A REPORT ON

DRUG CONSUMPTION

JULY 2018

Submitted to:

Dr. Ankush Mittal
Director
Raman Classes

Submitted By:

Rajat Tyagi
B.tech(CSE)
GEU

CANDIDATE'S DECLARATION

We hereby declare that I own the full responsibility for the information, results etc. provided in the projects in this report

submitted to **RWX Technology** for the award of certificate of **Certificate of Summer Internship** done during June-July 2018. We have taken care in all respect to honor the intellectual property right and have acknowledged the contribution of others for using them in academic purpose and further declare that in case of any violation of intellectual property right or copyright we shall be fully responsible for the same. Our supervisor should not be held responsible for full or partial violation of copyright or intellectual property right.

NAME : Rajat Tyagi

DATE : 30-7-18

PLACE : Roorkee

Abstract

Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed.
The Weka machine learning workbench is a modern platform for applied machine learning. Weka is an acronym which stands for Waikato Environment for Knowledge Analysis. It is also the name of a New Zealand bird the Weka.The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as the new data becomes available. In this mini project, various machine learning techniques have been applied on the data set taken from [1]. Among the operations performed upon the data set, Classification and selection of attributes or more precisely, feature selection are the major ones. The Weka 3.8 software is used for the analysis of the data set and all the operations mentioned above. Various graphs have been plotted for different attributes and parameters. The data set taken is drug consumption datase quantified data set donated in year 2016.

TABLE OF CONTENT

CLASSIFICATION OF DRUG CONSUMPTION DATASE:

INTRODUCTION

MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

SOME MACHINE LEARNING METHODS

Machine learning algorithms are often categorized as supervised or unsupervised.

- **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
- In contrast, **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
- **Semi-supervised machine learning algorithms** fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiringunlabeled data generally doesn't require additional resources.
- **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

WEKA (3.8)

What is Weka? Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.
Datasets in Weka Each entry in a dataset is an instance of the java class: − weka.core.Instance
Each instance consists of a number of attributes.

DATA SET DESCRIPTION

Database contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real-valued. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day. Database contains 18 classification problems. Each of independent label variables contains seven classes: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day".

Problem which can be solved:
* Seven class classifications for each drug separately.
* Problem can be transformed to binary classification by union of part of classes into one new class. For example, "Never Used", "Used over a Decade Ago" form class "Non-user" and all other classes form class "User".
* The best binarization of classes for each attribute.
* Evaluation of risk to be drug consumer for each drug.

ATTRIBUTES

1. ID is number of record in original database. Cannot be related to participant. It can be used for reference only.
2. Age (Real) is age of participant
3. Gender (Real) is gender of participant:
4. Education (Real) is level of education of participant
5. Country (Real) is country of current residence of participant
6. Ethnicity (Real) is ethnicity of participant
7. Nscore (Real) is NEO-FFI-R Neuroticism.
8. Escore (Real) is NEO-FFI-R Extraversion.
9. Oscore (Real) is NEO-FFI-R Openness to experience
10. Ascore (Real) is NEO-FFI-R Agreeableness.
11. Cscore (Real) is NEO-FFI-R Conscientiousness
12. Impulsive (Real) is impulsiveness measured by BIS-11.
13. SS (Real) is sensation seeing measured by ImpSS
14. Alcohol is class of alcohol consumption.
15. Amphet is class of amphetamines consumption.

16. Amyl is class of amyl nitrite consumption
17. Benzos is class of benzodiazepine consumption.
18. Caff is class of caffeine consumption
19. Cannabis is class of cannabis consumption.
20. Choc is class of chocolate consumption.
21. Coke is class of cocaine consumption.
22. Crack is class of crack consumption.
23. Ecstasy is class of ecstasy consumption.
24. Heroin is class of heroin consumption.
25. Ketamine is class of ketamine consumption.
26. Legalh is class of legal highs consumption.
27. LSD is class of alcohol consumption
28. Meth is class of methadone consumption.
29. Mushrooms is class of magic mushrooms consumption.
30. Nicotine is class of nicotine consumption.
31. Semer is class of fictitious drug Semeron consumption.
32. VSA is class of volatile substance abuse consumption.

TECHNIQUES USED

J48 DECISION TREE

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found . This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable . J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Basic Steps in the Algorithm

(i) In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class.

