

# **FIN559**

**End-of-Course Assessment – January Semester 2022**

## **Big Data, Cloud Computing and Machine Learning**

---

### **INSTRUCTIONS TO STUDENTS:**

1. This End-of-Course Assessment paper comprises **FIVE (5)** 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.

<p><b>IMPORTANT NOTE</b></p> <p><b>ECA Submission Deadline: 30 March 2022, 12 noon</b></p>
--

## **ECA Submission Guidelines**

*Please follow the submission instructions stated below:*

### **A - What Must Be Submitted**

*You are required to submit the following TWO (2) items for marking and grading:*

- *A Report (you **should submit this item first** as it carries the highest weightage).*
- *A well-documented Jupyter Notebook Script (.ipynb).*

*Please verify your submissions after you have submitted the above TWO (2) items.*

### **B - Submission Deadline**

- *The TWO (2) items of Report and the well-documented Jupyter Notebook Script (.ipynb) are to be submitted **by 12 noon** on the submission deadline.*
- *You are allowed multiple submissions till the cut-off date for each of the TWO (2) items.*
- *Late submission of any of the TWO (2) items **will be subjected to mark-deduction scheme** by the University. Please refer to Section 5.2 Para 2.4 of the Student Handbook.*

### **C - How the TWO (2) Items Should Be Submitted**

- *The Report: submit online to Canvas via TurnItIn (for plagiarism detection), folder name as ECA.*
- *The Jupyter Notebook Script: submit online to Canvas, folder name as Working File for ECA (.ipynb).*
- *Avoid using a public WiFi connection for submitting large video files. If you are using public wireless (WiFi) connection (e.g. SG Wireless at public areas), you might encounter a break in the connection when sending large files.*

### **D - Additional guidelines on file formatting are given as follows:**

<b>1. Report</b>	<ul style="list-style-type: none"><li>• <i>Please ensure that your Microsoft Word document is generated by Microsoft Word 2007 or higher.</i></li><li>• <i>The report must be saved in <b>.docx</b> format.</i></li></ul>
<b>2. Jupyter Notebook Script</b>	<ul style="list-style-type: none"><li>• <i>The Script must be saved in <b>.ipynb</b> format.</i></li></ul>

### ***E - Please be Aware of the Following:***

*Submission in hardcopy or any other means not given in the above guidelines will not be accepted. You do not need to submit any other forms or cover sheets (e.g. form ET3) with your ECA.*

*You are reminded that electronic transmission is not immediate. The network traffic may be particularly heavy on the date of submission deadline and connections to the system cannot be guaranteed. Hence, you are advised to submit your work early. **Canvas will allow you to submit your work late but your work will be subjected to the mark-deduction scheme.** You should therefore not jeopardise your course result by submitting your ECA at the last minute.*

*It is your responsibility to check and ensure that your files are successfully submitted to Canvas.*

### ***F - Plagiarism and Collusion***

*Plagiarism and collusion are forms of cheating and are not acceptable in any form in a student's work, including this ECA. Plagiarism and collusion are taking work done by others or work done together with others respectively and passing it off as your own. You can avoid plagiarism by giving appropriate references when you use other people's ideas, words or pictures (including diagrams). Refer to the 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 will be screened by plagiarism detection software. For more information about plagiarism and collusion, you should refer to the Student Handbook (Section 5.2.1.3). You are reminded that SUSS takes a tough stance against plagiarism or collusion. Serious cases will normally result in the student being referred to SUSS's Student Disciplinary Group. For other cases, significant mark penalties or expulsion from the course will be imposed.*

**You must answer ALL the questions. (100 marks)**

### **Question 1**

- (a) Read about the term “Black box ML”. In your own words explain this term and appraise the need to interpret Black box ML. (5 marks)
- (b) Read about DOSS from <https://cloud.google.com/blog/topics/startups/doss-makes-home-buying-and-selling-smarter-with-digital-real-estate-marketplace>.
- (i) Formulate the problem statement that is being solved in DOSS. (4 marks)
- (ii) From the information provided in the blog, identify the technologies being used by DOSS. Assess whether the above identified technologies be used in Singapore. What would be the challenges and how would you overcome it. (6 marks)
- (iii) What prompted DOSS to move from AWS to Google and what are the Google products currently used by DOSS? (3 marks)
- (iv) Identify **TWO (2)** Google Cloud products not listed in Question 1(b)(iii) that you will use to offer a service similar to DOSS in Singapore. State your reason for selecting the Google Cloud product. (2 marks)

### **Question 2**

Read about AWS IoT SiteWise <https://aws.amazon.com/iot-sitewise/>. You plan to adopt AWS IoT SiteWise to monitor the temperature and oxygen level in a remote fish farm in Singapore. There are 20,000 sensors deployed to measure the oxygen level and 10,000 sensors deployed to measure the temperature. Both oxygen and temperature are measured every 10 minutes. Six auto-computed aggregates for oxygen level and temperature are calculated over three intervals, viz., 20 minutes, 1 hour and 1 day. AWS charges you based on the messaging, data processing and data storage consumption. Appraise the following costs by making appropriate assumptions wherever needed.

- (a) Messaging cost incurred per month. (8 marks)
- (b) Data processing cost incurred per month (3 marks)
- (c) Data storage cost incurred per month (8 marks)

(d) Total cost of AWS IoT SiteWise service per month

(1 mark)

### Question 3

Download the MAGIC gamma telescope data 2004 dataset available in Kaggle (<https://www.kaggle.com/abhinand05/magic-gamma-telescope-dataset>). Understand the dataset and perform exploratory data analysis and implement a decision tree using 'entropy' criterion for identifying whether the pattern was caused by gamma signal or not. Get the tree depth, performance metrics and number of leaves in the tree before and after optimisation for the tree depth. For optimisation, use 4-fold cross validation. Show the optimised decision tree. Compare the performance of the decision tree before and after depth optimisation.

(20 marks)

### Question 4

Load the wine dataset from sklearn package. Perform exploratory data analysis and set up a KNN classifier. Propose an appropriate value for K. Compare the performance metrics of the KNN classifier with an appropriate algorithm.

(20 marks)

### Question 5

Use the Fashion MNIST dataset from the keras package. Perform exploratory data analysis. Show a random set of **SIX (6)** images from each class in the dataset with their corresponding class names. Prepare the dataset by normalising the pixel values to be between 0 and 1. Design a CNN with **three (3)** convolutional layers and **three (3)** dense layers (including the final output layer). Employ 'ReLU' activation and 'MaxPooling'. Keep 20% of the train dataset for validation. Rate the performance of the algorithm and provide necessary plots. Pick a random image from the test dataset, pass it to the algorithm and compare the algorithm output with the actual class label.

(20 marks)

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