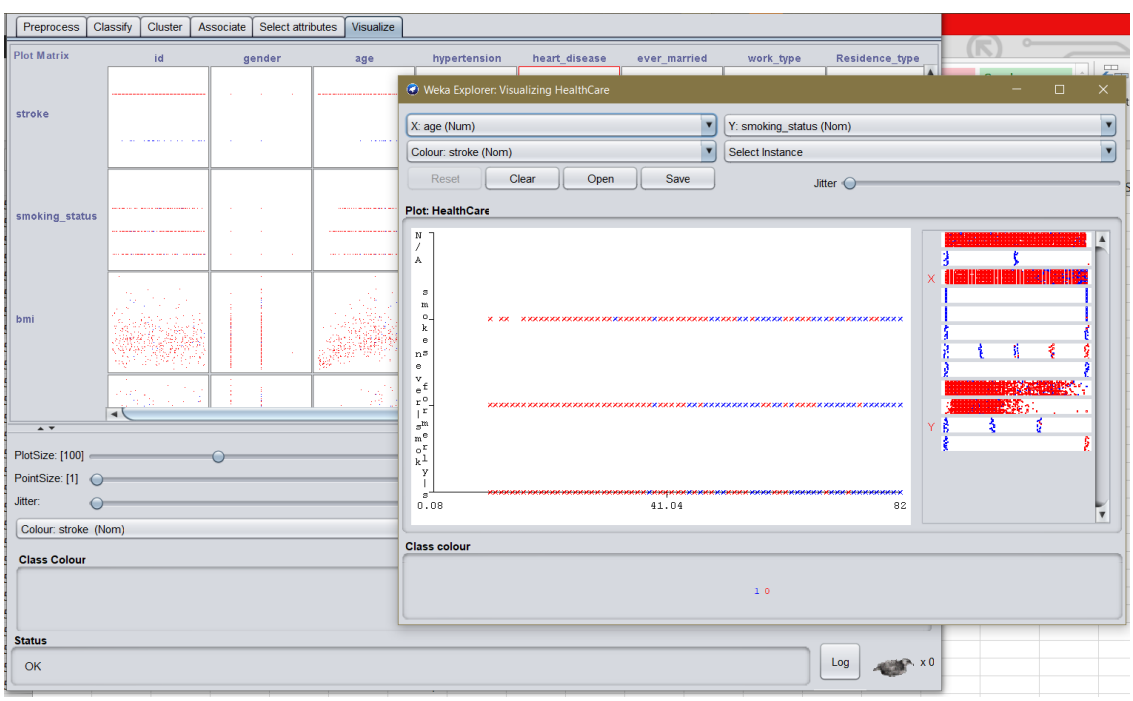
This time we chose RandomForest for the first iteration. Just select it and click start. It'll take some time then show the output.



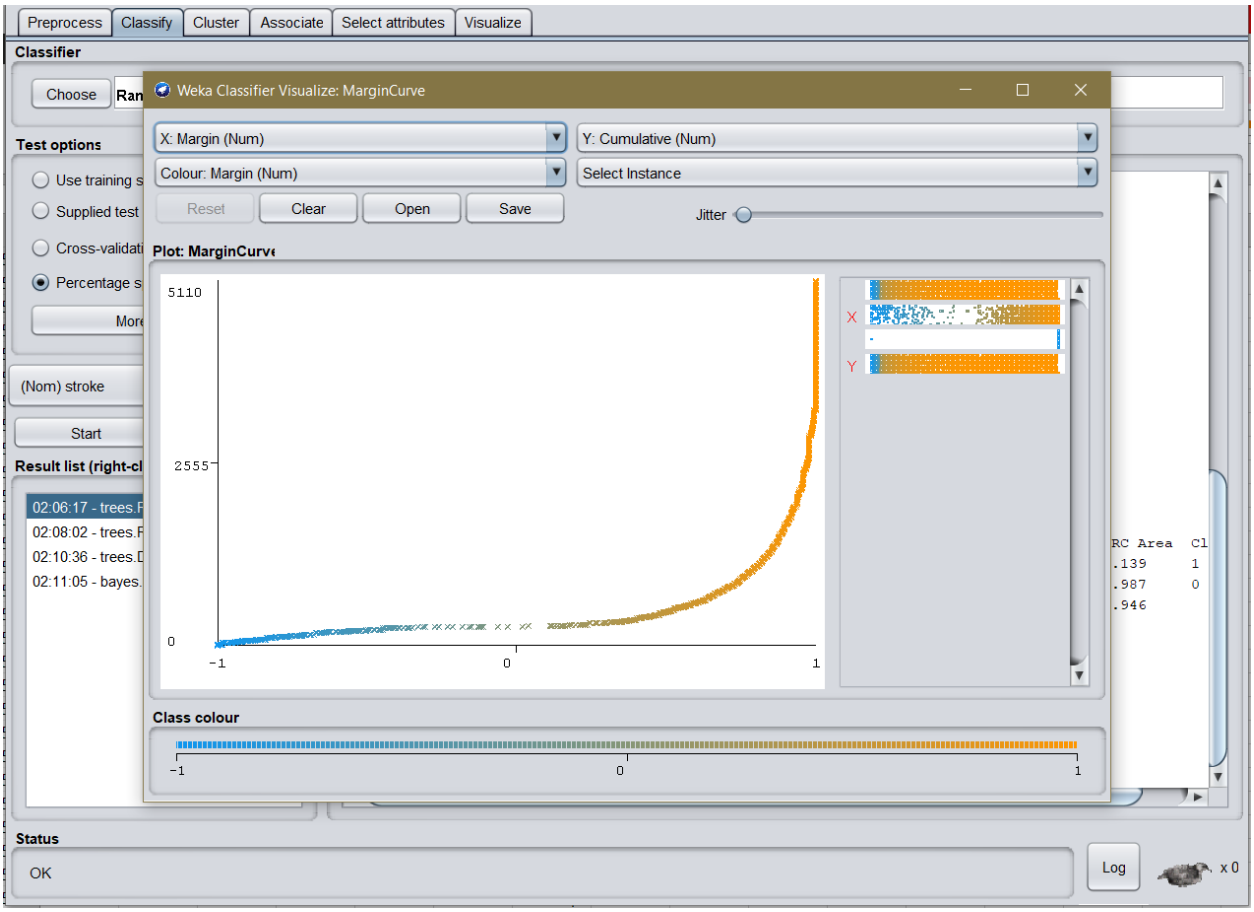
This classifier algorithm classified our trepanning set and shows a accuracy rate of 94.9706% we can also run the data set by changing percentage split from 66 to 70. Because we know the much we train a data set , it'll give a more accurate output.



