UNIVERSITY OF GHANA

College of Basic and Applied Sciences
School of Engineering Sciences
Department of Computer Engineering
First Semester 2022 Academic Year
CPEN 405: Artificial Intelligence Course Project

PART 1: Machine Learning

GROUP 5

Emmanuel Amoah Kwame – 10464434
Amanfo Oforiwaa Abena – 10714326
Dumenu Ernest Mawusu – 10726400
Kayang Edwin Pelpuo: 10728521
El-Karece Amoakoa Asiedu – 10735358

INTRODUCTION

Artificial intelligence is all about equipping machines with human-like intelligence. Artificial intelligence is now in vogue in the technology world. From smart devices through robotics and even to personalized disease diagnosis and drug design. Because of the huge potential benefits, it has received a lot of attention and its growth is sponsored by several entities. PART 1 is mainly about machine learning using WEKA. The project's goal is to gain practical experience with machine learning methods by using software such as WEKA to solve real-world data mining problems, as well as to gain a better understanding of some of the algorithmic issues that arise when designing and applying various machine learning algorithms. For our experiments, we used 10-fold cross validation with 5 classification schemes for Soybean Disease Diagnosis. After which, we used the WEKA Experiment Paired Corrected T-Tester to compare the classification schemes for their performance. The dataset used was Soybean, with its training and test database combined into a single file, from the UCI Machine Learning Repository, and 5 classification schemes were applied to it for evaluation. These data sets were selected because they are large enough to allow moderate size train and validation sets, and still have data left for large final test sets. It proved to be the best fit for the constraints provided in the instructions given.

PROBLEM FORMULATION

The following are its characteristics/ the rationale behind the datasets's selection:
There are 683 Instances, 19 classes (different diseases in soybean plant), however only the first 15 have been utilized previously. Because there are so few cases, the consensus seems to be that the last four classes are invalidated by the evidence. After performing Attribute Selection utilizing Attribute Ranking Search Method on it, there are 36 attributes, but 35 categorical attributes, some nominal and others sorted.

Data Set Characteristics	Multivariate	Number of Instances	683	Area	Life
Attribute Characteristics	Categorical	Number of Attributes	35	Date Donated	1988/07/11
Associated Tasks	Classification	Missing Values?	Yes	No. of Web Hits	153160

Source/Origin:

R.S. Michalski and R.L. Chilausky

"Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis", International Journal of Policy Analysis and Information Systems, Vol. 4, No. 2, 1980.

Donor: Ming Tan & Jeff Schlimmer (Jeff.Schlimmer%cs.cmu.edu), 11 July 1988

Link: https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)

Attribute Information:

Classes: diaporthe-stem-canker, charcoal-rot, rhizoctonia-root-rot, phytophthora-rot, brown-stem-rot, powdery-mildew, downy-mildew, brown-spot, bacterial-blight, bacterial-pustule, purple-seed-stain, anthracnose, phyllosticta-leaf-spot, alternarialeaf-spot, frog-eye-leaf-spot, diaporthe-pod-&-stem-blight, cyst-nematode, 2-4-d-injury, herbicide-injury.

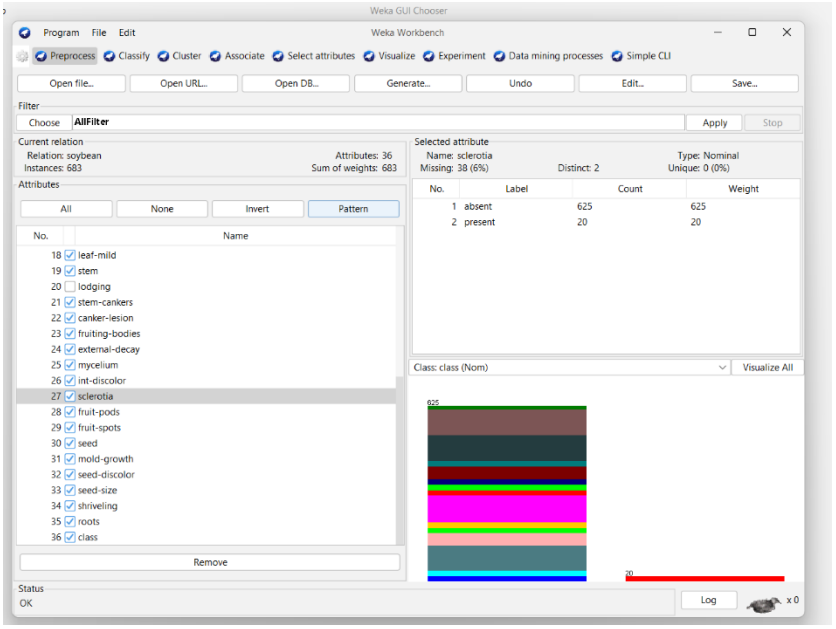
Attributes:

