AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Choose an item.

Assignment Cover Sheet

Assignment Title:	Data mining final assignment			
Assignment No:	01		Date of Submission:	29 April 2021
Course Title:	Data warehousing and data mining			
Course Code:	00837		Section:	B
Semester:	Spring	2020-2021	Course Teacher:	Tohedul Islam

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:

No	Name	ID	Program	Signature
1	Ara anjuman	18-39015-3	BSc [CSE]	
2			Choose an item.	
3			Choose an item.	
4			Choose an item.	
5			Choose an item.	
6			Choose an item.	
7			Choose an item.	
8			Choose an item.	
9			Choose an item.	
10			Choose an item.	

Faculty use only		
FACULTYCOMMENTS	Marks Obtained	
	Total Marks	

TASK -1

Introduction :

The data pertains to the recruitment industry in India for the years 2014-2016 and deals with candidate interview attendance for various clients. There are a set of questions that are asked by a recruiter while scheduling the candidate. The answers to these determine whether expected attendance is yes, no or uncertain. The Dataset consists of details of 1047 candidates and the interviews they have attended during the course of the period 2014-2016.

Number of Instance : 1047

Number of attributes : 23

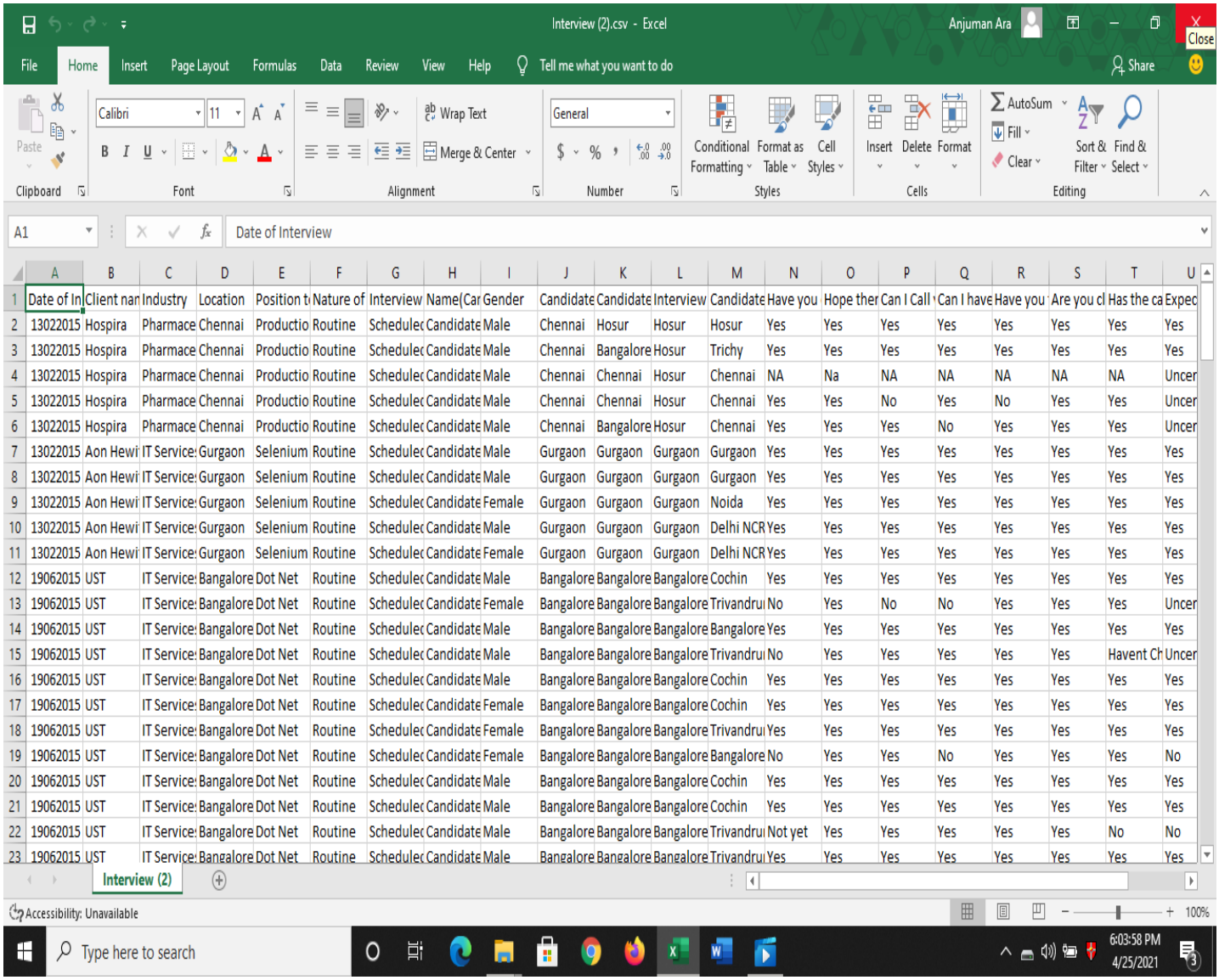
Attributes are :

1. Date of interview
2. Client name
3. Industry
4. Location
5. Position to be closed
6. Nature of skillset
7. Interview type
8. Name
9. Gender
10. Candidate currenct location
11. Candidate job location
12. Interview venue
13. Candidate native location
14. Have you obtained the necessary permission to start at the required time ?
15. Hope there will be no unscheduled meetings
16. Can I call you three hours before the interview and follow up on your attendance for the interview ?
17. Have you taken a printout of your updated resume have you read the JD and understood the same ?
18. Are you clear with the venue details and the landmark ?
19. Has the call letter been shared?
20. Expected attendance
21. Observed attendance
22. Marital status
23. Can I have an alternative number desk number I assure you that I will not trouble you too much?

Available classifier :

1. Naïve Bayes (bayesNet)
2. KNN (Lazy.IBK)
3. Decision tree (REPTree)

Dataset :

