### **Hong Kong Baptist University**

### Department of Computer Science

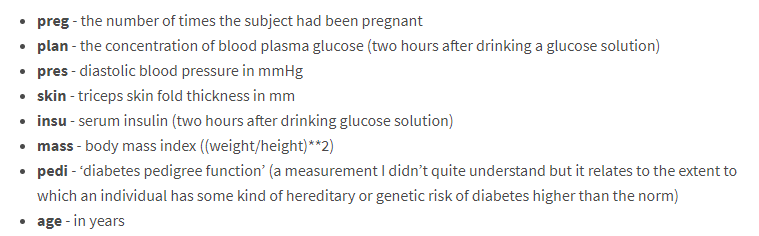
*COMP 7990 Principles and Practices of data analytics (2022-23)*

*Assignment 1*

Name: \_\_\_\_\_\_\_\_\_\_Li Tian\_\_\_\_\_\_\_\_\_\_\_\_ SID: \_\_\_\_\_\_\_\_22401792\_\_\_\_\_\_

**Exercise 1 - Classification using diabetes dataset**

1. Open the **Weka** **Explorer**. Under **Preprocess** panel, open the file **diabetes.arff**. There are some attributes as shown below and the last attribute **class** is used to indicate whether a person has diabetes or not :



1. Answer the following questions:
2. How many instances are there in the diabetes dataset? \_\_\_\_768\_\_\_\_\_\_\_\_\_
3. How many attributes are there in the diabetes dataset? \_\_\_\_\_\_9\_\_\_\_\_\_\_
4. How many people tested positive? \_\_\_\_\_268\_\_\_\_\_\_\_
5. Use IBk algorithm to create different training models. Test option is **10-fold cross-validation.** Select **(Nom) class** as class attribute (which).Fill in the results by using the table below.
6. IBk (with KNN value = 5)
7. IBk (with KNN value = 10)
8. IBk (with KNN value = 15)

**Test option: 10-fold cross-validation**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **IBk**  **(KNN value = 5)** | **IBk**  **(KNN value = 10)** | **IBk**  **(KNN value = 15)** |
| **Classification accuracy (in %)** | 73.1771 | 71.0938 | 74.0885 |
| **Number of Correctly Classified Instances** | 562 | 546 | 569 |
| **Confusion matrix (paste the Screenshot)** |  |  |  |

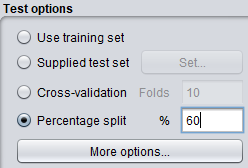
1. In the last model with the classifier **IBk** (**KNN value = 15**):
2. How many instances of type positive are misclassified as negative? \_\_\_\_\_130\_\_\_\_\_\_\_\_
3. How many instances of type negative are misclassified as positive? \_\_\_\_\_\_69\_\_\_\_\_

**Exercise 2 - Classification using diabetes dataset (Percentage split)**

1. Continue to use the file **diabetes.arff** with the classifier **IBk** (**KNN value = 15**):



1. To split diabetes dataset into training and testing set. Select **Percentage split** and type **60** as shown below.



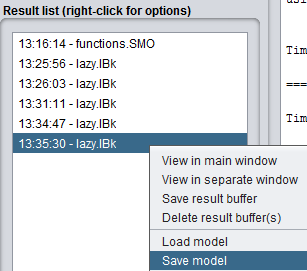
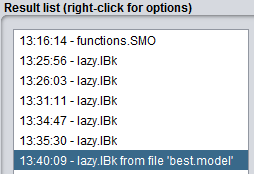
1. Click **Start**. Fill in the answers for the percentage split **60%**. Repeat the steps by using percentage splits **70%** and **80%** with the same KNN value. Then **close** the **Weka Explorer**.

|  |  |  |  |
| --- | --- | --- | --- |
| **IBk**  **(KNN value = 15)** | **Training set = 60%** | **Training set = 70%** | **Training set = 80%** |
| **Testing data %** | 40 | 30 | 20 |
| **Classification accuracy (in %)** | 74.5928 | 76.087 | 77.2727 |
| **Number of instances for training** | 461 | 538 | 614 |
| **Number of instances for testing** | 307 | 230 | 154 |
| **Correctly Classified Instances** | 229 | 175 | 119 |

