Faculty of Applied Sciences

Wayamba University of Sri Lanka

CMIS 3253 – Data Mining

Practical Assignment

In this task, you are required to explore machine learning algorithms in Python for classification problems. You should perform the following tasks.

1. Download and install Python.
   https://www.python.org/downloads/

2. Download the dataset relevant to your Registration number (as mentioned below) from UCI machine learning repository datasets.
   https://archive.ics.uci.edu/ml/index.php

Ser.No	Reg. No	Name	Data set
1	182005	Ms. P.A.I. Anjana	Abalone
2	182009	Ms. W.B.A.N. Bandara	Higher Education Students Performance Evaluation Dataset
3	182011	Mr. P.S. Bazanbeg	Hepatitis
4	182022	Ms. A.H.T. Dilshani	Adult
5	182029	Mr. P.S.H. Ekanayaka	Lenses
6	182036	Mr. W.K.S.A.T.D. Fernando	Maternal Health Risk Data Set
7	182039	Mr. G.G.A.D.Y.S. Ganegoda	Algerian Forest Fires Dataset
8	182047	Ms. B.K. Hasara	Anticancer peptides
9	182061	Mr. A.L.D. Kanisha	Leaf
10	182065	Ms.A.H.O.L. Kaushani	Audit Data
11	182072	Mr. G.V.A.K. Kumara	Raisin Dataset

12	182085	Ms. W.M.N.W.B. Manike	Bone marrow transplant: children
13	182103	Mr. Perera W.M.S.T.	Caesarian Section Classification Dataset
14	182107	Mr. H.K.T. Pushpashan	Chemical Composition of Ceramic Samples
15	182121	Mr. H.J. Sandaruwan	Chronic_Kidney_Disease
16	182126	Ms. H.A.A. Thathsarani	Credit Approval Data Set
17	182127	Mr. T.M.T.S. Thennakoon	Diabetic Retinopathy Debrecen Data Set
18	182152	Ms. E.H.D.M.N. Ehelapitiya	Dry Bean Dataset
19	182160	Ms. Kaumadi I.A.S.	Estimation of obesity levels based on eating habits and physical condition
20	182164	Ms. B.R. Madhurangi	Forest type mapping
21	182175	Ms. D.M.R. Wickramanayaka	Glass Identification

3. Using Python, apply the following six classifiers on the dataset.
    - Decision Tree
    - K-Nearest Neighbors (KNN)
    - Linear Discriminant Analysis
    - Logistic Regression
    - Naive Bayes
    - Support Vector Machine (SVM)

4. You should study and configure the parameters of each classifier and obtain the best classification performance.

5. Create a report including the performance of each classifier on the dataset. For each classification task, you should include the confusion matrix, true positive rate, false positive rate, precision, recall, F-measure, and ROC area (AUC).

6. Identify the best classifier that describes your dataset. Give your reasons for choosing it.