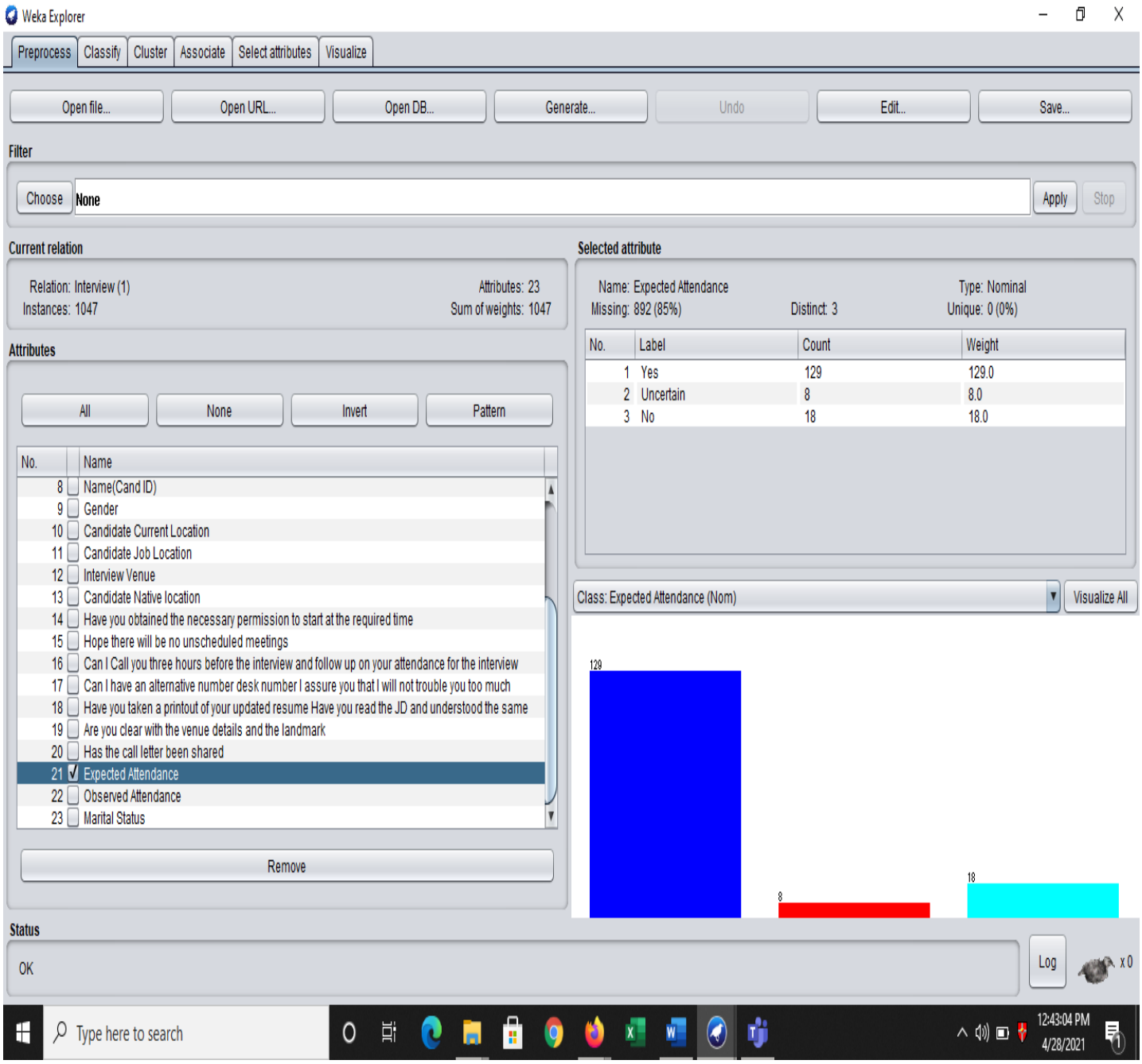


Date of In	Client nan	Industry	Location	Position t	Nature of	Interview	Name(Car	Gender	Candidate	Candidate	Interview	Candidate	Have you	Hope ther	Can I Call	Can I have	Have you	Are you cl	Has the ca	Expec
13022015	Hospira	Pharmace	Chennai	Productio	Routine	Schedulec	Candidate	Male	Chennai	Hosur	Hosur	Hosur	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
13022015	Hospira	Pharmace	Chennai	Productio	Routine	Schedulec	Candidate	Male	Chennai	Bangalore	Hosur	Trichy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
13022015	Hospira	Pharmace	Chennai	Productio	Routine	Schedulec	Candidate	Male	Chennai	Chennai	Hosur	Chennai	NA	Na	NA	NA	NA	NA	NA	Uncer
13022015	Hospira	Pharmace	Chennai	Productio	Routine	Schedulec	Candidate	Male	Chennai	Chennai	Hosur	Chennai	Yes	Yes	No	Yes	No	Yes	Yes	Uncer
13022015	Hospira	Pharmace	Chennai	Productio	Routine	Schedulec	Candidate	Male	Chennai	Bangalore	Hosur	Chennai	Yes	Yes	Yes	No	Yes	Yes	Yes	Uncer
13022015	Aon Hewi	IT Service	Gurgaon	Selenium	Routine	Schedulec	Candidate	Male	Gurgaon	Gurgaon	Gurgaon	Gurgaon	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
13022015	Aon Hewi	IT Service	Gurgaon	Selenium	Routine	Schedulec	Candidate	Male	Gurgaon	Gurgaon	Gurgaon	Gurgaon	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
13022015	Aon Hewi	IT Service	Gurgaon	Selenium	Routine	Schedulec	Candidate	Female	Gurgaon	Gurgaon	Gurgaon	Noida	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
13022015	Aon Hewi	IT Service	Gurgaon	Selenium	Routine	Schedulec	Candidate	Male	Gurgaon	Gurgaon	Gurgaon	Delhi NCR	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
13022015	Aon Hewi	IT Service	Gurgaon	Selenium	Routine	Schedulec	Candidate	Female	Gurgaon	Gurgaon	Gurgaon	Delhi NCR	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Male	Bangalore	Bangalore	Bangalore	Cochin	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Female	Bangalore	Bangalore	Bangalore	Trivandru	No	Yes	No	No	Yes	Yes	Yes	Uncer
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Male	Bangalore	Bangalore	Bangalore	Bangalore	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Male	Bangalore	Bangalore	Bangalore	Trivandru	No	Yes	Yes	Yes	Yes	Yes	Havent Ch	Uncer
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Male	Bangalore	Bangalore	Bangalore	Cochin	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Female	Bangalore	Bangalore	Bangalore	Cochin	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Female	Bangalore	Bangalore	Bangalore	Trivandru	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Female	Bangalore	Bangalore	Bangalore	Bangalore	No	Yes	Yes	No	Yes	Yes	Yes	No
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Male	Bangalore	Bangalore	Bangalore	Cochin	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Male	Bangalore	Bangalore	Bangalore	Cochin	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Male	Bangalore	Bangalore	Bangalore	Trivandru	Not yet	Yes	Yes	Yes	Yes	Yes	No	No
19062015	UST	IT Service	Bangalore	Dot Net	Routine	Schedulec	Candidate	Male	Bangalore	Bangalore	Bangalore	Trivandru	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Figure : Supervised learning Dataset

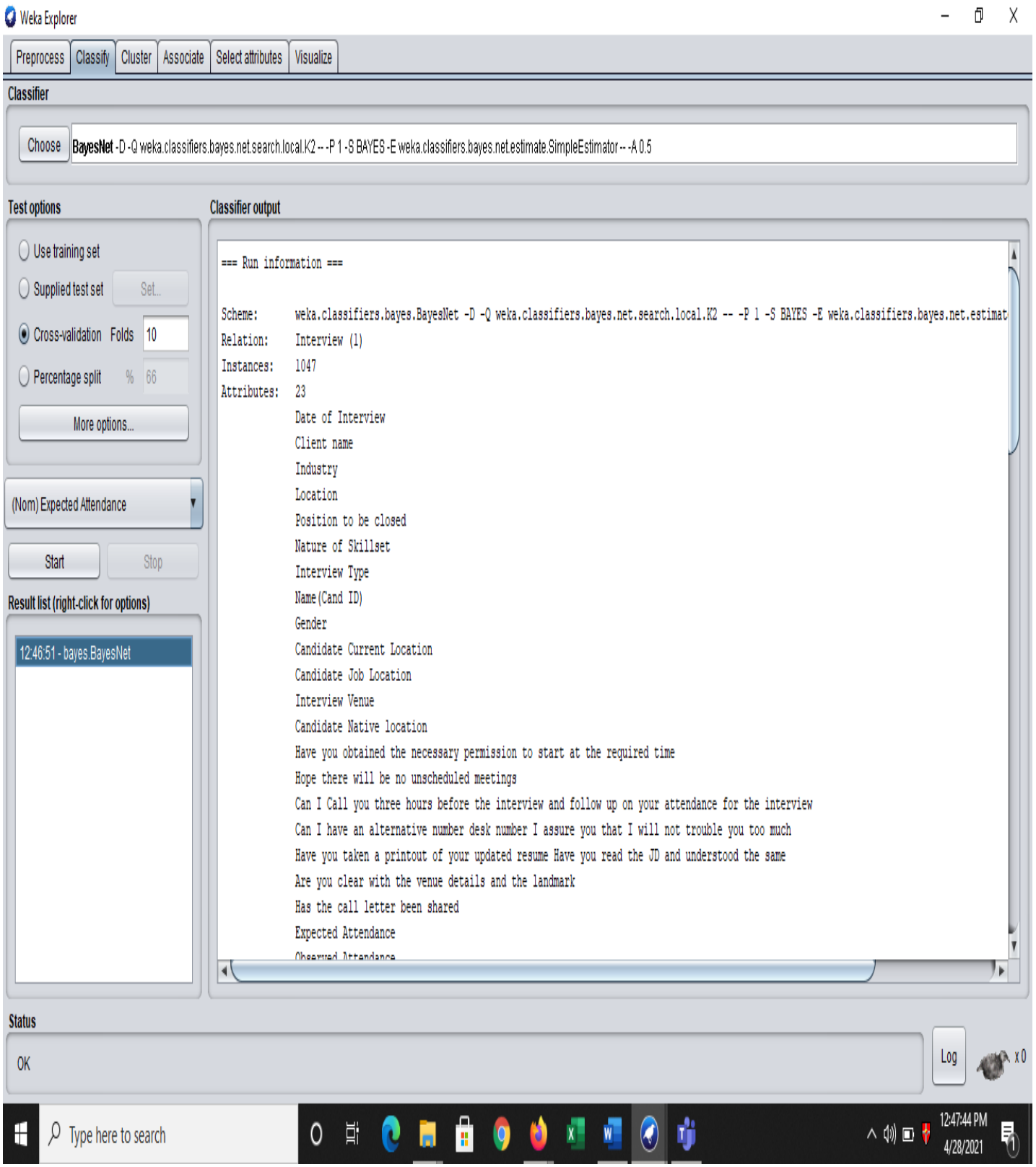
The dataset was collected from Kaggle to test different kind classifires.All fields are already filtered with actual data without null values .so clean data was already prepared.