We can find minimum point value by visualize.

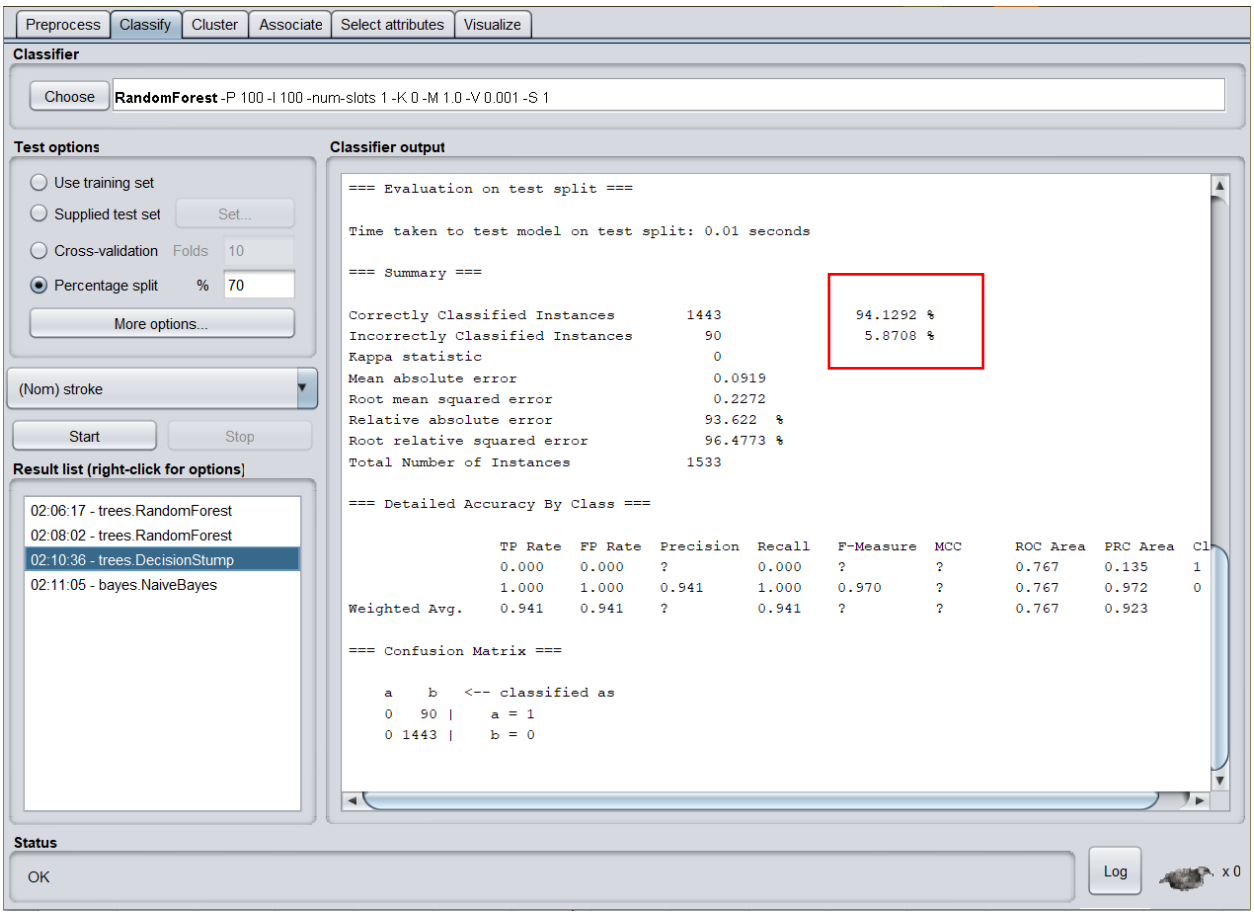


A margin carve also can illustrate the trade of the data set.