(ii) The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.

(iii) Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

The steps performed for data mining in WEKA are:

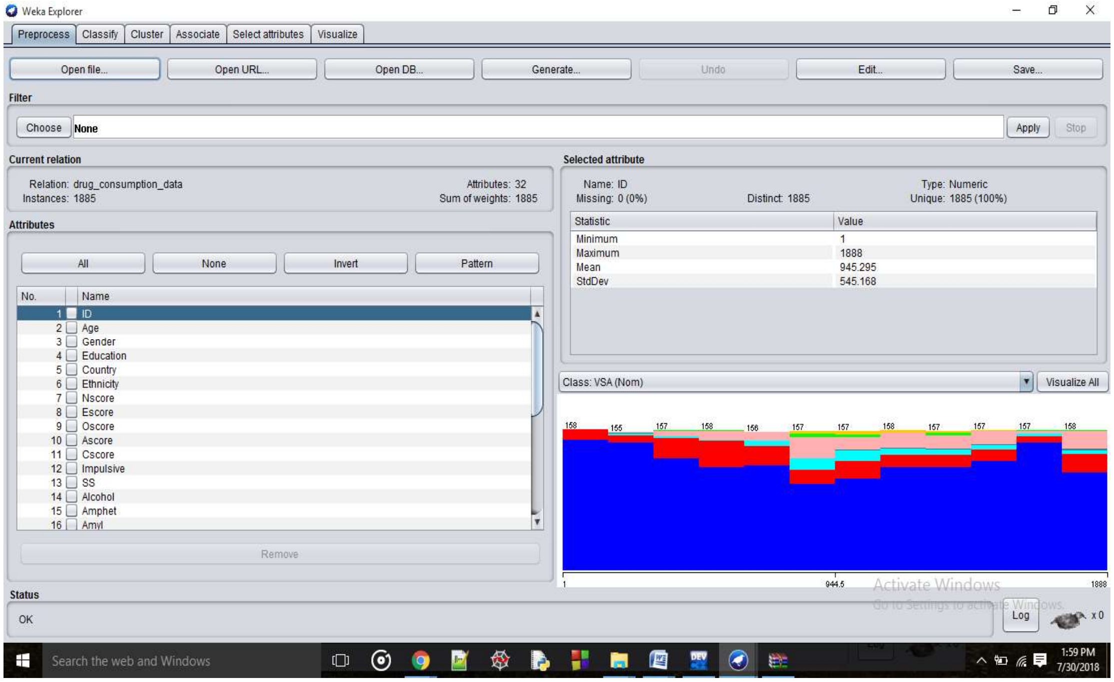
- Data pre-processing and visualization
- Attribute selection
- Classification (Decision trees)
- Prediction (Nearest neighbour)
- Model evaluation
- Clustering (Cobweb, K-means) • Association rules

Test Options included using training set, supplying test set, cross validation and percentage splitting. Accuracy of the machine was tested through the data in many methods: by selecting all the original attributes and then training the machine, then by deselecting some features and then calculating the accuracy. For feature selection the attribute evaluator used is **CfsSubsetEvaluator**and the method used is **BestFirst method.**

RESULTS AND DISCUSSION

LOADING DATASET OR PRE –PROCESSING

This dataset loaded in comma separated value (.csv) format



Filter: None

Current relation

Relation: drug_consumption_data Attributes: 32
Instances: 1885 Sum of weights: 1885

Attributes

No.	Name
1	☐ ID
2	☐ Age
3	☐ Gender
4	☐ Education
5	☐ Country
6	☐ Ethnicity
7	☐ Nscore
8	☐ Escore
9	☐ Oscore
10	☐ Ascore
11	☐ Cscore
12	☐ Impulsive
13	☐ SS
14	☐ Alcohol
15	☐ Amphet
16	☐ Amyl

Selected attribute

Name: ID Type: Numeric
Missing: 0 (0%) Distinct: 1885 Unique: 1885 (100%)

Statistic	Value
Minimum	1
Maximum	1888
Mean	945.295
StdDev	545.168

Class: VSA (Nom)

