Jay Patel
CS422 – Assignment 2 Report
09 – 16 – 2018

ASSIGNMENT 2 REPORT

The given assignment require Weka Software in order to analyze the data provided along with the Weka Packages.

In this report, we will be analyzing data from 3 different dataset; Iris, Vote, and Diabetes.

Part I: Data Analysis

1.1 Attributes and Types

	Number of Attributes	Type of Attribute	Total No. of Instances
Iris	5	Numeric	150
Vote	17	Nominal	435
Diabetes	9	Numeric	768

1.2 Detail Report for Iris

- The Iris dataset shows the classification of the Iris flower based on certain attributes such as listed below:
  - Class – There are 150 instances of the data and its evenly distributed among 3 different types of Iris flower, which are categorized into “classes”, which basically are like identifiers that will help categorized test data into these classes based on their attributes.
  - Sepal Length
    - classified based on the length of the Sepal.
    - Min size: 2 | Max size: 4.4 | Mean: 3.054 | Std. Dev: 0.434
    - Looking at the graph, we can notice that most of the data points overlap in the range [4.8, 6.1] approx. This will become rather difficult for the classifier to sort the instance into a category based on this attribute.
    - We can also see that the Iris-setosa class flowers have smaller Sepal length, while the other two ; Iris-versicolor and Iris-virginica have length spread out.
  - Sepal Width
    - classified based on the width of the Sepal.
    - Min size: 4.3 | Max size: 7.9 | Mean: 5.843 | Std. Dev: 0.828
    - Similar to Length, the Width of the Sepal is one of the attributes. Attribute graph shows that most of the Virginica are almost same width while others are spread throughout the range. Classifier requires the sepal width and can sort the instance accordingly.

- Petal Length
    - In order to classify it as a specific type of Iris flower, we also need the size of Petal as one of the attributes.
    - Min size: 1 | Max size: 6.9 | Mean: 3.759 | Std. Dev: 1.764
    - The Petal attributes graphs are visually distinct. The key point in this graph is that Iris-setosa's entire 50 instances have shorter length, which reduces classifier's error rate.
- Petal Width
    - Petal Width is also key feature in order to classify iris.
    - Min size: 0.1 | Max size: 2.5 | Mean: 1.199 | Std. Dev: 0.763
    - As seen in the petal length, the width of Setosa is smaller as well, which becomes unique to this class, and increases classifier success %age.

- Weka Workbench provides great visualization tool which can be used to analyse data. Using the plot matrix, major overlap of instances can be notices, are between Versicolor, and Virginica. Setosa, on the other hand is quite distinct. This will help the classifier sort Setosa with 100% accuracy, while get some set backs, on the other two.

Part II Classifiers:

Classifier (Tree)	Parameters
Decision Stump	Batch Size (Can change the size of dataset used)
J48	- Confidence factor: Builds complete tree and the work back from leaves. Smaller number leads more pruning. - minNumObj: Stops splitting if the nodes get too little. Set a limit of instances per leaf.
Random Forest	- Number of Trees: Specify how many trees to use in randomization - Max depth of trees: alters the speed of the classifier. - Random forest picks K random options rather than best ones. - Number of Iteration : basically points to number of trees generated

Jay Patel
CS422 – Assignment 2 Report
09 – 16 – 2018

1. Decision Stump:
    a. Iris:
        i. The classes in the Iris data set are equally divided; 50 each, making 150 total instances.
        ii. It took 0.01 secs to build the model, with accuracy of 66.67%. There were no Iris-virginica detected by the Stump. They were classified as Versicolor. This could have been because the data points for these two have major overlap (Observed at Confusion Matrix).
    b. Vote:
        i. A little different than Iris, Vote Data set has more number of Democrats than Republicans. Also, they are nominal in time so there is less variation, although we can notice over lap.
        ii. Decision Stump provided with better performance than with Iris, with 95.6% success rate.
        iii. Classification done based on physician-fee-freeze. Based on the graph. It seems that they data is well sorted in yes and no with a minute overlap. This might have caused the ease in classification.
    c. Diabetes
        i. Plas attribute was used in order to classify the model. We can see pattern over here, is that not all the attributes in the model. The algorithm only uses features that are appropriate and predictable.
        ii. 71% success rate, Stump classified 552 out of 768 correctly.
        iii. Diabetes Dataset is different from the other two because there are only 2 classes while there are many features / attributes of the instances that depicts the class.
        iv. Plas ranges from 1 to 199 and the Decision Tree uses 127.7 as cutting point for classification.

