CDS503: Machine Learning
Academic Session: Semester 1, 2021-2020

School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Task Type

Assignment 1 is an individual assignment.

The student is required to choose one of the listed problems/data sets in **Appendix A** and propose the solution to the problem.

You are NOT required to perform an experiment on the chosen algorithms.

Assignment Description

Topic Selection: See *Appendix A*.

Based on the title chosen:

1. **Data set background and characteristics**. Study the data set chosen carefully. Based on the data set:
   a. The background study of the data set. Perform the literature that covers (not limited to):
      (i) The usage of the dataset and the trend, including the proposed approaches
      (ii) Search for the literature for the past 5 years (2016 and above) that has used the chosen dataset or similar ones. The required information includes the title, the authors, year, journal/conferences, and age number.
   b. Report the class distribution of the given dataset (dataset in Appendix A).
   c. Determine whether the data set is balanced or unbalanced. Give your justification and explain how this condition will affect the performance.
   d. **Feature correlations**. Discuss and do a research on how the features correlations affect the performance of your classifier.
2. **Pre-processing** options. Discuss and select the suitable pre-processing options.
3. **Model Evaluation Technique**. Based on the size and characteristics of the data set, choose the suitable *Model Evaluation* techniques to be used in your machine learning evaluation. Explain the reason for using that option (hold-out, cross validation etc.).
4. **Choice of the classifier**. Choose **one** classifier that you have learned in the first half of the semester.

   a. State the reasons of your choice based on the data set characteristics (and any others) and how these choices will affect the performance. You may justify your choice based on the literature done in 1(a).

   b. Based on data set, select, and justify a suitable metric to evaluate the performance of your classification model (Confusion matrix, F1 Score etc.).

CDS503: Machine Learning
Academic Session: Semester 1, 2021-2020
School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Report Requirement and Format

- A report must be prepared using Microsoft Word, font type Arial, size 12 in single line. Every chapter should start with a **new** page (Chapter 1.0 to 5.0, and references).
- A **cover page** should contains course name (including semester and year), assignment title, name, matrix no and dataset title.
- Table of Contents

1.0 Dataset Background
2.0 Pre-processing options
3.0 Model Evaluation Technique
4.0 Choice of the classifier
5.0 Conclusion

References

Note: You may create additional subsection as deemed necessary**.**

Report Submission Instruction

- Submit soft copy (zip/rar to eLearn@USM).
- The zip/rar package must be named according to the following notation: ***CDS503_Assignment 1_Name_MatrixNo_TitleNo.***

Assignment Evaluation

This assignment will be graded (A to F scale).

IMPORTANT: Students who copied or plagiarized other's work or let their work be copied or plagiarized will be given an F grade. The student may be barred from sitting for final exam and reported to the university's disciplinary board.

Assignment Due Date: Tuesday, 14 December 2021 5:00 pm.

CDS503: Machine Learning
Academic Session: Semester 1, 2021-2020

School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Grading Rubric – Assignment 1

Course Learning Outcome (CLO):

- CLO1 Describe concepts, theories, and implementation of machine learning algorithms.
- CLO3 Apply relevant machine learning algorithms for typical real-world problems.

Rubrics

Component	2-1 (Poor)	5-3 (Average)	8-6 (Good)	10-9 (Excellent)	Weight
Dataset set background and characteristics	Dataset description is **absent**.	Dataset description is **minimal**.	Dataset description is **adequately** complete.	Dataset description is **complete** and **comprehensive**.	35%
Pre-processing options	Pre-processing options are **minimally** discussed and justified.	Pre-processing options are **fairly** discussed and justified.	Pre-processing options are **adequately** discussed and justified.	Pre-processing options are **clearly** discussed and justified.	10%
Model Evaluation Technique	The model is **poorly** presented, and discussion of the model is **absent**. Insights and contributions are **poorly** discussed or absent.	The best-suited model is **minimally** discussed and justified. Insights from the analysis are **vague.**	The best-suited model is **fairly** discussed and justified. Insights from the analysis are **less evident,** and contribution is **fairly** discussed.	The best-suited model is **clearly** discussed and justified. Insights from the analysis evident and contribution are discussed and well-explained.	15%
Choice of the classifiers	The choice of classifiers is **minimally** discussed based	The choice of classifiers is **fairly** discussed based on the	The choice of classifiers is **adequately** discussed	The choice of classifiers is **clearly** discussed based on the LR and justified.	25%

	on the LR and justified.	LR and justified.	based on the LR and justified.		
Conclusion & References	The conclusion is **absent,** and no references provided.	The conclusion is of **simplistic** summary and **few** references are provided.	The conclusion is a **partially** complete summary and **adequately** references are given.	The conclusion contains a **comprehensive** summary and good references are provided.	10%
Report Formatting	Some writings are inaccurate and unclear. Follow the format given and somewhat organized.	Some writings are inaccurate and unclear. Follow the format given and somewhat organized.	Most writings are accurate, clear and concise. Somewhat follow the format and organized.	Most writings are accurate, clear and concise language used throughput. Report follows the format given and is properly arranged and well-organized.	5%

CDS503: Machine Learning
Academic Session: Semester 1, 2021-2020

School of Computer Sciences, USM, Penang

ASSIGNMENT 1

Appendix A

Assignment No	Name	Link
T01	Pima Indian Diabetes Dataset	https://www.kaggle.com/uciml/pima-indians-diabetes-database
T02	Rain in Australia	https://www.kaggle.com/jsphyg/weather-dataset-rattle-package
T03	The Estonia Disaster Passenger List	https://www.kaggle.com/christianlillelund/passenger-list-for-the-estonia-ferry-disaster
T04	Airline Passenger Satisfaction	https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction
T05	Wheat seeds dataset	https://www.kaggle.com/jmcaro/wheat-seedsuci
T06	Early-stage diabetes risk prediction dataset	https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.
T07	Autism Screening Adult Dataset	https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult
T08	Bank Marketing Dataset	https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
T09	Breast Cancer Wisconsin (Diagnostic) dataset	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
T10	Predict 5-Year Career Longevity for NBA Rookies	https://data.world/exercises/logistic-regression-exercise-1
T11	Diabetic Retinopathy Debrecen Data Set	https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set#
T12	Cervical cancer (Risk Factors) Data Set	https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
T13	Titanic Disaster Dataset	https://data.world/nrippner/titanic-disaster-dataset
T14	Car Evaluation Data Set	https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
T15	Leaf Data Set	https://archive.ics.uci.edu/ml/datasets/Leaf
T16	Tic-Tac-Toe Endgame Data Set	https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame

CDS503: Machine Learning
Academic Session: Semester 1, 2021-2020

School of Computer Sciences, USM, Penang

ASSIGNMENT 1

T17	Mushroom Data Set	https://archive.ics.uci.edu/ml/datasets/Mushroom
T18	Wine Data Set	https://archive.ics.uci.edu/ml/datasets/Wine
T19	Mobile Price Classification	https://www.kaggle.com/iabhishekofficial/mobile-price-classification?select=train.csv
T20	Fetal Health Classification	https://www.kaggle.com/andrewmvd/fetal-health-classification
T21	Ad Click Prediction	https://www.kaggle.com/jahnveenarang/cvdcvd-vd
T22	Paris Housing Classification	https://www.kaggle.com/mssmartypants/paris-housing-classification
T23	Water Quality	https://www.kaggle.com/mssmartypants/water-quality
T24	Symptoms and COVID Presence	https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence
T25	Start-up Success Prediction	https://www.kaggle.com/manishkc06/startup-success-prediction