



# **ENG335** **Machine Learning**

---

## **Group-based Assignment**

### **July 2023 Presentation**

---

## **GROUP-BASED ASSIGNMENT**

This mini-project assignment is worth **15%** of the final mark for **ENG335 Machine Learning**. Total mark assigned to this assignment is 100 marks.

The cut-off date for this assignment is **16 Sep 2023, 23 55 hrs.**

This is a group-based assignment. You should form a group of **maximum** 5 members from your seminar group. Each group is required to upload a single report to Canvas Turnitin via your respective seminar group. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. In your 1-page cover sheet, please include all project partners' names and student PI numbers.

### Note to Students:

You are to submit the GBA assignment i.e. using Canvas in the form of a single MS Word file. It should be saved as ENG335\_GBA01\_group\_number.doc 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.

### **Additional Instructions for Submission:**

*Please follow the submission instructions stated below:*

1. Please submit all Program Code / Answers in the form of a Jupyter Notebook file (i.e. .ipynb File) for all the programming questions via the additional submission link found under Assignments on ENG335 T01 course site.
  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 Q1a).
  3. Submit the file before the submission cut-off date/time via the Canvas T-group course site in **Assignments> GBA01**. You will be then directed to the submission page. Late submission will be subjected to the mark deduction scheme. Please refer to the Student Handbook for details.
-

*Answer all questions (100 marks)*


### Question 1

- (a) Read about generative models and discuss why they can be unsupervised or semi-supervised algorithms. (10 marks)
- (b) Learn about ChatGPT and appraise if it is a search engine. In your own words explain how ChatGPT works. Assess if ChatGPT or chatbots can replace search engines. Try prompting the above question to ChatGPT and provide snapshot of the response. (10 marks)
- (c) Read about Document AI (<https://cloud.google.com/document-ai>). Obtain a receipt or bill and upload it in the demo section. Take a snapshot of your observation. Based on the observation, explain how the document was processed using a maximum of **THREE (03)** sentences. Can Document AI perform document translation? (5 marks)

### Question 2


The dataset is available in the below link.

<https://www.kaggle.com/datasets/abrambeyer/openintro-possum>

- (a) Perform exploratory data analysis and understand the parameters. (7 marks)
- (b) The objective is to estimate the possum's total length. Construct a new dataset by only selecting the following possum's features: age, head length, tail length, skull width, foot length, ear conch length, distance from medial canthus to lateral canthus of right eye, chest girth and belly girth. Estimate the possum's total length using the best **THREE (03)** features from the above selection.  (12 marks)
- (c) Assess the performance of the linear regressor by getting the relevant performance metrics. You need to provide any **THREE (03)** metrics and explain the importance of these metrics. (6 marks)

### Question 3

Download the “Pistachio Types Detection” dataset from Kaggle (<https://www.kaggle.com/datasets/amirhosseinmirzaie/pistachio-types-detection?select=pistachio.csv> ).

- (a) Perform exploratory data analysis and understand the dataset. Use Python code to encode the target variable. Implement a suitable algorithm from what you have learned in the class for identifying the pistachio types. (14 marks)
- (b) Implement a Naïve Bayes algorithm for the above constructed dataset. (5 marks)
- (c)  Compare the performance metrics of the algorithm in Question 3(a) and Naïve Bayes classifier. Does the scaling of the parameters have any impact on the performance (Justify your answer)? (6 marks)

### Question 4

Use the train.csv dataset in <https://www.kaggle.com/datasets/iabhishekoofficial/mobile-price-classification?select=train.csv> . This train.csv will be your full dataset. You must construct the train and test datasets from train.csv for training and evaluation of your algorithm.

Understand the dataset by performing exploratory data analysis. Your objective is to identify the mobile phone price range by training a suitable algorithm. The target column in the dataset is already numbered as 0, 1, 2 and 3. Identify which number corresponds to the highest mobile price. Identify the algorithm (from what has been covered in the seminars) suitable for this identification of the mobile price range. Present appropriate performance metrics. (25 marks)

-----END OF GBA ASSIGNMENT-----