2. J48
    a. Iris:
        i. One of the best accuracies provided by J48. Only 6 incorrect classification over 150 instances. (96% success rate)
        ii. Most of the classification was done based on the length and width of Petal.
        iii. Decreasing the minNumObj to 1 decreases the success rate to 94%. Changing the confidence number doesn't bring any significant change in the classification.
    b. Vote:
        i. Using full training set, J48 Algorithm provides with better success rate of 96.3%. The Classifier created 6 leaves with size of tree – 11.
        ii. 5 out of 17 attributes used for pruning.
        iii. Manipulating the confidence number changes the time taken to build the model. Changing confidence number to 0.7 increases time taken to 0.13 secs. However, it doesn't change any performance rate.
    c. Diabetes:
        i. Very Low success rate compared to other two data set, with 73.82%.

ii. It took 0.03 secs to build the model which is slower as well.
iii. It can be also noticed that the size of the tree in this model is 39 with 20 leaves.
iv. Increasing the minNumObj parameter to 4 and decreasing confidence to 0.1, it gives faster result, and little growth in success rate (74.34%)

3. Random Forest**
   a. Iris:
      i. Given smaller number of instances, and compared to other trees, Random forest took more time in building the model; 0.03 secs.
      ii. With 95% correctly classified instances with default parameters, Model had high success.
      iii. Using the varied number of trees, below is the table for # of trees, Performance, and time.

Num. Trees	Performance(%)	Time(Secs)
10	95.33	0.01
20	95.33	0.02
50	94.66	0.03
100	95.33	0.05
200	95.33	0.04
500	95.33	0.07
1000	95.33	0.16

   b. Vote:
      i. With 418 instances, and 100 iteration, it took 0.15 secs to build the model. It provided with 96% success rate.
      ii. It is difficult to tell which attributes were chosen at excess, since selection changes at every iteration.
      iii. Using the varied number of trees, below is the table for # of trees, Performance, and time.

Num. Trees	Performance(%)	Time(Secs)
10	96.32	0.02
20	96.32	0.04
50	96.55	0.07
100	96.02	0.15
200	96.55	0.15
500	96.32	0.29
1000	96.55	0.62

   c. Diabetes:
      i. Giving higher success rate of 75.09%, higher than other two algorithm model for Diabetes.
      ii. Random forest at default parameters build in 0.4 secs which would be due to increased number of instances.

Jay Patel
CS422 – Assignment 2 Report
09 – 16 – 2018

iii. Using the varied number of trees, below is the table for # of trees, Performance, and time.

Num. Trees	Performance(%)	Time(Secs)
10	74.39	0.07
20	74.86	0.09
50	75.78	0.17
100	75.78	0.22
200	75.65	0.43
500	75.78	1.19
1000	75.39	1.95

Part III Feature Selection:

1. Iris:
   Iris is one of the easier ones to predict to which attributes are useful or not. The only way three classes mostly differentiate was based on the petals – length and width.
   a. Correlation Attribute Evaluation:
      ColAttEval ranks Petal Width at 1
   b. Information Gain Evaluation:
      InfoGainEval gives priority to Petal Length over width. However, as expected, Sepal sizes are not important to be included. But Removing these attributes from the dataset still gives the same performance with slight increase in time.
2. Vote:
   Vote contains a lot of features, and it has higher number of instances. Looking at the data, we noticed that the physician-fee attributes is an important one, since our J48 starts its pruning using that attribute.
   a. Correlation Attribute Evaluation:
      i. The attributes are ranked as predicted. Physicians-fee , adoption, el-salvador, education received high merit and rank.
   b. Information Gain Evaluation:
      i. We got the same ranking result as the correlation algorithm, atleast the top attributes.
   c. When attempted to remove unnecessary attribute, it changed the success with a little bit, however it changed the process time a lot.

3. Diabetes:
   a. Correlation Attribute Evaluation:
      i. The algorithm picked mainly the attributes that were bell curved graph, also they over lapping. Pled, Mass, and Age are the ones I predicted during data visualization process.