Step 05:

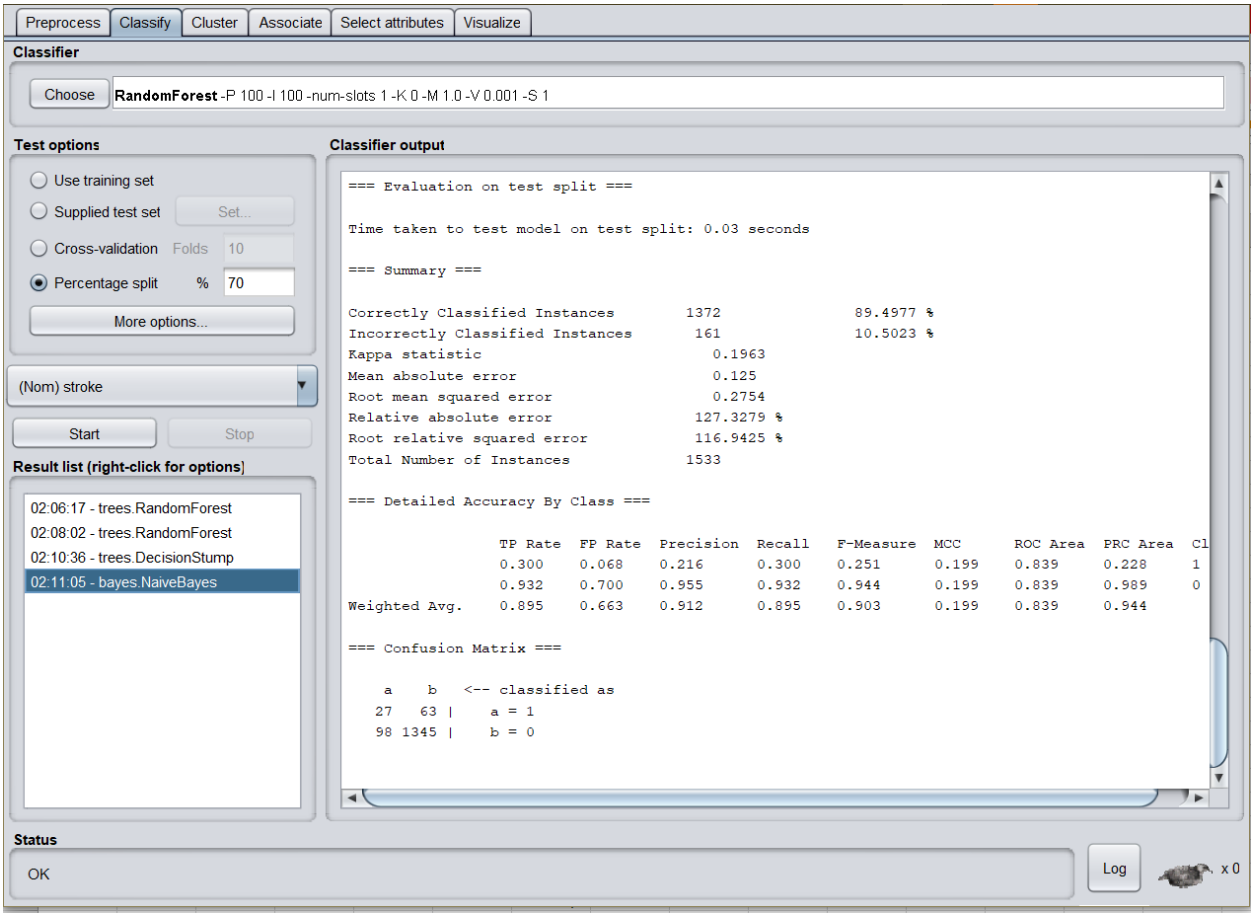
Now we'll test to more classifier algorithm into the data set for expecting more accurate value. For the second one we use DiscussionTree algorithm which shows nearly the similar percentage as RandomPhorest. Which is 94.1292%.



Despite having a little 0.7752% fractional difference DiscussionTree also can provide a batter accuracy as compared to other.

Step 06:

For the last and final test, we've run NevilBay algorithm in our data set. However this time its shows a massive difference, the accuracy rate fall from 94.9706% to 89.7982%.



To sum up we can consider RandomForest algorithm as the best classifying algorithm to develop a frame work that can predict stroke case 3 to 12 month earlier by using data mining technique.