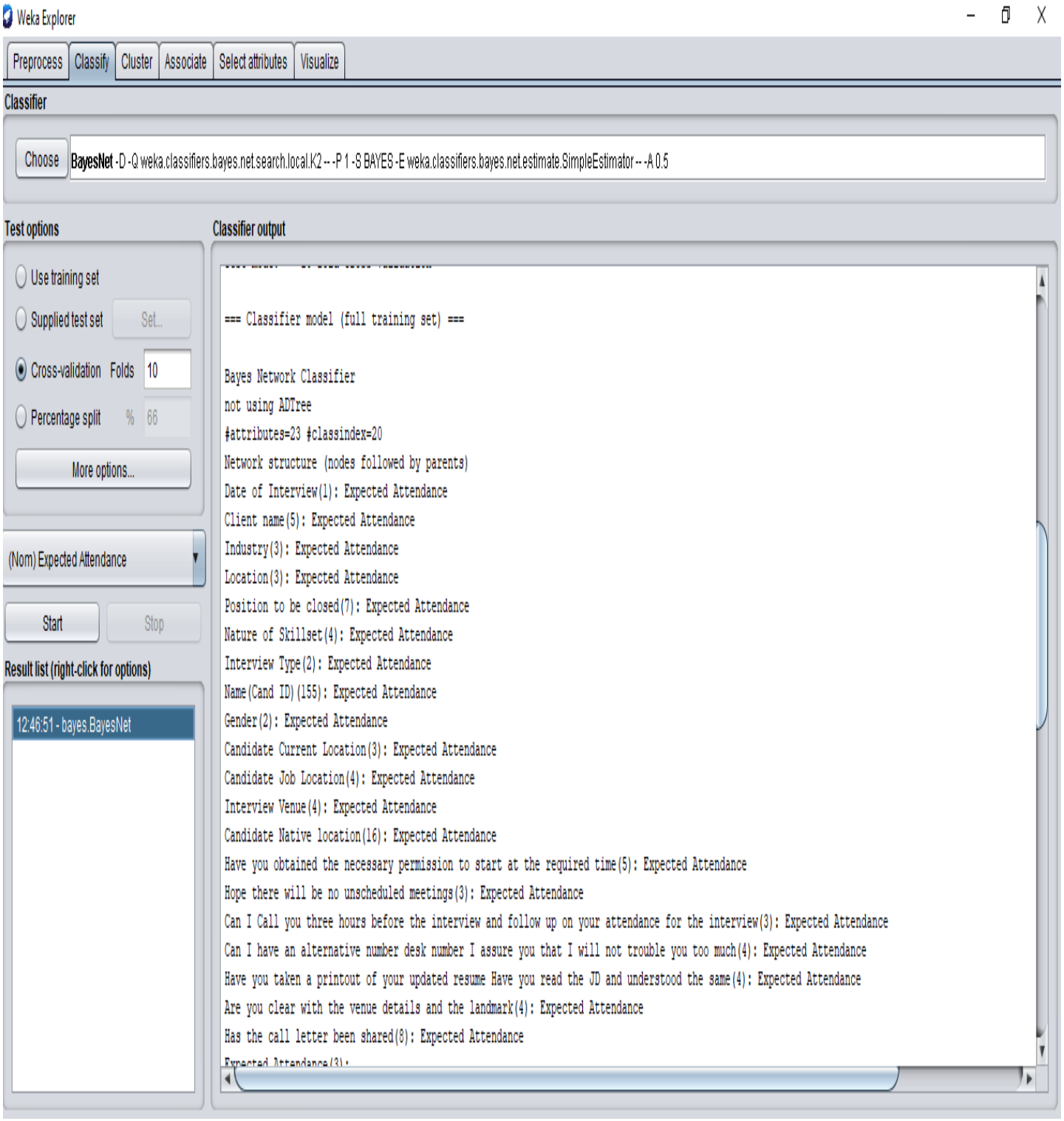
In this below picture we see that for expected attendance class is 3 attribute.the 3 attribute are yes,uncertain,no. For yes label is count 129 and weight 129 and label uncertain count is 8 and weight 8 and for no label count is 18 and weight 18.

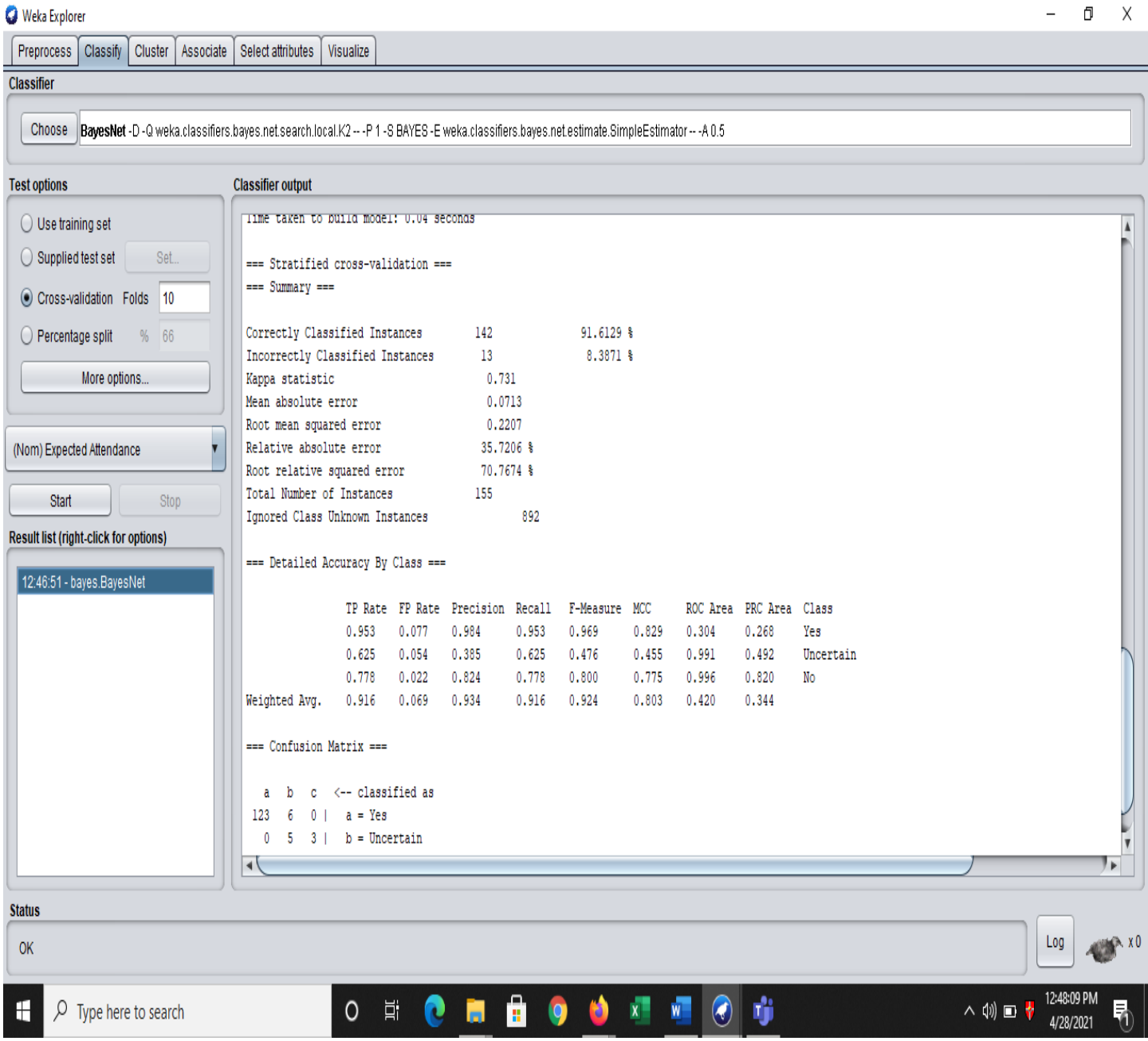


Now we apply 3 different classifier Naïve Bayes (bayesNet), KNN (Lazy.IBK) and Decision tree (REPTree).

Naïve bayes(Bayes.net) :


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Expected Attendance
Start
Stop
Result list (right-click for options)
12:46:51 - bayes.BayesNet
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimat
Relation: Interview (1)
Instances: 1047
Attributes: 23
Date of Interview
Client name
Industry
Location
Position to be closed
Nature of Skillset
Interview Type
Name(Cand ID)
Gender
Candidate Current Location
Candidate Job Location
Interview Venue
Candidate Native location
Have you obtained the necessary permission to start at the required time
Hope there will be no unscheduled meetings
Can I Call you three hours before the interview and follow up on your attendance for the interview
Can I have an alternative number desk number I assure you that I will not trouble you too much
Have you taken a printout of your updated resume Have you read the JD and understood the same
Are you clear with the venue details and the landmark
Has the call letter been shared
Expected Attendance
Observed Attendance
Status
OK
Log
x 0
Type here to search
12:47:44 PM
4/28/2021


