SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ENG335

End-of-Course Assessment - July Semester 2023

Machine Learning

INSTRUCTIONS TO STUDENTS:

1. This End-of-Course Assessment paper comprises **FIVE (05)** pages (including the cover page).

2. You are to include the following particulars in your submission: Course Code, Title of the ECA, SUSS PI No., Your Name, and Submission Date.

3. Late submission will be subjected to the marks deduction scheme. Please refer to the Student Handbook for details.

IMPORTANT NOTE

ECA Submission Deadline: 06 Nov 2023, Monday, 12 noon

ECA Submission Guidelines

Please follow the submission instructions stated below:

This ECA carries 70% of the course marks and is a compulsory component. It is to be done individually and not collaboratively with other students.

Submission

You are to submit the ECA assignment in exactly the same manner as your tutor-marked assignments (TMA), i.e. using Canvas. Submission in any other manner like hardcopy or any other means will not be accepted. You are to ensure that the file to be submitted does not exceed 20MB in file size.

Electronic transmission is not immediate. It is possible that the network traffic may be particularly heavy on the cut-off date and connections to the system cannot be guaranteed. Hence, you are advised to submit your assignment the day before the cut-off date in order to make sure that the submission is accepted and in good time.

Once you have submitted your ECA assignment, the status is displayed on the computer screen. You will only receive a successful assignment submission message if you had applied for the e-mail notification option.

ECA Marks Deduction Scheme

Please note the following:

a. Submission Cut-off Time – Unless otherwise advised, the cut-off time for ECA submission will be at 12:00 noon on the day of the deadline. All submission timings will be based on the time recorded by Canvas.
b. Start Time for Deduction – Students are given a grace period of 12hours. Hence calculation of late submissions of ECAs will begin at 00:00 hrs the following day (this applies even if it is a holiday or weekend) after the deadline.
c. How the Scheme Works – From 00:00 hrs the following day after the deadline, 10 marks will be deducted for each 24-hour block. Submissions that are subject to more than 50 marks deduction will be assigned zero mark. For examples on how the scheme works, please refer to Section 5.2 Para 1.7.3 of the Student Handbook.

Any extra files, missing appendices or corrections received after the cut-off date will also not be considered in the grading of your ECA assignment.

Plagiarism and Collusion

Plagiarism and collusion are forms of cheating and are not acceptable in any form of a student's work, including this ECA assignment. You can avoid plagiarism by giving appropriate references when you use some other people's ideas, words or pictures (including diagrams). Refer to the American Psychological Association (APA) Manual if you need reminding about quoting and referencing. You can avoid collusion by ensuring that your submission is based on your own individual effort.

The electronic submission of your ECA assignment will be screened through a plagiarism detecting software. For more information about plagiarism and cheating, you should refer to the Student Handbook. SUSS takes a tough stance against plagiarism and collusion. Serious cases will normally result in the student being referred to SUSS's Student Disciplinary Group. For other cases, significant marking penalties or expulsion from the course will be imposed.

Additional Instructions for Submission:

1. Please submit a Word document with screenshots of the Jupyter notebook and a zipped folder container Jupyter notebook file (.ipynb) for each question in the respective folders via the Canvas T- group.
2. All answers for each question should be indicated clearly using the Comments section/markups in the Notebook so that the marker can see clearly which code is for which Question. (e.g. # Answer for Q1).

You are required to prepare the dataset wherever necessary.

Answer all questions. (100 marks)

Question 1

Download the Lyft Inc dataset from Kaggle (https://www.kaggle.com/datasets/dermisfit/lyft-inc-dataset). Understand the dataset by performing exploratory analysis. Prepare a new dataset by excluding the "date and day" and "year" attributes. You need to also drop **TWO (02)** more attributes. If you don't exclude these two attributes, you will get a perfect/ideal estimator. Design a linear regression model to estimate the bike demand using only **FOUR (04)** best attributes from the newly constructed dataset. Discuss your results and the relevant metrics. If you include all the features of the new dataset, does that give a better model. Would you use the model that employs all the features for the prediction of the bike demand?

(20 marks)

Question 2

Load the Wrestling World Tournament dataset from Kaggle (https://www.kaggle.com/datasets/julienjta/wrestling-world-tournament). The objective is to detect the gender of the wrestler given the other parameters. Perform exploratory data analysis. Analyze and drop the appropriate features and suitably encode the categorical features. Design a simple neural network classifier with **ONE (01)** hidden layer. Construct the Naïve Bayes classifier for the above problem. Adjust the parameters of the neural network algorithm such that it has the same or better performance than the Naïve Bayes classifier.

(20 marks)

Question 3

Download the Sloan Digital Sky Survey DR16 dataset available in Kaggle (https://www.kaggle.com/datasets/muhakabartay/sloan-digital-sky-survey-dr16). Prepare the dataset by dropping the features ['objid', 'run', 'rerun', 'camcol', 'plate', 'field', 'mjd', 'fiberid', 'specobjid', 'redshift'] and perform exploratory data analysis. Propose optimal values for the depth and number of trees in the random forest.

(20 marks)

Question 4

Use the cat vs rabbit dataset available in the Kaggle (https://www.kaggle.com/datasets/muniryadi/cat-vs-rabbit). You can use example codes (from Kaggle or other resources) to download and load the data properly into the programming environment. Perform exploratory data analysis and show a random sample of **SIX (06)** images each for the cat and the rabbit. Design a CNN with **TWO (02)** convolutional layers and **THREE (03)** dense layers (including the final output layer). Employ 'tanh' activation and MaxPooling. Keep 18% of the training dataset for validation and use at least 10 epochs. Note: Use the data in train-cat-rabbit folder to create your training and validation datasets. Use the data in val-cat-rabbit as your test dataset to rate the performance of the algorithm.

(20 marks)

Question 5

Select any stock listed in Singapore stock exchange. Using Yahoo finance, download the daily stock data (Open, High, Low, Close, Adj Close, Volume). Download the data such that 8 years of data up to the last working day of December 2021 can be used for training and the data from the 1st working day of 2022 till the last working day of year 2022 can be used as test data. Use the previous 52 days of stock information (High and Volume) to predict the next day stock price (High). Design an LSTM network to do the predictions. You are required to use LSTM with a cell state of at least 100 dimensions and do at least 50 epochs of training. Rate the performance of the LSTM classifier and provide necessary plots.

(20 marks)

----- END OF ECA PAPER -----