1. date: april, may, june, july, august, september, october,?.
2. plant-stand: normal, lt-normal,?.
3. precip: lt-norm,norm,gt-norm,?.
4. temp: lt-norm, norm, gt-norm,?.
5. hail: yes, no,?.
6. crop-hist: diff-lst-year, same-lst-yr, same-lst-two-yrs, same-lst-sev-yrs,?.
7. area-damaged: scattered,low-areas,upper-areas,whole-field,?.
8. severity: minor,pot-severe,severe,?.
9. seed-tmt: none,fungicide,other,?.
10. germination: 90-100%,80-89%,lt-80%,?.
11. plant-growth: norm,abnorm,?.
12. leaves: norm,abnorm.

13. leafspots-halo: absent,yellow-halos,no-yellow-halos,?.
14. leafspots-marg: w-s-marg,no-w-s-marg,dna,?.
15. leafspot-size: lt-1/8,gt-1/8,dna,?.
16. leaf-shread: absent,present,?.
17. leaf-malf: absent,present,?.
18. leaf-mild: absent,upper-surf,lower-surf,?.
19. stem: norm,abnorm,?.
20. lodging: yes,no,?.
21. stem-cankers: absent,below-soil,above-soil,above-sec-nde,?.
22. canker-lesion: dna,brown,dk-brown-blk,tan,?.
23. fruiting-bodies: absent,present,?.
24. external decay: absent,firm-and-dry,watery,?.
25. mycelium: absent,present,?.
26. int-discolor: none,brown,black,?.
27. sclerotia: absent,present,?.
28. fruit-pods: norm,diseased,few-present,dna,?.
29. fruit spots: absent,colored,brown-w/blk-specks,distort,dna,?.
30. seed: norm,abnorm,?.
31. mold-growth: absent,present,?.
32. seed-discolor: absent,present,?.
33. seed-size: norm,lt-norm,?.
34. shriveling: absent,present,?.
35. roots: norm,rotted,galls-cysts,?.

SOLUTION APPROACH AND ALGORITHMS

We used 10-fold cross validation with 5 classification schemes for Soybean Disease Diagnosis. After which, we used the WEKA Experiment Paired Corrected T-Tester to compare 3 classification schemes for their performance. The raw data was pre-processed in various ways. Firstly, the ordinal inputs were normalized to have zero mean and unit standard deviation on the training data. Part of the inputs are categorical and these are mapped to a 1-of-c coding, thus increasing the number of attributes from 35 to 36. The worth of "dna" stands for "does not apply." The values for characteristics are encoded numerically, with "0" being the first value, "1" being the second, and so on. The value of an unknown value is encoded as "?" for each of the five algorithms.



Data Preprocessing

env(precipitation) = [l,n]

The following results were obtained (using the ranked rules and [prop] strategy):

% identification = 50
Indecision Ratio = 1.9

The data was modified so that:

env(precipitation) = g
env(precipitation) = n

and the rule condition used was:

env(precipitation) = [g,n]

The results obtained (using the same rules and strategy) were:

% identification = 50
Indecision Ratio = 2.3
Specificity Index = 7.5

The changes in the Indecision Ratio were due to three extra false positive identifications of cases of phyllosticta leaf spot as brown spot and one false positive identification of phyllosticta leaf spot as frog eye leaf spot. The changes in the Specificity Index were due to eight cases of brown spot, and three cases of alternaria leaf spot being incorrectly identified as phyllosticta leaf spot.