Status

OK

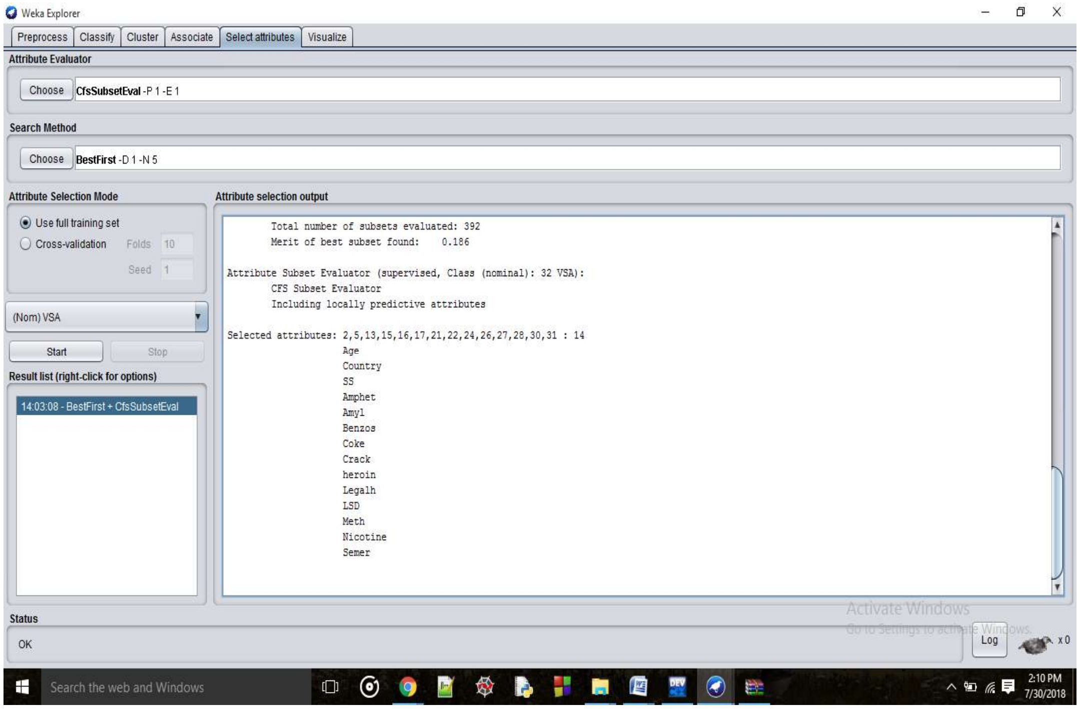
Activate Windows
Go to Settings to activate Windows.

Figure 1: Pre-processing Window where the data is loaded.

ATTRIBUTE SELECTION

- **Selected attributes** : 2,5,13,15,16,17,21,22,24,26,27,28,30,31 : 14

  Age
  Country
  SS
  Amphet
  Amyl
  Benzos
  Coke
  Crack
  heroin
  Legalh
  LSD
  Meth
  Nicotine
  Semer



Attribute Evaluator: CfsSubsetEval -P 1 -E 1

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set
○ Cross-validation Folds 10 Seed 1

