**UNIT-2 (Data Mining with WEKA)**

**Q1. Use the following learning schemes to analyse the zoo data (in**[**zoo.arff**](http://csis.pace.edu/~benjamin/teaching/cs619/webfiles/zoo.arff)**):**

**Decision stump - weka.classifiers.DecisionStump**

**OneR - weka.classifiers.OneR**

**Decision table - weka.classifiers.DecisionTable -R**

**C4.5 - the J48 classifier**

**PART - under "rules"**

**How do the classifiers determine whether an animal is a mammal, bird, reptile, fish, amphibian, insect, or invertebrate? Do the decisions made by the classifiers make sense to you? What can you say about the accuracy of these classifiers when classifying an animal that has not been used for training? Why does OneR perform so badly?**

**ANSWERS-**

1. **Decision Stump** - The classification of animal was made on the basis of following assumptions:

Milk = false: bird

Milk! = false: mammal

Milk is missing mammal

Based on Confusion Matrix, classification of animals is incorrect as all the animals are classified as birds except mammal.

Correctly Classified instances i.e. accuracy rate is (61/101) = **60.396%.**

1. **OneR** - OneR is a classifier with one parameter, which provides results that were inaccurate except for mammal.

Correctly Classified Instances - (43/101) = **42.5743 %**

OneR performs badly than other classifiers because it produces rules only slightly less accurate than classification algorithms while producing rules that are simple for humans to interpret.

1. **Decision table –** Decision are made to determine the type of animals based on Best first and feature selection from set. These decisions are based on the milk, fins, legs, and tail and that’s why there is an error in each class except mammal, bird and fish.

Correctly Classified Instances - (87/101) = **86.1386 %**

1. **J48 -** The classification of animal was based on J48 pruned tree which has,

No. of leaves = 9, and size of tree = 17.

Based on Confusion Matrix, assumptions for mammal, bird and fish are classified correctly.

Correctly Classified Instances – (93/101) = **92.0792 %**

1. **PART** –

Number of Rules: - 8

PART decision list: -

1. feathers = false AND

milk = true: mammal (41.0)

1. feathers = true: bird (20.0)
2. backbone = false AND

airborne = false AND

predator = true: invertebrate (8.0)

1. backbone = false AND

legs > 2: insect (8.0)

1. fins = true: fish (13.0)
2. backbone = true AND

tail = true: reptile (6.0/1.0)

1. aquatic = true: amphibian (3.0)
2. : invertebrate (2.0)

The animals are classified based on the 8 decision rules. On observing the Confusion matrix, we can see that it is similar to the previous case of J48 tree classifier. Hence, the animal types ‘mammal’, ‘bird’ and ‘fish’ were classified correctly whereas other types were inaccurately predicted. Also, it has the same highest accuracy rate as the previous classifier.

Correctly Classified Instances – (93/101) = **92.0792 %**

The animals are classified based on the 8 decision rules. On observing the

Confusion Matrix, we can see that it is similar to the previous case of J48 tree

classifier. Hence, the animal types ‘mammal’, ‘bird’, and ‘fish’ were correctly

classified whereas other types were inaccurately predicted. Also, it has the same

highest accuracy rate as the previous class

**Q2. Use the following learning schemes to analyse the bolts data (**[**bolts.arff**](http://csis.pace.edu/~benjamin/teaching/cs619/webfiles/bolts.arff)**without the TIME attribute):**

* **Decision stump - weka.classifiers.DecisionStump**
* **Decision table - weka.classifiers.DecisionTable -R**
* **Linear regression - weka.classifiers.LinearRegression**
* **M5' - weka.classifiers.M5'**

**The dataset describes the time needed by a machine to produce and count 20 bolts. (More details can be found in the file containing the dataset.) Analyse the data. What adjustments have the greatest effect on the time to count 20 bolts? According to each classifier, how would you adjust the machine to get the shortest time to count 20 bolts?**

**ANSWERS -**

1. **Decision stump –** For greatest effect on time to count 20 bolts, it is based on attribute SPEED1. If SPEED1 is less than 5 then it will take the least amount of time to count 20 bolts.
2. **Decision table -** SPEED1 and TOTAL attributes have the greatest effect. Adjusting SPEED1 to 4 and TOTAL to 20, machine will take the least time.
3. **Linear Regression -** SPEED1 and SENS has greatest effect on time. Low value of SPEED1 and high value of SENS will make the machine to take the least time.
4. **M5’ -** It will first check the SPEED1 attribute and if SPEED1 is less than 5, machine will take the least time if we increase SPEED1 and RUN and lower value of TOTAL and SENS attribute.

If SPEED1 is more than 5, machine will take least time if we increase SPEED1 value and lower value of TOTAL and SENS.