ALGORITHMS

Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve high accuracy levels.

Bagging, also known as bootstrap aggregation or Random Forest, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement meaning that the individual data points can be chosen more than once.

K-means clustering algorithm computes the centroids and iterates until it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Stacked Generalization, or stacking for short, is an ensemble machine learning algorithm. Stacking involves using a machine learning model to learn how to best combine the predictions from contributing ensemble members.

RESULTS AND DISCUSSION

Ranked Attributes:
Evaluator: weka.attributeSelection.InfoGainAttributeEval
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation: soybean, Instances: 683, Attributes: 36
Evaluation mode: Evaluate on all training data
=== Attribute Selection on all input data ===

Search Method: Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 36 class): Information Gain Ranking Filter

Ranked attributes:
1.1517 22 canker-lesion
1.0129 15 leafspot-size
0.9852 29 fruit-spots
0.8684 13 leafspots-halo
0.8535 21 stem-cankers
0.8504 14 leafspots-marg
0.8437 28 fruit-pods
0.6918 19 stem
0.6715 1 date
0.6265 11 plant-growth
0.5853 3 precip
0.5392 35 roots
0.5245 26 int-discolor
0.4829 24 external-decay
0.4808 7 area-damaged
0.4241 4 temp
0.4133 30 seed
0.3614 18 leaf-mild
0.3568 12 leaves

0.3517 23 fruiting-bodies
0.3432 31 mold-growth
0.3106 8 severity
0.2981 33 seed-size
0.2862 2 plant-stand
0.2688 32 seed-discolor
0.2629 16 leaf-shread
0.2465 17 leaf-malf
0.2173 34 shriveling
0.1883 27 sclerotia
0.0987 20 lodging
0.0787 6 crop-hist
0.0784 5 hail
0.0742 9 seed-tmt
0.0554 10 germination
0.0461 25 mycelium

Selected attributes:
22,15,29,13,21,14,28,19,1,11,3,35,26,24,7,4,30,18,12,23,31,8,33,2,32,16,17,34,27,20,6,5,9,10,25 : 35

Ranked Attributes 2(Best First):
Evaluator: weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search: weka.attributeSelection.BestFirst -D 1 -N 5
Relation: soybean, Instances: 683, Attributes: 36
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method: Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 552
Merit of best subset found: 0.702
Attribute Subset Evaluator (supervised, Class (nominal): 36 class): CFS Subset Evaluator, Including locally predictive attributes
Selected attributes: 1,3,4,5,7,8,9,10,11,12,13,15,17,18,19,22,23,24,26,28,30,35 : 22

- **date**
- **precip**
- **temp**
- **hail**
- **area-damaged**
- **severity**
- **seed-tmt**
- **germination**
- **plant-growth**
- **leaves**
- **leafspots-halo**
- **leafspot-size**
- **leaf-malf**
- **leaf-mild**

- **stem**
- **canker-lesion**
- **fruiting-bodies**
- **external-decay**
- **int-discolor**
- **fruit-pods**
- **seed**
- **roots**

CLASSIFICATION RESULTS

Double Click to view a detailed report of results. This includes: Scheme, Test mode, Classifier model, Predictions on training set, Time taken to build model, Actual prediction, Error prediction, Summary on Correctly Classified Instances, Incorrectly Classified Instances, Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, and Detailed Accuracy By Class(TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area, PRC Area, and Weighted Avg.), as well as the Confusion Matrix.

Naive Bayes:


NBtest.txt


NaiveBtrain.txt

Bagging:


BaggingTr.txt


bagginTe.txt

K-means:


Cluster,KMeans.txt

Random forest:


RandomForest.txt


RFtest.txt

Stacked Generalization:


StackingTrain.txt


StackingTest.txt

Plot Of Arrtributes:

