# Course CDS527 Big Data Analytics 2024-2025 Term2 Group Project (60%)

## Group Project (40%)

There are two tasks in this project: a **system development** and a **case study**. For system development, your group (5-6 people) will develop a decision support system. For case study, your teacher will give you a company case with business problems. Based on the Big-Data knowledge you know, suggest a solution to the case. The details of the two tasks are as follow:

## Task 1: System Development (30%)

Your group will be given a business textual dataset, with a standardized data splitting and evaluation metric. Your task is to develop a “*good-performing*” decision support system that classifies text. To achieve this, you will need to:

1. Build a **baseline** model (i.e., a model with basic setup and no fine-tuning, e.g., logistic regression)
2. Compare with **a range of different** **data analytics models** (e.g., Gradient boosting tree, decision tree, etc) and **different** **word embeddings models** (BERT, GloVE, Word2vec, etc)
3. Fine-Tune different **hyper-parameter** for each model above and report the best score
4. **Visualize with a range of different data statistics methods**. There are many ways that can describe patterns found in data include chart and figures (e.g., bar charts, line graphs, word cloud, word distribution, etc), as well as statistical metrics like central tendency (mean, mode and median) and dispersion (range, variance, maximum, minimum, quartiles), correlation and standard deviation.
5. Explore different methods to improve your baseline models. Marks will be given based on: ***the novelty of the methods*** (e.g., if the methods not been covered in classes/in the project notes), ***the variety of methods*** (e.g., develop multiple, feasible methods), ***the difficulty of implementation*** and the ***improvement*** brought about by the methods.

In your submission (i.e., Jupyter Notebook file), you need to include t**he code implementation for Point 1-5;** and **a short table and/or text (less than 300 words)** that describe Point 1-5 (e.g., what model, hyper-parameter, visualization techniques you have tried and/or their scores and details).

You can use any visualization and model improvement techniques learnt in the classes and/or include other techniques that have not been covered in the classes. However, you have to use **the SAME data splitting and evaluation metric throughout the project** (marks will be deduced if these three conditions are found changed in the submission).

## Task 2: Case Study (10%)

Each group will be assigned one of the following company cases:

* Airbnb
* Amazon
* Apple
* BBC
* Facebook
* Fitbit
* Google
* Ralph Lauren
* Twitter
* Walmart
* Others

Your teacher will give you the case study file. The case study will mention the background of a company and the data/problem it current has. As a Big Data professional, your group will be responsible to provides **a background description of the case study**, **how you approach the problem and/or utilize their data**, **a description of the solution** and **a critical evaluation of the solution**.

When you describes the case study. You may consider the characteristics of Big-Data (e.g., 5Vs) the company has; the types of data / data formats (e.g., any structured, un/semi-structured data); the business incentive for Big-Data solution adoption. When you approach the business problem and/or data, you can consider the solution from Big-Data **analytics** perspectives. For example, can statistical analysis (e.g., A/B testing), visual analysis (e.g., Heat maps), machine learning (e.g., clustering) and semantic analysis (e.g., text analysis) help the company? Finally, to evaluate your solution, you may consider mentioning some KPI (key performance indicators) like SMART.

## Oral Presentation (20%)

Each student will present. The contents of the oral presentation should be consistent with that of the written report. The presentation will be about 15 mins. Your teacher will tell you the timeslot your group is assigned to present.

## Submission (25-April-2025 23:59 HK time)

Each group should submit three files onto the Moodle (you can find a representative groupmate and upload the three files for the team **with your group number in your filename**):

1. A Jupyter Notebook file containing all deliverables regarding the system development task (e.g., the model performance, evaluation description and visualization)
2. A written report in MS Word file (**3 pages max**) that provides a background description of the case study, how you approach the problem, a description of the solution, and a critical evaluation of the solution.
3. A PowerPoint file that you use in the presentation covering highlights in (1) and (2).

## **Each** student needs to upload **a work distribution file** like below in word file to the Moodle saying the work distribution of you and your groupmate (you don’t need to let your group mate to know the work distribution you feel)

Group X:

|  |  |  |
| --- | --- | --- |
| Members (Full name | Student ID | Work Distribution |
| Chan Tai Man | 000000 | 20% |
| Wong Tai Man | 000001 | 20% |
| Chen Tai Man | 000002 | 20% |
| Tsang Tai Man | 000003 | 20% |
| Ng Tai Man | 000004 | 20% |

## 

## Assessment

Your submission will be assessed based on the following mark schemes.

Marking Scheme for Group project (40%)

* Implementation of a big analytical data management system and decision support system using professional software (20%): Whether your group have a good-performing system and a comprehensive evaluation
* Problem definition (demonstrate the understanding of the problem and formulate alternative solutions) (10%): Within the scope of the Big-Data area, whether your group have clearly describe the problem, list out related constraints, and able to formulate alternative solutions.
* Creative solution design (10%): Whether you can link your Big-Data solution back to the context of the company case.

Marking Scheme for Presentation (20%)

* Appropriate time allocation and pace (2%): Did your group allocate time appropriately, and mange time effectively, with smooth progression? Did your group use appropriate pace? Did the presentation start punctually?
* Clear, logically organized and relevant content (4%): Was information included always relevant? Were presented points clearly stated and developed? Did the materials flow extremely well? Were the materials well organized? Were there any ambiguities are left unexplained?
* Uses good body language, eye contact, and appropriate voice tone (4%): Did the group show balanced posture, enthusiasm and confidence? Did the group make good eye contact with audience? Did the group use voice tone effectively?
* Gains/holds attention (4%): Did the group provide good motivation to engage the audience’s interest? Did the group present the contents in a manner that captivates the audience’s attention?
* Uses instructor defined role appropriate dress (2%): Was the group dressed as expected by the instructor?
* Clarity of speech/Accuracy of grammar & pronunciation (4%): Was the voice consistently comprehensible? Were grammar and pronunciation accurate?

**Plagiarism will be penalized. Late submission will be subject to a deduction of 1% per day.**