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Expected Attendance
Start
Stop
Result list (right-click for options)
12:46:51 - bayes.BayesNet
Classifier output
=== Classifier model (full training set) ===
Bayes Network Classifier
not using ADTree
#attributes=23 #classindex=20
Network structure (nodes followed by parents)
Date of Interview(1): Expected Attendance
Client name(5): Expected Attendance
Industry(3): Expected Attendance
Location(3): Expected Attendance
Position to be closed(7): Expected Attendance
Nature of Skillset(4): Expected Attendance
Interview Type(2): Expected Attendance
Name(Cand ID)(155): Expected Attendance
Gender(2): Expected Attendance
Candidate Current Location(3): Expected Attendance
Candidate Job Location(4): Expected Attendance
Interview Venue(4): Expected Attendance
Candidate Native location(16): Expected Attendance
Have you obtained the necessary permission to start at the required time(5): Expected Attendance
Hope there will be no unscheduled meetings(3): Expected Attendance
Can I Call you three hours before the interview and follow up on your attendance for the interview(3): Expected Attendance
Can I have an alternative number desk number I assure you that I will not trouble you too much(4): Expected Attendance
Have you taken a printout of your updated resume Have you read the JD and understood the same(4): Expected Attendance
Are you clear with the venue details and the landmark(4): Expected Attendance
Has the call letter been shared(8): Expected Attendance



Cross validation summary :

Correctly classified instances : 142 91.6129

Incorrectly classified instances : 13 8.3871

Kappa statistic : 0.731

Mean absolute error : 0.0713

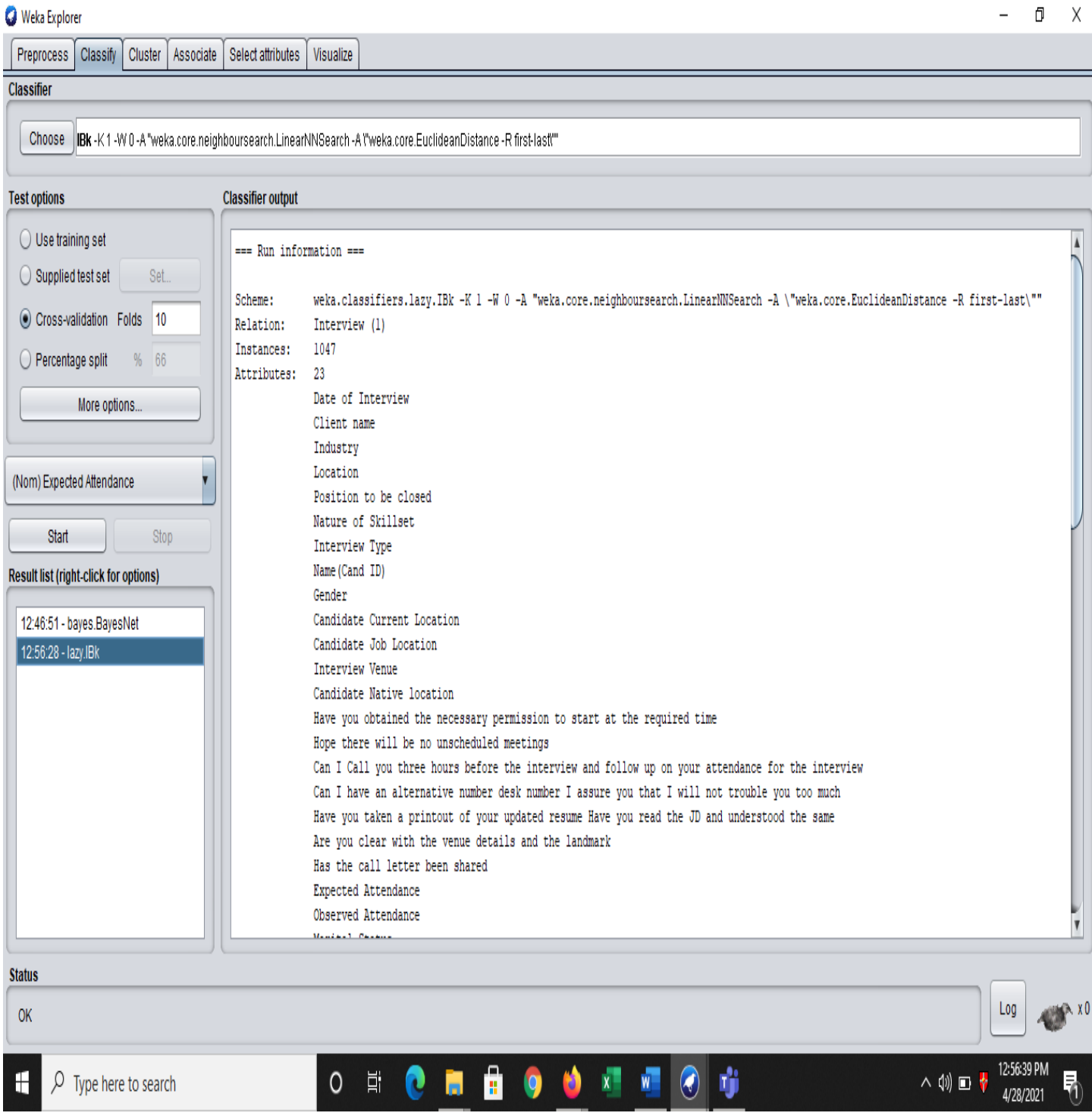
Root mean squared error : 0.2207

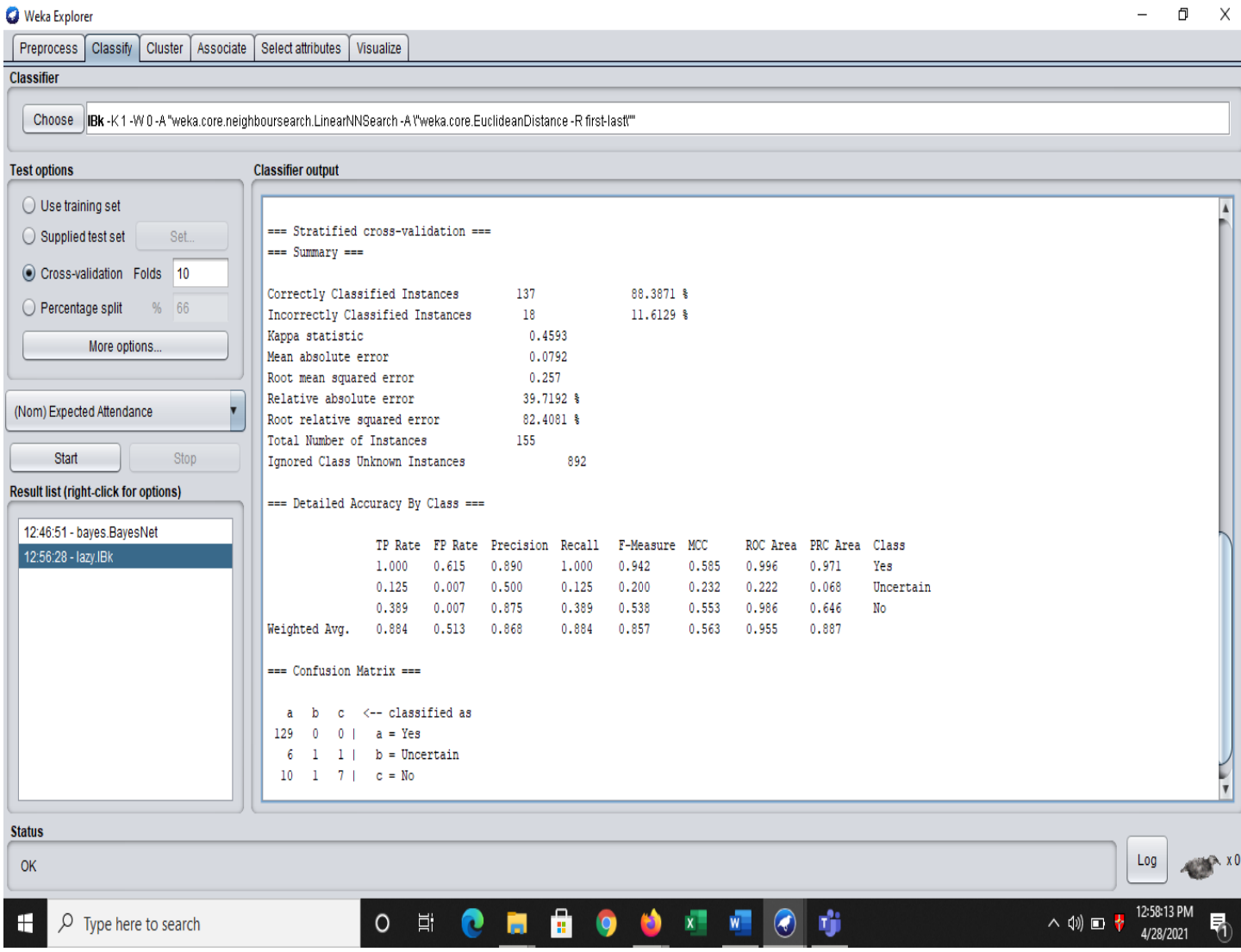
Relative absolute error : 35. 7206

Root relative squared error : 70.7674

Total number of instances : 155

KNN(K nearest neighbors algorithm) / Lazy.IBK :