Below is the visualization of all the attributes

Plot.arff

Comparing 3 Classification Schemes

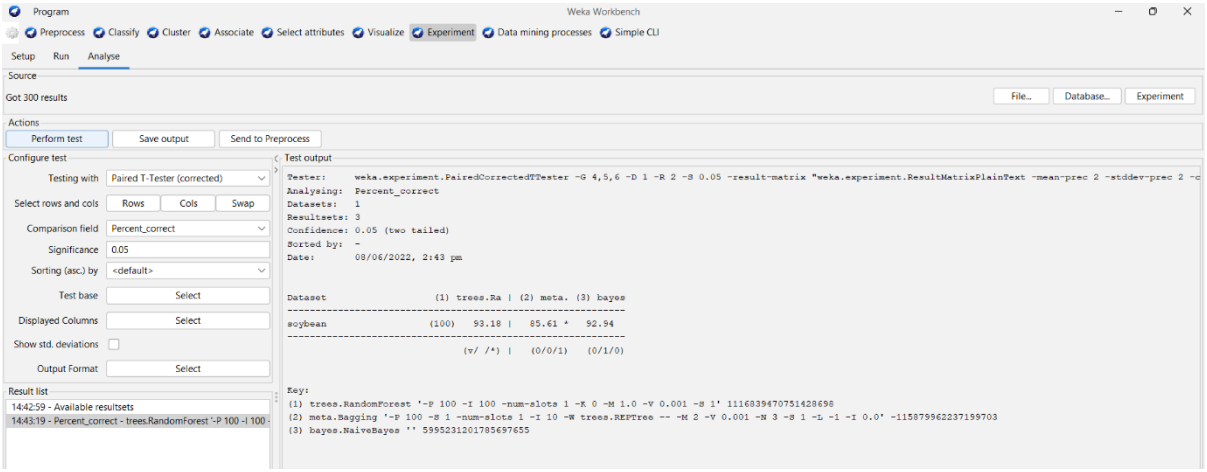
Using the WEKA Experiment, Paired Corrected T-Tester to compare the classification schemes for their performance, 10-fold cross validation was run on Random Forest, Bagging and Naive Bayes for 25 times. The results below show that the Random Forest(%93.18) and Naive Bayes(%92.94) schemes are very competitive. However, Bagging(%85.61) was the least performing. The files below hold detailed reports of how they performed, why they had such accuracies and why Bagging was the least performed.

