Homework 04

1. **Data Exploration**

The number of rows (instances) = 1000
The number of columns (attributes)= 100
Types of Attributes: Nominal and Numeric.
Data type is Numeric for attributes from A5 – A99 and is Nominal for attributes from A1 to A4

Class distributions:

The class attribute has 2 class labels which are distributed uniformly. The count and weight of two classes are equally distributed. Attribute 1 to Attribute 4 is Uniformly distributed.

Pattern: From Attribute A4 to Attribute A99, We see the **101 distinct values** and **Mean and Standard deviation** which provides the information on how spread of the data and **SD** helps determine the outliers in the data. The Descriptive statistics of the attribute and histogram provides the spread of the data points in each attribute.

2. The table consists of 3 classifiers and Its respective Classification Accuracy and RMSE.

No	Classifier	Classification accuracy	RMSE
1	DecisionStump	52% (520)	0.5004
2	J48(pruned)	91% (910)	0.2923
3	J48(unpruned)	69.6% (696)	0.5417
4	Lazy-IBK**(k=8)**	100% (1000)	0.1121

3 a. Decision Stump:

A *decision stump* is a Decision Tree, which uses only a single attribute for splitting. For discrete attributes, this typically means that the tree consists only of a single interior node (i.e., the root has only leaves as successor nodes).It is a single based on one feature. The classifier model is constructed based on the following Classifications and its respective confusion matrix :

a0079 <= 0.875 : 2
a0079 > 0.875 : 1
a0079 is missing : 1

a b <-- classified as
147 353 | a = 1
127 373 | b = 2

The classification was depending on single **attribute of a0079** for missing and greater than 0.875 = 1, else 2.The accuracy rate is 52% for which are correctly classified for.

b.Pruned vs Unpruned:

	Number of Leaves	Size of the tree	Accuracy
Pruned	4	7	91%
Unpruned	41	81	69.90%

The performance of pruned seems better compared to Unpruned because if pruning is enabled, an additional step looks at what nodes/branches can be removed without affecting the performance too much thus reducing the risk **of overfitting the model.**The accuracy of the Pruned seems more than Unpruned.

c.LazyIBK:

K value	Classification Accuracy (%)	RMSE
1	97.1	0.1701
2	94.8	0.1466
3	99.2	0.1321
4	98.9	0.1212
5	99.9	0.1192
6	99.9	0.1149
7	100	0.1138
8	**100**	**0.1121**
9	100	0.1127
10	100	0.1125

The values of K as follows. The accuracy of the model increases when K =8 and RMSE IS 0.1121. The lower the RMSE value, the better it fits the model. The below stated confusion Matrix has correctly classified with 500 belong to class 1 and The Class 2 =500 which correctly classified.

```
a   b   <-- classified
500   0 |   a = 1
  0 500 |   b = 2
```