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Expected Attendance
Start
Stop
Result list (right-click for options)
12:46:51 - bayes.BayesNet
12:56:28 - lazy.IBk
Classifier output
=== Run information ===
Scheme: weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation: Interview (1)
Instances: 1047
Attributes: 23
Date of Interview
Client name
Industry
Location
Position to be closed
Nature of Skillset
Interview Type
Name(Cand ID)
Gender
Candidate Current Location
Candidate Job Location
Interview Venue
Candidate Native location
Have you obtained the necessary permission to start at the required time
Hope there will be no unscheduled meetings
Can I Call you three hours before the interview and follow up on your attendance for the interview
Can I have an alternative number desk number I assure you that I will not trouble you too much
Have you taken a printout of your updated resume Have you read the JD and understood the same
Are you clear with the venue details and the landmark
Has the call letter been shared
Expected Attendance
Observed Attendance
Status
OK
Log
x 0
Type here to search
12:56:39 PM
4/28/2021



Cross validation summary :

Correctly classified instances : 137 88.3871

Incorrectly classified instances : 18 11.6129

Kappa statistic : 0.4593

Mean absolute error : 0.0792

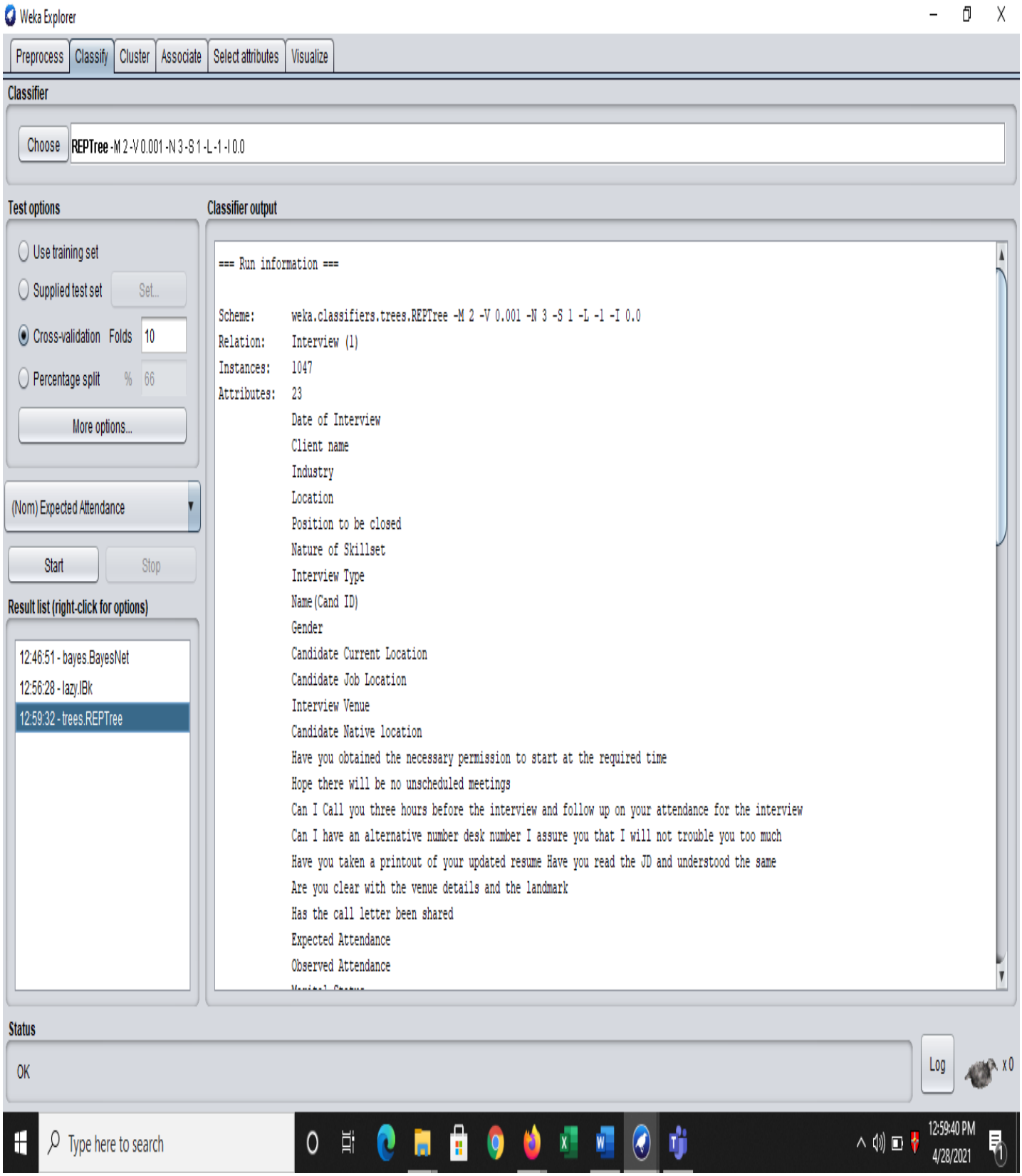
Root mean squared error : 0.257

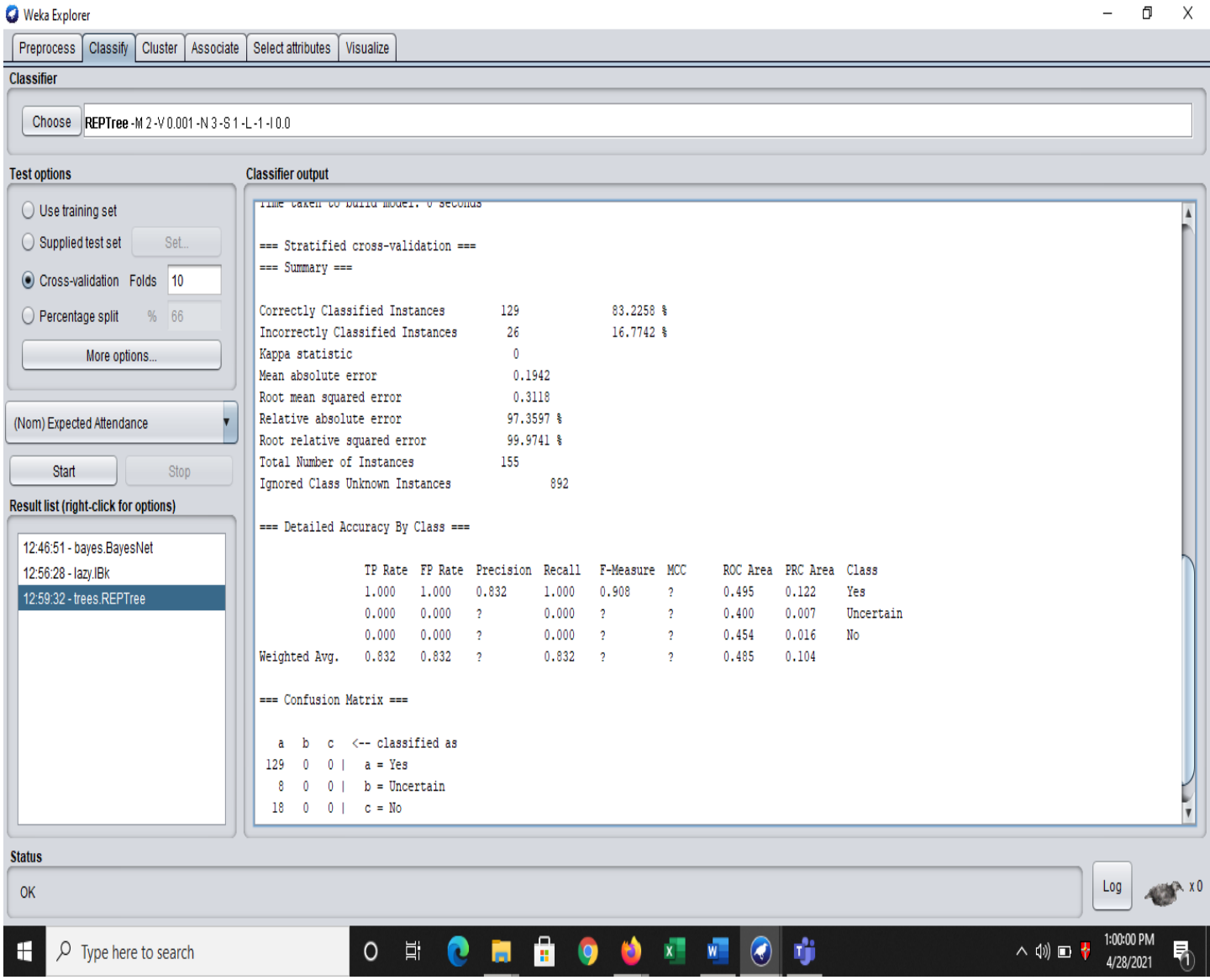
Relative absolute error : 39.7192

Root relative squared error : 82.4081

Total number of instances : 155

Decision tree(REPTree) :