comparison experiment

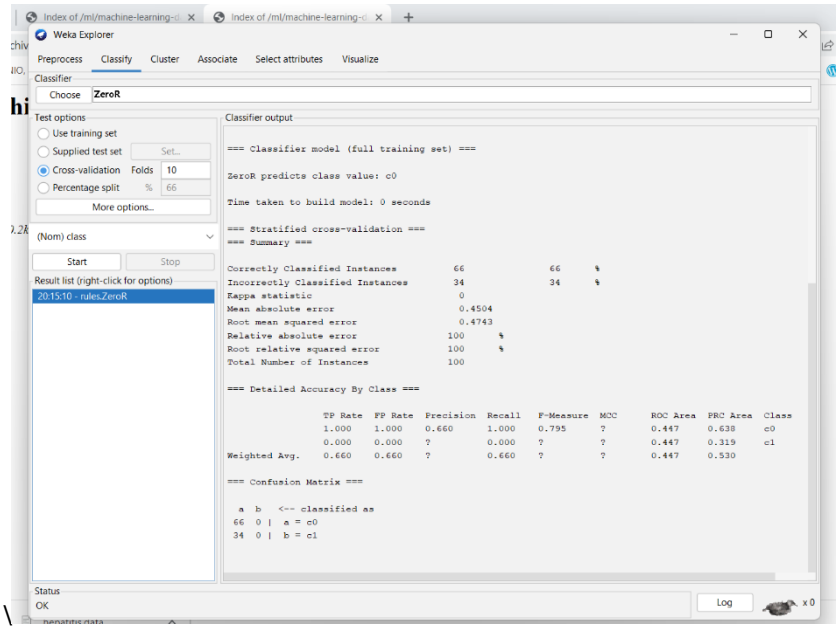
Comparison.exp

soybeanExp.arff

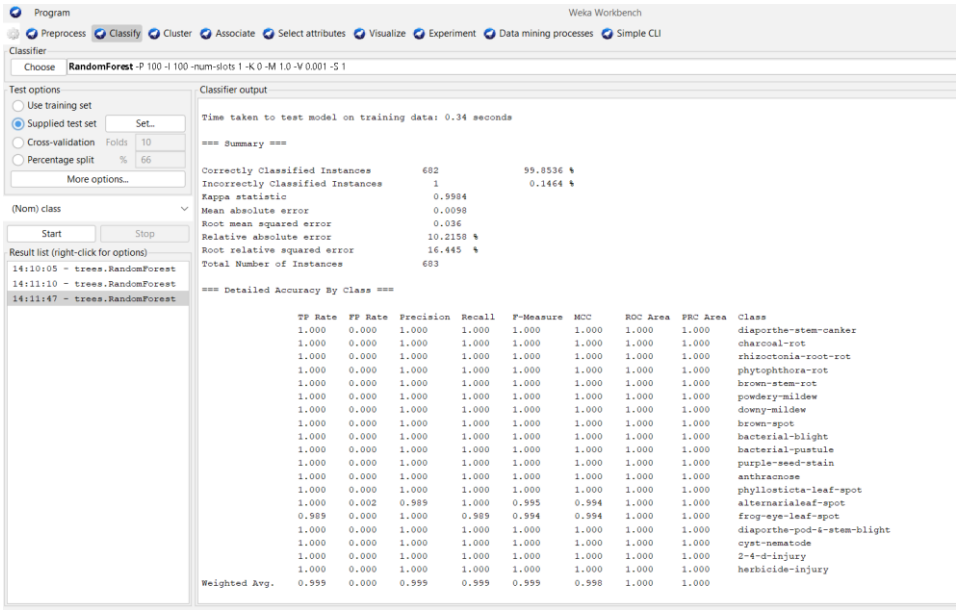
APPENDICES (Program screenshots)