(Nom) VSA

Result list (right-click for options): 14:03:08 - BestFirst + CfsSubsetEval

Attribute selection output

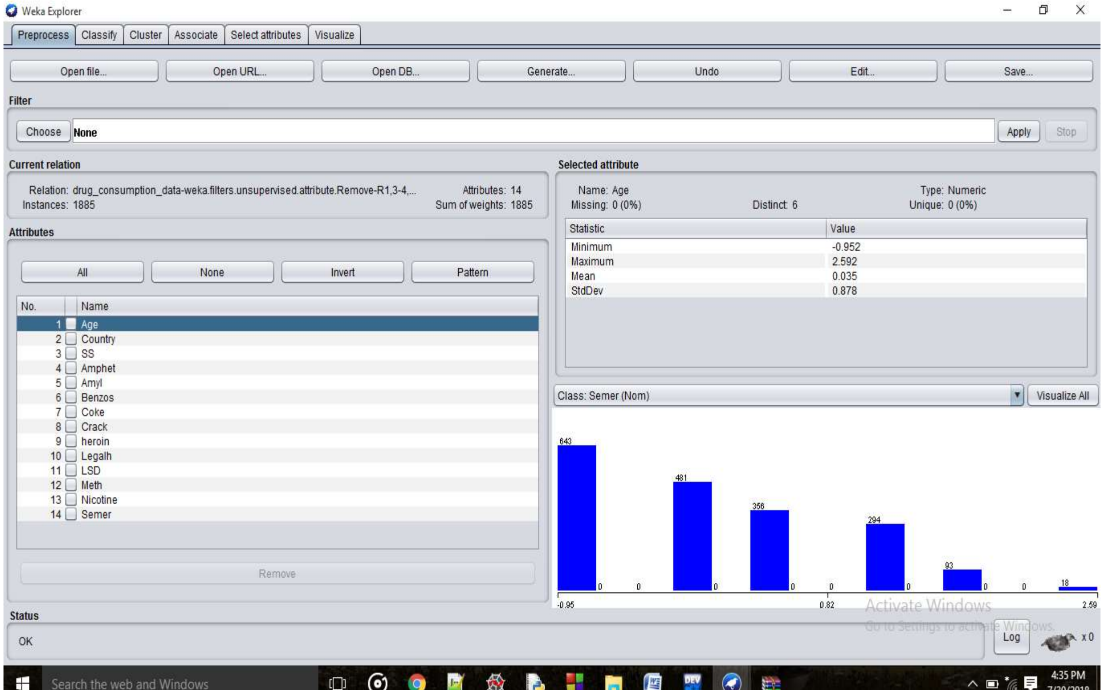
```
        Total number of subsets evaluated: 392
        Merit of best subset found:    0.186

Attribute Subset Evaluator (supervised, Class (nominal): 32 VSA):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,5,13,15,16,17,21,22,24,26,27,28,30,31 : 14
                     Age
                     Country
                     SS
                     Amphet
                     Amyl
                     Benzos
                     Coke
                     Crack
                     heroin
                     Legalh
                     LSD
                     Meth
                     Nicotine
                     Semer
```

Status: OK

Figure 2: Features selection process



Filter

Filter: None

Current relation

Relation: drug_consumption_data-weka.filters.unsupervised.attribute.Remove-R1,3-4,... Attributes: 14
Instances: 1885 Sum of weights: 1885

Attributes

No.		Name
1	□	Age
2	□	Country
3	□	SS
4	□	Amphet
5	□	Amyl
6	□	Benzos
7	□	Coke
8	□	Crack
9	□	heroin
10	□	Legalh
11	□	LSD
12	□	Meth
13	□	Nicotine
14	□	Semer

Selected attribute

Name: Age Type: Numeric
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

Statistic	Value
Minimum	-0.952
Maximum	2.592
Mean	0.035
StdDev	0.878

Class: Semer (Nom)