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Expected Attendance
Start
Stop
Result list (right-click for options)
12:46:51 - bayes.BayesNet
12:56:28 - lazy.IBk
12:59:32 - trees.REPTree
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0
Relation: Interview (1)
Instances: 1047
Attributes: 23
Date of Interview
Client name
Industry
Location
Position to be closed
Nature of Skillset
Interview Type
Name(Cand ID)
Gender
Candidate Current Location
Candidate Job Location
Interview Venue
Candidate Native location
Have you obtained the necessary permission to start at the required time
Hope there will be no unscheduled meetings
Can I Call you three hours before the interview and follow up on your attendance for the interview
Can I have an alternative number desk number I assure you that I will not trouble you too much
Have you taken a printout of your updated resume Have you read the JD and understood the same
Are you clear with the venue details and the landmark
Has the call letter been shared
Expected Attendance
Observed Attendance
Status
OK
Log
x 0
Type here to search
12:59:40 PM
4/28/2021



Cross validation summary :

Correctly classified instances : 129 83.2258

Incorrectly classified instances : 26 16.7742

Kappa statistic : 0

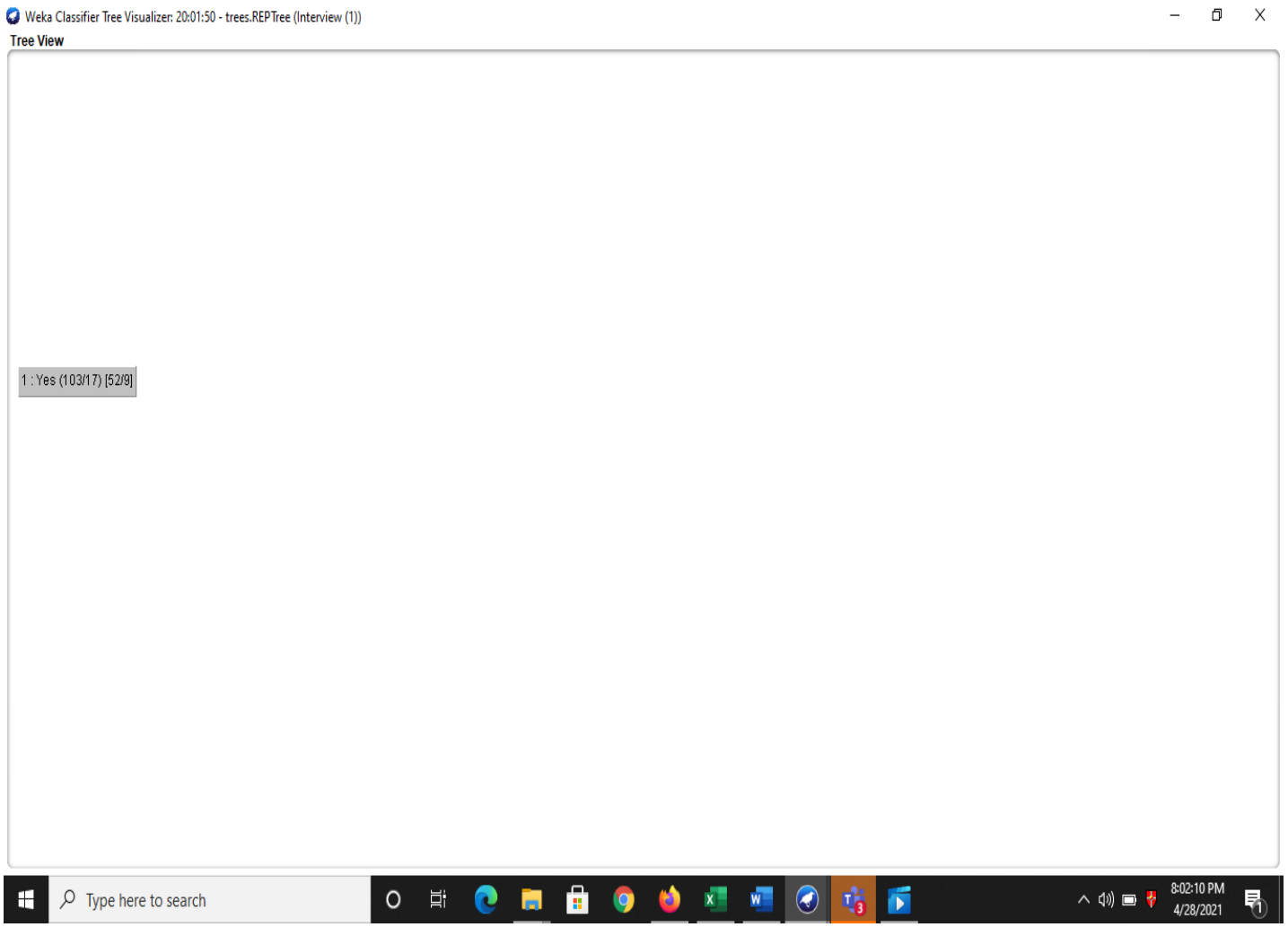
Mean absolute error : 0.1942

Root mean squared error : 0.3118

Relative absolute error : 97.3597

Root relative squared error : 99.9741

Total number of instances : 155



Model : REPTree tree view model visualization

Result : The main objective was to find the best relational algorithm to find which algorithm was best for this dataset.So we can see that bayes,lazy.IBK ,trees.REPTress has different correctly classified instances about expected attendance.

Discussion : For this data set correctly classified instances are

1. Naive bayes = 91.6129%
2. lazy.IBK = 88.3871%

3. REPTree = 83.2258%

So we can say that for this data set fastest dataset is Naïve Bayes .The perchentange rate for correctly classified instances are high for Naïve bayes.

So ,

Naïve Bayes > lazy.IBK > REPTree.

Reference :

https://www.kaggle.com/vishnusraghavan/the-interview-attendance-problem.

Task -2

Introduction :

The dataset covers 13 drugs cover 17 age group .This directory contains data behind the story how baby boomers get high.Drug addiction typically begins at a young age with higher rates of addiction seen in adolescents and young adults.The data set was collected from from Kaggle in csv format.In this data set there are total 28 attributes.

Number of instances : 17

Attributes : 28

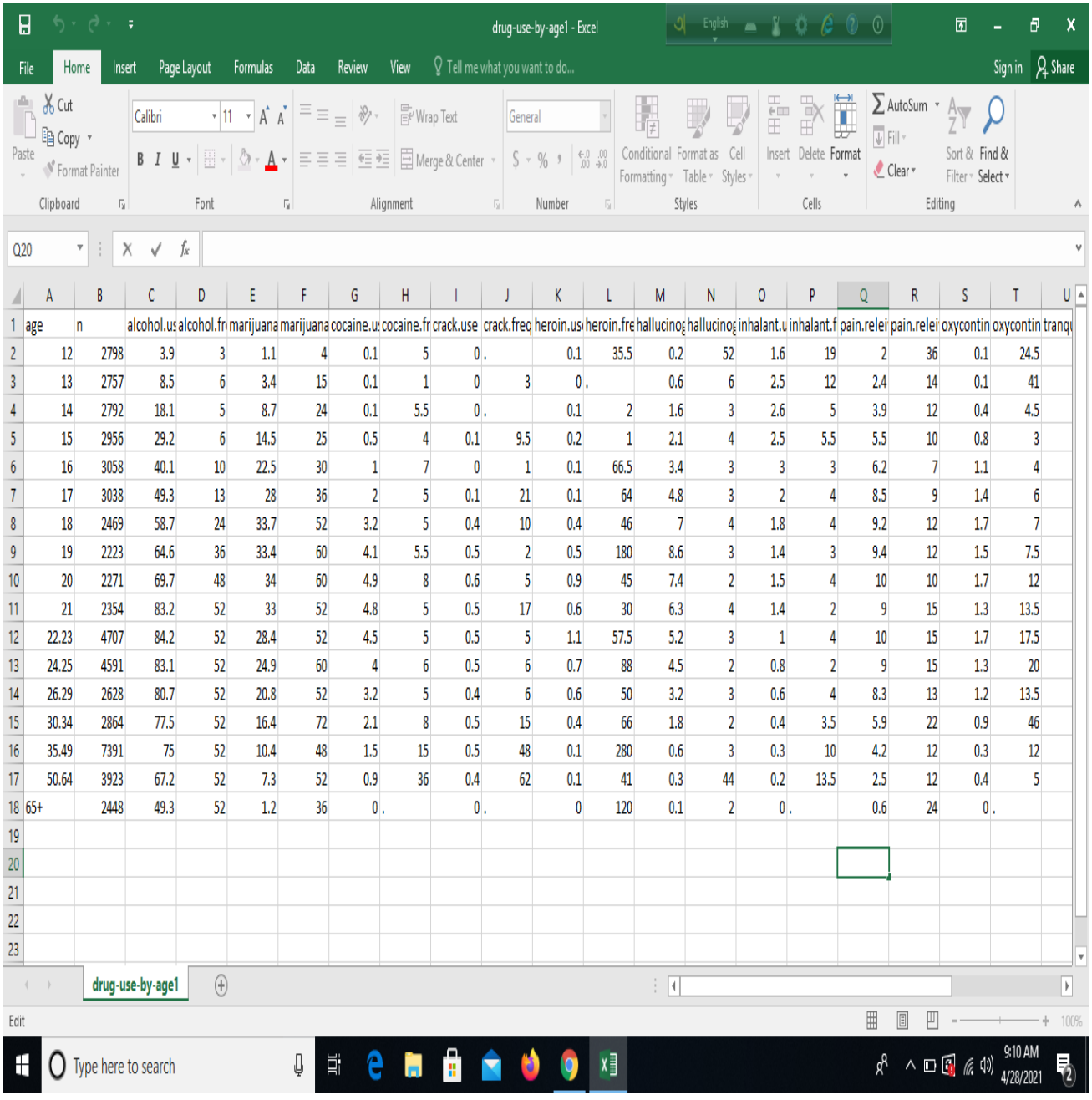
Attributes are :