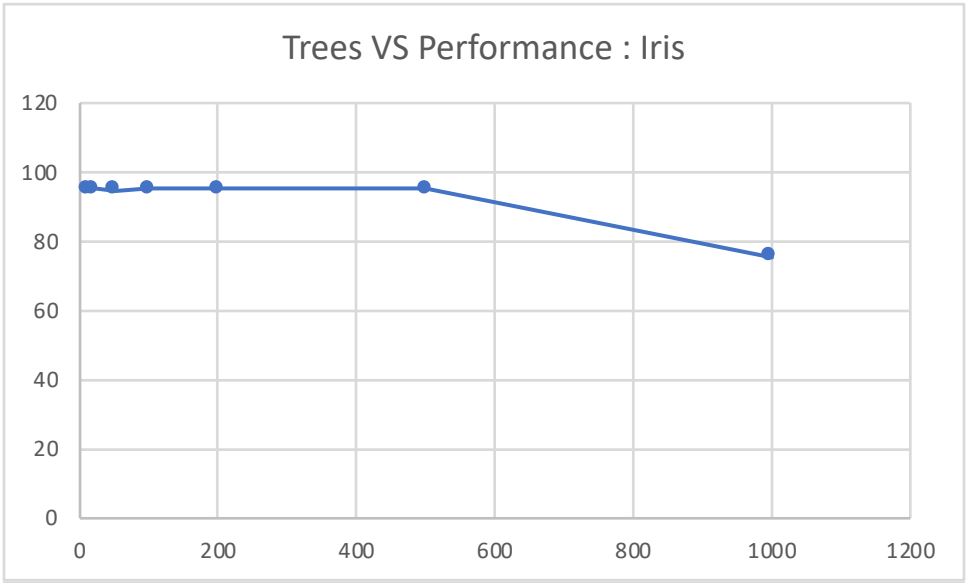
b. Information Gain Evaluation:
    i. Similar to previous Algorithm, the InfoGainEval works on ranking algorithm, and outputs similar pattern of attribute selection.
c. Removing the features doesn't provide with any significant growth in the success rate.

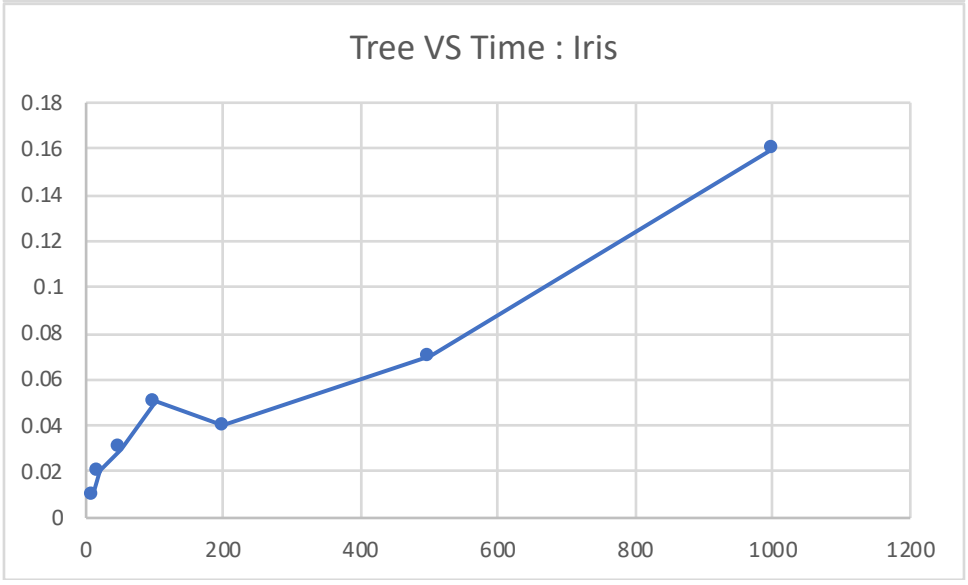
Part IV Noise and Missing Values:

- Introduced Missing values using the question marks. Since the size of the data set was 150, we added 7, 15 75 missing values in order and saw the change in performance.
- Used Random Forest as our training model, while increasing the size of the missing values, we could see that the training time, increased signifanctly as we put more and more missing values.
- At 15% the time increased from 0.02 to 0.04, and the performance decreased to 94.66 %.
- After the missing value experiment, introducing some mismatched values, and added noise.
- Introducing noise, significantly dropped the success rate to 82% using the same random forest model at default parameter. Even though the time taken by the classifier was same, the performance significantly decreased,
- Furthermore, after removing the missing values, tested the selected attribute feature, and saw that there has been some changes in the rank and merit as well.