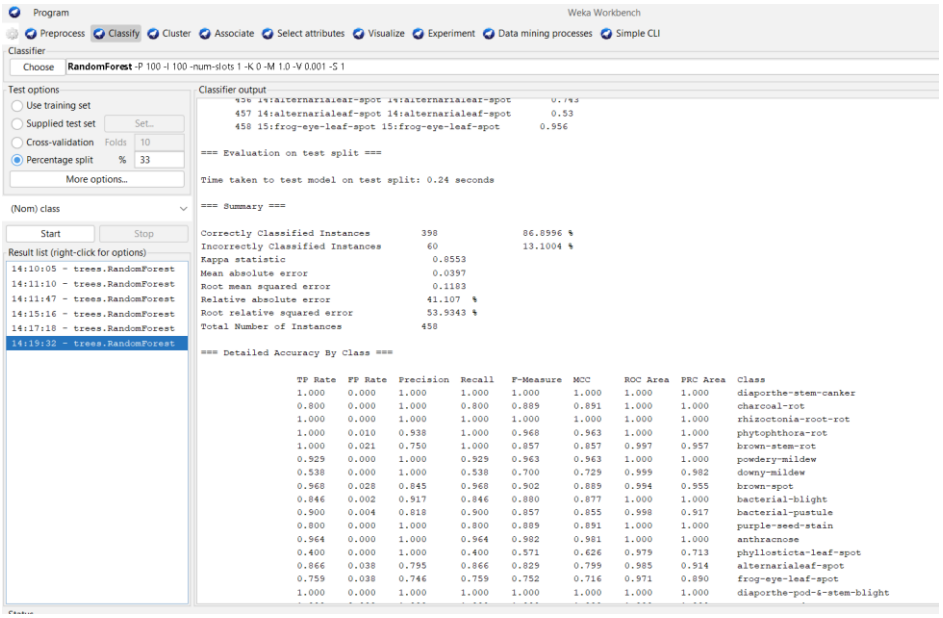


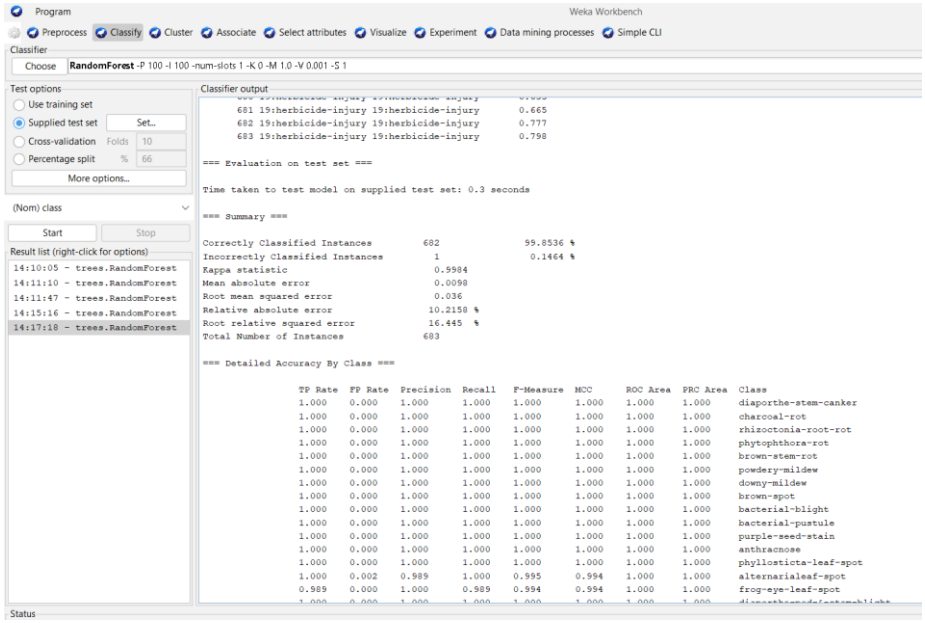
Comparing Schemes



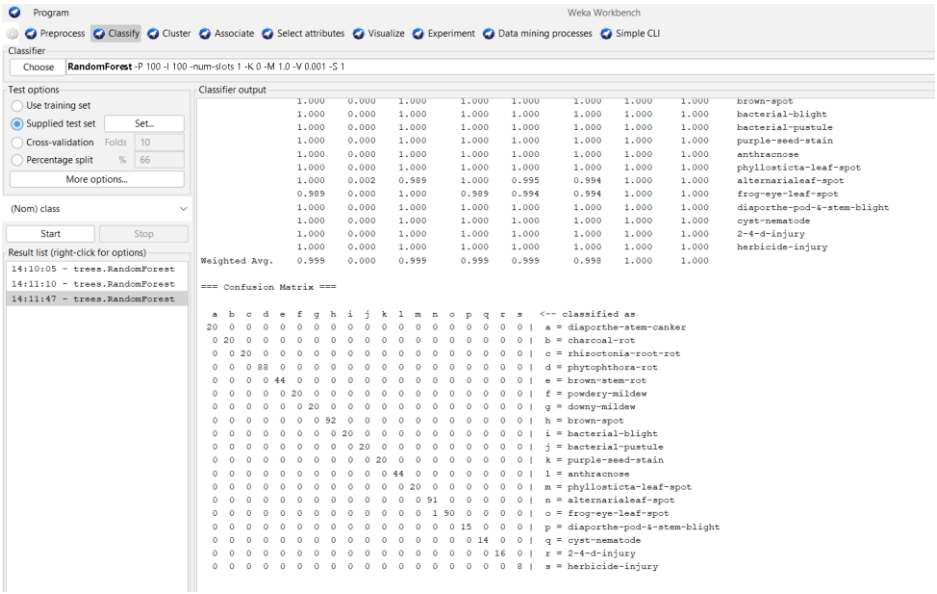
Cross Validation with ZeroR Rules