1. Age
2. Alcohol .frequency
3. Alcohol.use
4. Marijuana.use
5. Marijuana.frequency
6. Cocaine.use
7. Cocaine.frquency
8. Crack.use
9. Crack.frequency
10. Heroin.use
11. Heroin.frequency
12. Hallucinogen.use
13. Hallucinogen.frequency
14. Inhalant.use
15. Inhalant.frequency
16. Pain.releiver.use
17. Pain.releiver.frequency
18. Oxycontin.use
19. Oxycontin.frequency
20. Tranquilizer.use
21. Tranquilizer.frequency
22. Stimulant.use
23. Stimulant.frequency
24. Meth.use
25. Meth.frequency
26. Sedative.use

27. Sedative.frequency
28. N

Available classifier : Hierarchicalclusterer

Dataset :

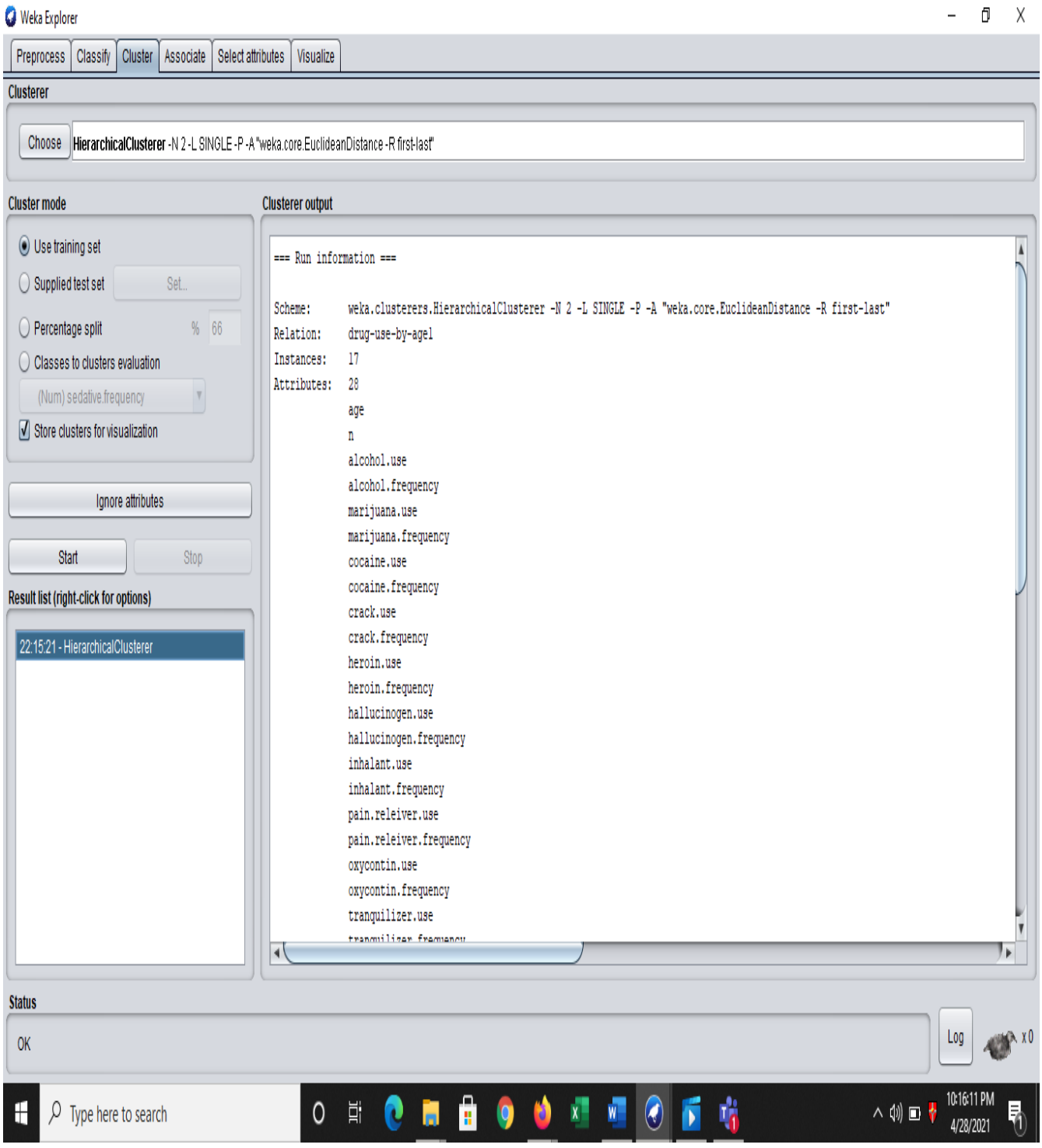


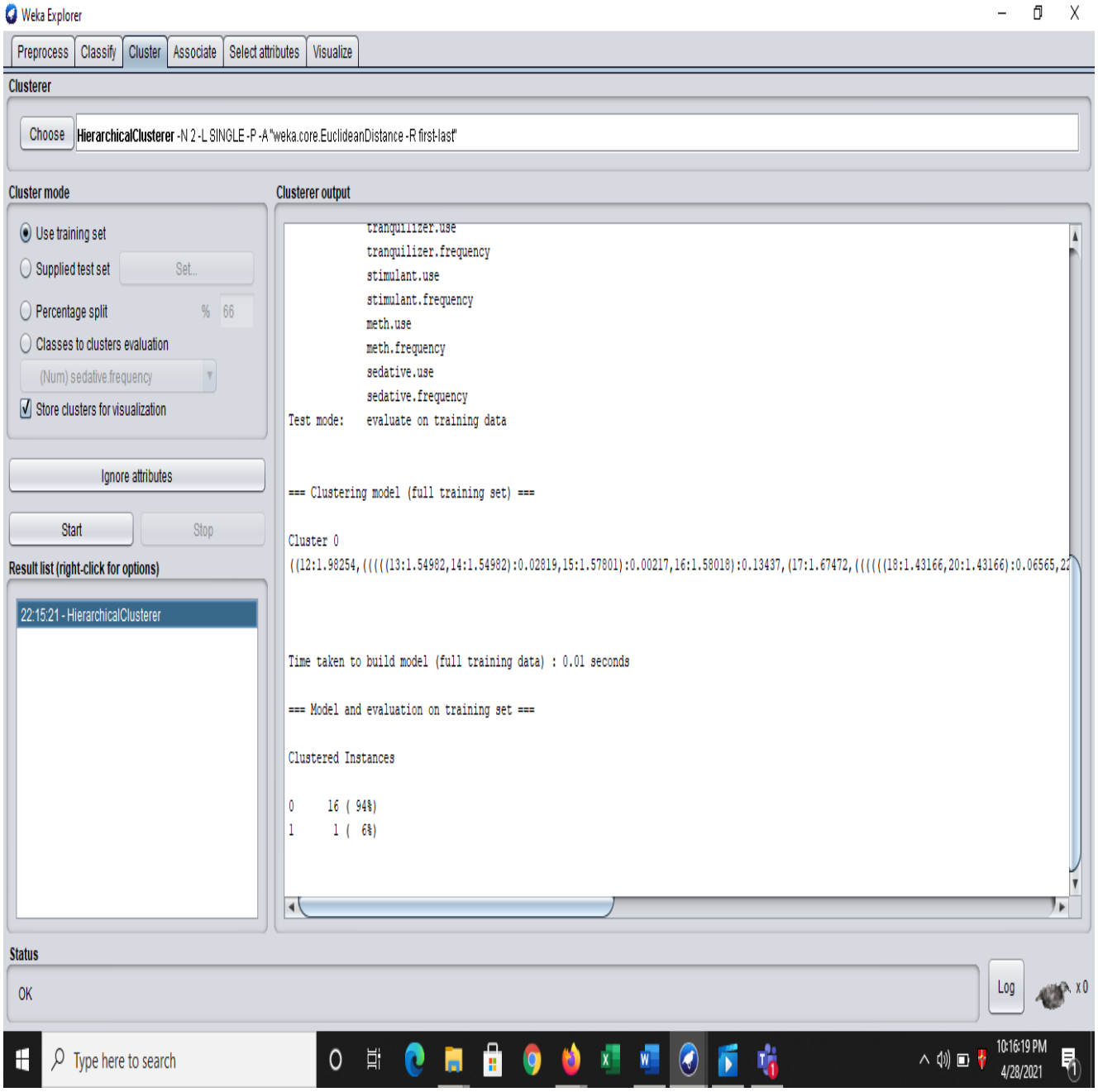
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	age	n	alcohol.us	alcohol.fr	marijuana	marijuana	cocaine.u	cocaine.fr	crack.use	crack.freq	heroin.us	heroin.fre	hallucino	hallucino	inhalant.u	inhalant.f	pain.relei	pain.relei	oxycontin	oxycontin	tranq
2	12	2798	3.9	3	1.1	4	0.1	5	0	.	0.1	35.5	0.2	52	1.6	19	2	36	0.1	24.5	
3	13	2757	8.5	6	3.4	15	0.1	1	0	3	0	.	0.6	6	2.5	12	2.4	14	0.1	41	
4	14	2792	18.1	5	8.7	24	0.1	5.5	0	.	0.1	2	1.6	3	2.6	5	3.9	12	0.4	4.5	
5	15	2956	29.2	6	14.5	25	0.5	4	0.1	9.5	0.2	1	2.1	4	2.5	5.5	5.5	10	0.8	3	
6	16	3058	40.1	10	22.5	30	1	7	0	1	0.1	66.5	3.4	3	3	3	6.2	7	1.1	4	
7	17	3038	49.3	13	28	36	2	5	0.1	21	0.1	64	4.8	3	2	4	8.5	9	1.4	6	
8	18	2469	58.7	24	33.7	52	3.2	5	0.4	10	0.4	46	7	4	1.8	4	9.2	12	1.7	7	
9	19	2223	64.6	36	33.4	60	4.1	5.5	0.5	2	0.5	180	8.6	3	1.4	3	9.4	12	1.5	7.5	
10	20	2271	69.7	48	34	60	4.9	8	0.6	5	0.9	45	7.4	2	1.5	4	10	10	1.7	12	
11	21	2354	83.2	52	33	52	4.8	5	0.5	17	0.6	30	6.3	4	1.4	2	9	15	1.3	13.5	
12	22.23	4707	84.2	52	28.4	52	4.5	5	0.5	5	1.1	57.5	5.2	3	1	4	10	15	1.7	17.5	
13	24.25	4591	83.1	52	24.9	60	4	6	0.5	6	0.7	88	4.5	2	0.8	2	9	15	1.3	20	
14	26.29	2628	80.7	52	20.8	52	3.2	5	0.4	6	0.6	50	3.2	3	0.6	4	8.3	13	1.2	13.5	
15	30.34	2864	77.5	52	16.4	72	2.1	8	0.5	15	0.4	66	1.8	2	0.4	3.5	5.9	22	0.9	46	
16	35.49	7391	75	52	10.4	48	1.5	15	0.5	48	0.1	280	0.6	3	0.3	10	4.2	12	0.3	12	