Status

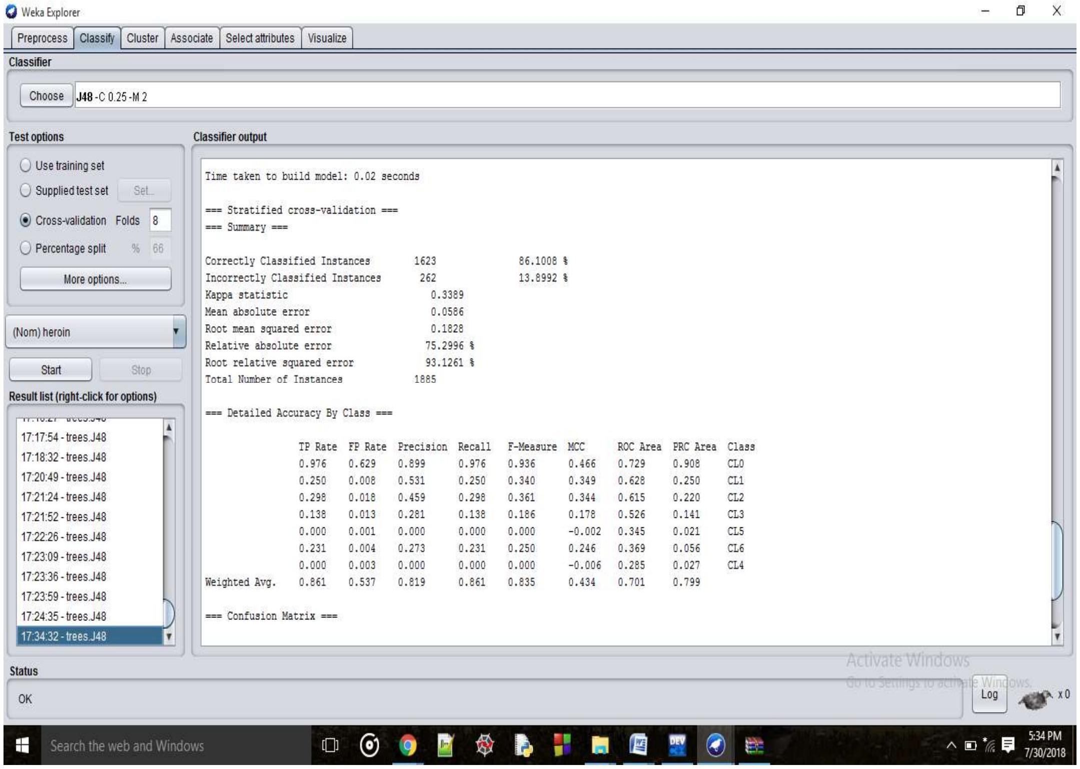
OK

- Figure 3: Remaining Attributes after Feature selection

SELECTED ATTRIBUTES TAKEN, CLASSIFICATION APPLIED, CROSS VALIDATION USED, ABOUT HEROIN ATTRIBUTE

No. of Cross Validation Folds	Accuracy obtained	Error observed
10	85.7825	14.2175
9	85.4642	14.5358
8	86.1008	13.8992
7	85.2520	14.7480
6	85.7294	14.2706
5	85.9947	14.0053
4	85.4111	14.5889
3	85.8886	14.1114
2	85.6764	14.3236
11	86.1008	13.8992
12	85.6233	14.3767
13	85.7294	14.2706
14	85.7294	14.2706
15	85.5703	14.4297

16	85.8355	14.1645
17	85.7294	14.2706
18	85.3581	14.6419
19	85.5703	14.4297
20	85.6764	14.3236



J48 -C 0.25 -M 2

● Cross-validation Folds 8

(Nom) heroin

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1623               86.1008 %
Incorrectly Classified Instances       262               13.8992 %
Kappa statistic                          0.3389
Mean absolute error                      0.0586
Root mean squared error                  0.1828
Relative absolute error                 75.2996 %
Root relative squared error             93.1261 %
Total Number of Instances             1885

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.976    0.629    0.899      0.976    0.936      0.466    0.729     0.908     CL0
                 0.250    0.008    0.531      0.250    0.340      0.349    0.628     0.250     CL1
                 0.298    0.018    0.459      0.298    0.361      0.344    0.615     0.220     CL2
                 0.138    0.013    0.281      0.138    0.186      0.178    0.526     0.141     CL3
                 0.000    0.001    0.000      0.000    0.000      -0.002   0.345     0.021     CL5
                 0.231    0.004    0.273      0.231    0.250      0.246    0.369     0.056     CL6
                 0.000    0.003    0.000      0.000    0.000      -0.006   0.285     0.027     CL4
Weighted Avg.    0.861    0.537    0.819      0.861    0.835      0.434    0.701     0.799