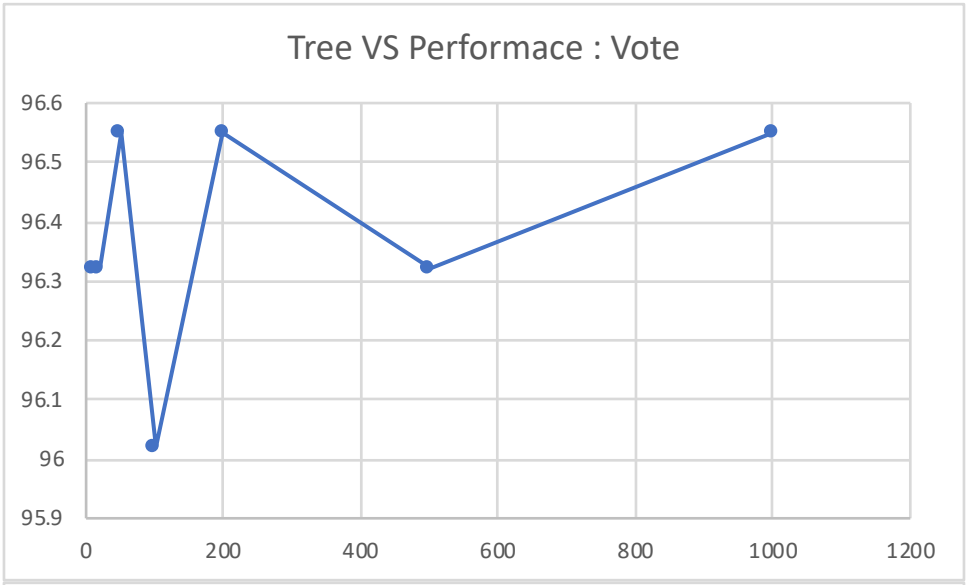
** Provided are the graphs for the Varied Forest Size

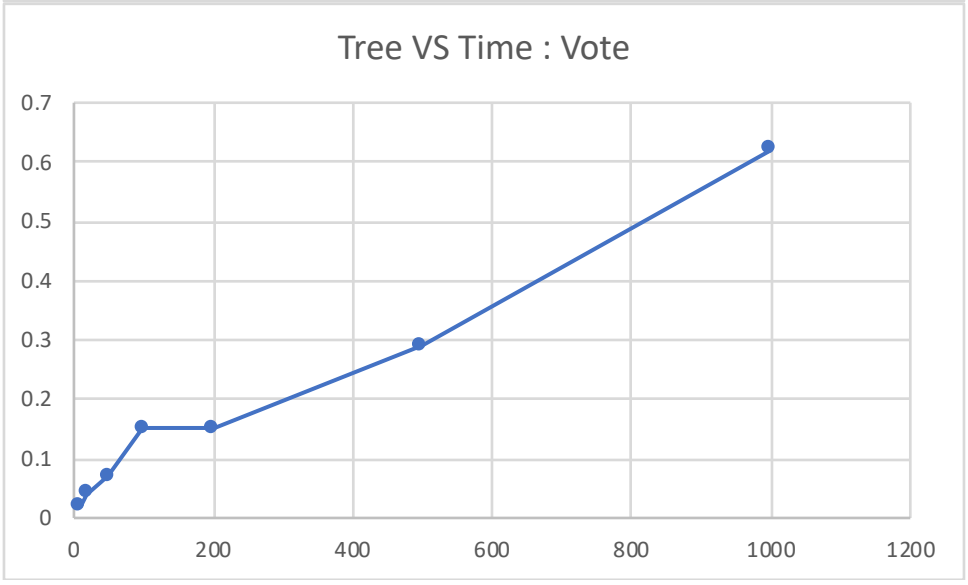
Jay Patel
CS422 – Assignment 2 Report
09 – 16 – 2018





Jay Patel
CS422 – Assignment 2 Report
09 – 16 – 2018





Jay Patel
CS422 – Assignment 2 Report
09 – 16 – 2018