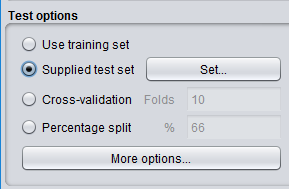
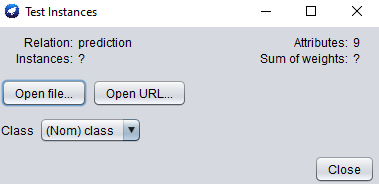
1. Which of the following statement(s) is/are true (based on your observation in step 3)? \_\_\_a\_\_\_
2. In general, the more the training data, the greater the classification accuracy.
3. Training set size has no influence on the classification accuracy.
4. The more the test data, the greater the classifier’s success rate/accuracy.
5. None of the above

**Exercise 3 - Making prediction for diabetes using new instances**

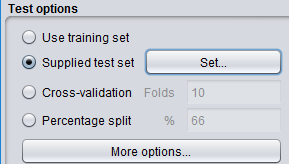
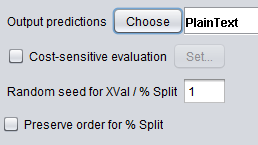
1. **Save** the last model (**IBk** with **KNN value = 15, training set = 80%**) by using the filename **best.model** (save on Desktop). Then load the saved model using **Load model** option

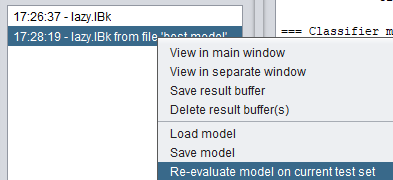
1. Under **Test options**, select **Supplied test set,** click **Set** button next to it to predict the class for some new instances. Open the file **prediction.arff** for new prediction. Keep  Then press **Close.**

1. Click **More options** under **Test options**. Output prediction as **PlainText**. Press **OK**

1. **Right click** the model you just load and choose **Re-evaluate model on current test set**.



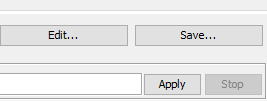
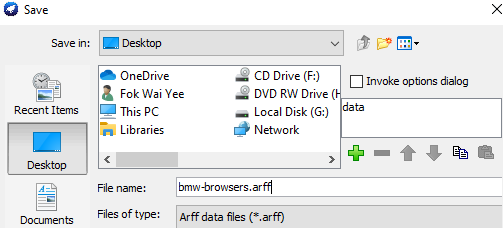
1. Fill in the prediction result:

|  |  |
| --- | --- |
|  | **Prediction result (negative or positive)** |
| **Instance 1** | **tested\_negative** |
| **Instance 2** | **tested\_positive** |
| **Instance 3** | **tested\_negative** |
| **Instance 4** | **tested\_positive** |
| **Instance 5** | **tested\_negative** |
| **Instance 6** | **tested\_positive** |

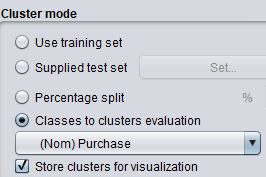
1. **Close** the **Weka Explorer**.

**Exercise 4 - Clustering using SimpleKMeans**

1. Open the **Explorer**. Under **Preprocess** panel, open the file **bmw-browsers.csv**.
2. Click **Save** button to save the file with the name **bmw-browsers.arff**.

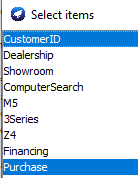
 

1. Go to the **Cluster** panel. Choose **SimpleKMeans** asalgorithm. **Keep** **the number of clusters = 2 (seed = 10)**. In the **Cluster mode** tab, select “**Classes to clusters evaluation**”



1. Click on **Ignore attributes** button to **ignore** the **CustomerID** and **Purchase** fields by holding down **Ctrl** key.





1. Fill in the answers for the following table. Try other configurations to create and evaluate the models.

**Clustering results:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Number of clusters** | **2** | **3** | **4** | **4** | **3** | **2** |
| **seed** | **10** | **10** | **10** | **11** | **11** | **11** |
| Number of iterations | 7 | 6 | 5 | 5 | 4 | 2 |
| Within cluster sum of squared errors | 127.143 | 105.444 | 95.758 | 104.6111 | 116.182 | 138.865 |
| Incorrectly clustered instances | 43.0 | 62.0 | 59.0 | 64.0 | 61.0 | 50.0 |

For the last model (number of clusters = 2, seed = 11):

Number of instances of type **yes** that are misclassified as no (cluster 0) is \_\_\_\_\_\_22\_\_\_\_\_\_\_\_\_

Number of instances of type **no** that are misclassified as yes (cluster 1) is \_\_\_\_\_\_28\_\_\_\_\_\_\_\_\_\_

# **Assignment Submission**

Submit the file **lab3b-assignment-ans.docxs** to bulearning website