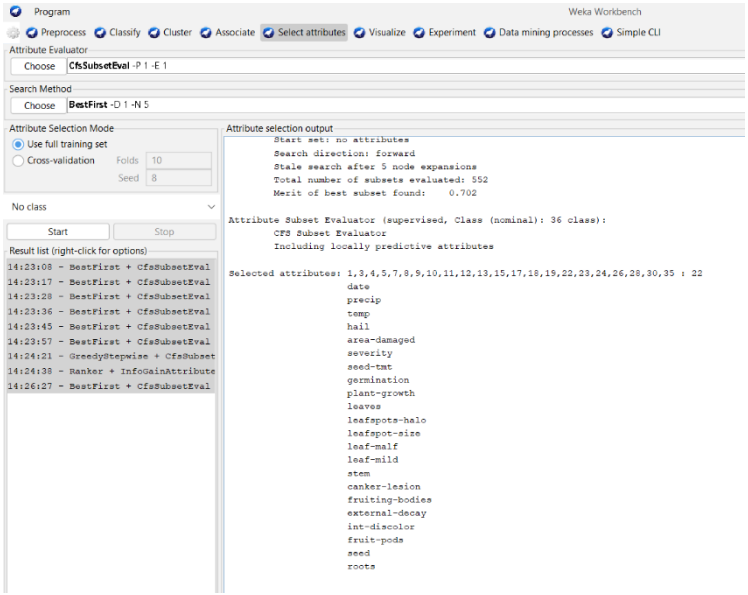




Random Forest Classifier(Best Performing)

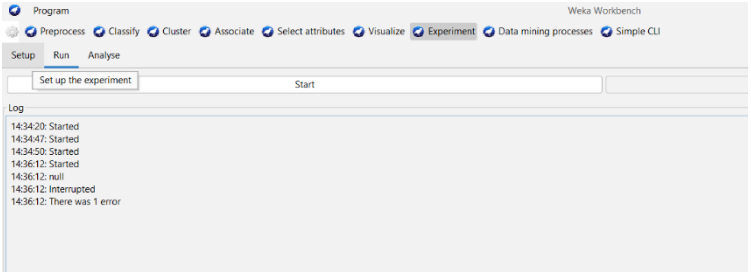


Confusion Matrix, Random Forest Classifier(Best Performing)