17	50.64	3923	67.2	52	7.3	52	0.9	36	0.4	62	0.1	41	0.3	44	0.2	13.5	2.5	12	0.4	5	
18	65+	2448	49.3	52	1.2	36	0	.	0	.	0	120	0.1	2	0	.	0.6	24	0	.	
19																					
20																					
21																					
22																					
23																					

Figure : unsupervised learnning dataset

The dataset was collected from Kaggle to test hierarchical cluster .so clean data was already prepared.

Hierarchical cluster :


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) sedative.frequency
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
22:15:21 - HierarchicalClusterer
Clusterer output
=== Run information ===
Scheme: weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Relation: drug-use-by-age1
Instances: 17
Attributes: 28
age
n
alcohol.use
alcohol.frequency
marijuana.use
marijuana.frequency
cocaine.use
cocaine.frequency
crack.use
crack.frequency
heroin.use
heroin.frequency
hallucinogen.use
hallucinogen.frequency
inhalant.use
inhalant.frequency
pain.releiver.use
pain.releiver.frequency
oxycontin.use
oxycontin.frequency
tranquilizer.use
Status
OK
Log
x 0
Type here to search
10:16:11 PM
4/28/2021



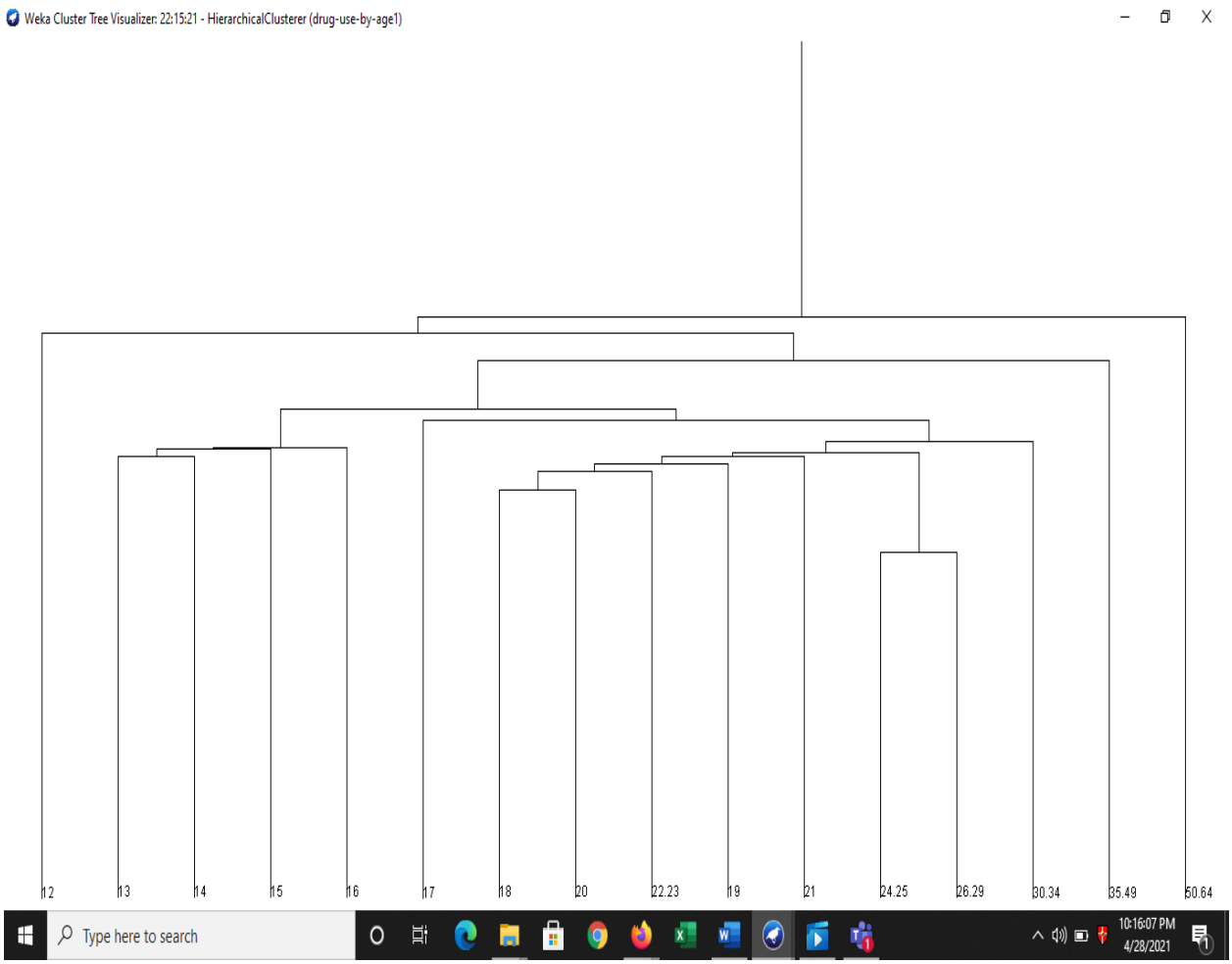
Result :

Clustered instances

1 16(94%)

2 1(6%)

Hierarchical cluster dendrogram :



In the dendogram above,the height of the dendogram indicates the order in which the clusters are joined.It is important to appreciate that the dendogram is a summary of the distance matrix. In this dendogram ages 24.25 and 26.29 is much closer than 12 and 50.64.also 13 and 14 are much closer.To use some jargon ,a dendogram is only accurate when data satisfies the ultrametric tree inequality,and this is unlikely for any real world data.the consequence of the information loss is that the dendograms are most accurate at the bottom,showing which items are very similar.

Discussion : By applying hierarchical cluster we find clustered instances 0 16(94%) And 1 1(6%).And dendogram shows which clusterd are joined.

Reference : https://www.kaggle.com/tunguz/drug-use-by-age