=== Confusion Matrix ===
```

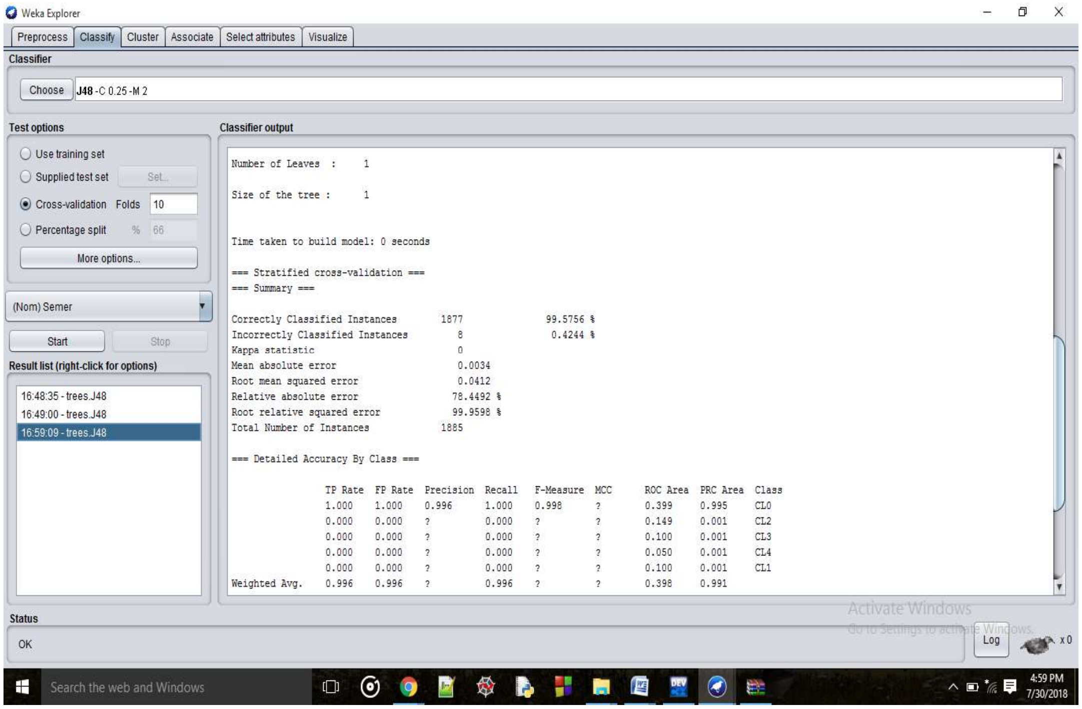
- **Discussion for heroin attribute**

As per the above table , it is clear that the maximum accuracy is 86.0001% when we take heroin as root attribute and cross-validation folds are 8 .
This highlighted the importance of Feature Selection in classification. Attribute selection is done in the project using CfsSubsetEval evaluator.

SELECTED ATTRIBUTES TAKEN, CLASSIFICATION APPLIED, CROSS VALIDATION USED, ABOUT SEMER ATTRIBUTE

No. of Cross Validation Folds	Accuracy obtained	Error observed
10	99.5756	0.4244
9	99.5756	0.4244
8	99.5756	0.4244
7	99.5756	0.4244
6	99.5756	0.4244
5	99.5756	0.4244
4	99.5756	0.4244
3	99.5756	0.4244
2	99.5756	0.4244
11	99.5756	0.4244
12	99.5756	0.4244
13	99.5756	0.4244
14	99.5756	0.4244
15	99.5756	0.4244

16	99.5756	0.4244
17	99.5756	0.4244
18	99.5756	0.4244
19	99.5756	0.4244
20	99.5756	0.4244



Classifier: J48 -C 0.25 -M 2

Test options: ● Cross-validation Folds 10

(Nom) Semer

Result list (right-click for options)

16:48:35 - trees.J48
16:49:00 - trees.J48
16:59:09 - trees.J48

Classifier output

```
Number of Leaves  :     1

Size of the tree :      1


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1877               99.5756 %
Incorrectly Classified Instances         8                0.4244 %
Kappa statistic                          0
Mean absolute error                      0.0034
Root mean squared error                  0.0412
Relative absolute error                 78.4492 %
Root relative squared error             99.9598 %
Total Number of Instances             1885

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.996      1.000    0.998      ?        0.399     0.995     CL0
                 0.000    0.000    ?          0.000    ?          ?        0.149     0.001     CL2
                 0.000    0.000    ?          0.000    ?          ?        0.100     0.001     CL3
                 0.000    0.000    ?          0.000    ?          ?        0.050     0.001     CL4
                 0.000    0.000    ?          0.000    ?          ?        0.100     0.001     CL1
Weighted Avg.    0.996    0.996    ?          0.996    ?          ?        0.398     0.991
```

Status

OK

- **Discussion for semer attribute**

  As per the above table , it is clear that the accuracy is 99.5756% when we take semer as root attribute .
  This highlighted the importance of Feature Selection in classification. Attribute selection is done in the project using CfsSubsetEval evaluator.

CONCLUSION

In this project we learn how to select important features of a dataset and reduce the computational work.
We also learn how attributes play important role for accuracy.
Example : When we take heroin as root attribute we got maximum 86.0001% accuracy but in case of semer we got 99.5756% accuracy.
Above example shows how attribute selection is important in weka.