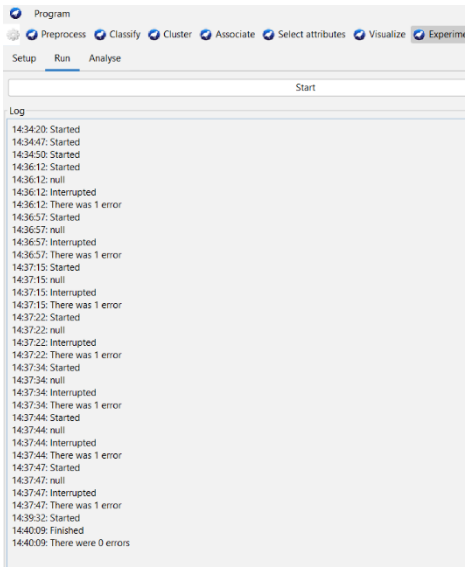


Best First + CFS Subset Evaluator

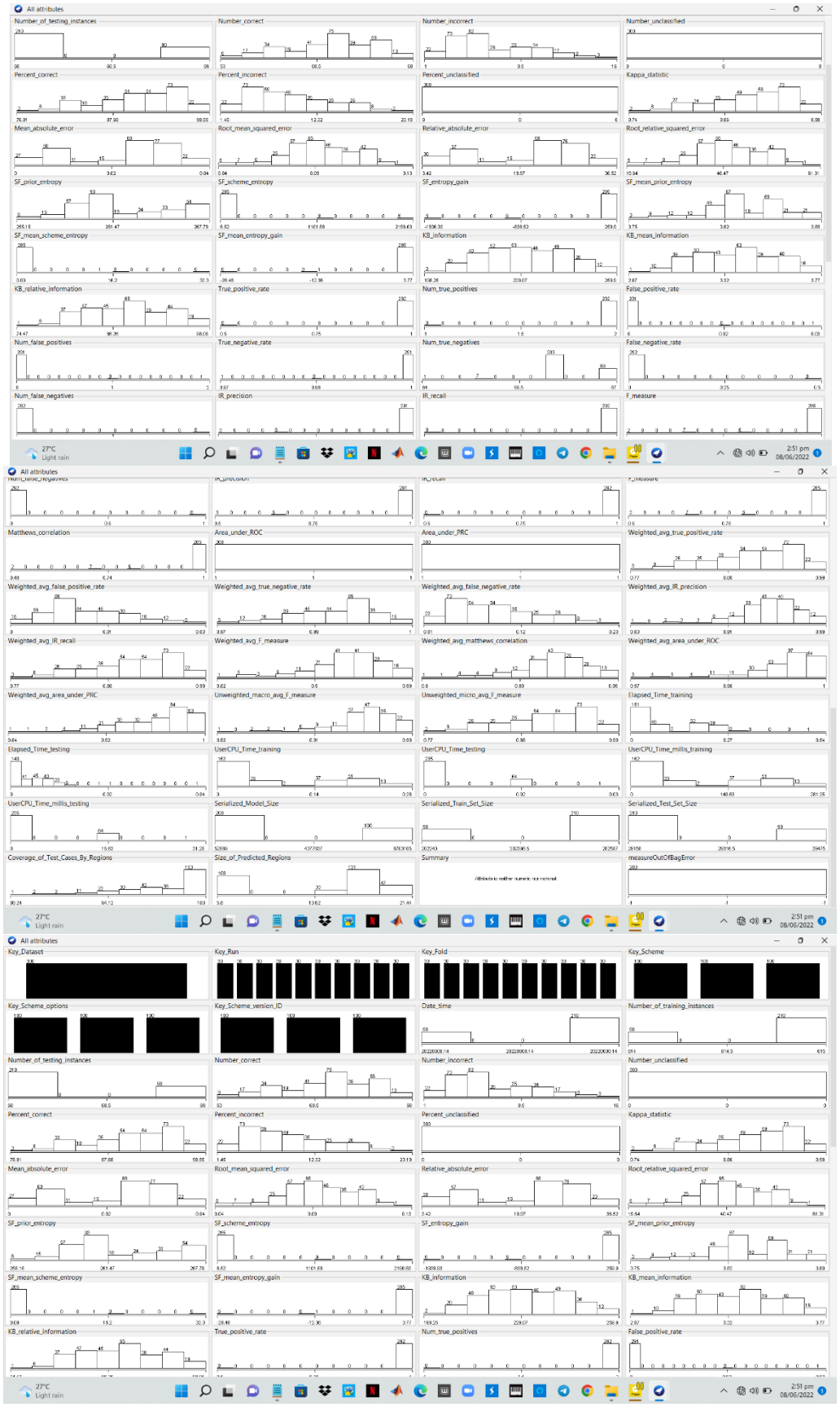
Plots from Raw Data



Error obtained from using raw data



Feedback after normalizing data



Plots from Normalized Data

CONCLUSION

The WEKA explorer is used for a variety of tasks, beginning with preprocessing. Preprocessing takes an.arff file as input, processes it, and outputs a file that may be utilized by other computer applications. The preprocessing output in WEKA provides the properties available in the dataset, which can then be used for statistical analysis and comparison with class labels.

WEKA also has a number of decision tree classification techniques. J48 is a well-known classification algorithm that generates a decision tree. The user can visualize the decision tree using the Classify tab. If the decision tree has become overly filled, tree pruning can be performed from the Preprocess tab by eliminating non-essential attributes and restarting the categorization process.

The results show that the RandomForest and NaiveBayes schemes are very competitive. However, Bagging was the least performing because it had most of the networks for the data set containing a maximum of 20 hidden units(average number of hidden units was 19.44). Hence, it might be possible to improve the overall predictive accuracy of these networks.

REFERENCES

[1] R.S. Michalski and R.L. Chilausky "Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis", International Journal of Policy Analysis and Information Systems, Vol. 4, No. 2, 1980.

[2] Tan, M., & Eshelman, L. (1988). Using weighted networks to represent classification knowledge in noisy domains. Proceedings of the Fifth International Conference on Machine Learning (pp. 121-134). Ann Arbor, Michigan: Morgan Kaufmann.

[Web Link (http://rexa.info/paper/9caf1d9fd8292532ba2a5348c6f381ca5421b59e)]

[3] Fisher,D.H. & Schlimmer,J.C. (1988). Concept Simplification and Predictive Accuracy. Proceedings of the Fifth International Conference on Machine Learning (pp. 22-28). Ann Arbor, Michigan: Morgan Kaufmann.

[Web Link (http://rexa.info/paper/71b78822eab4b70819ca479b23c5a84a70185605)]