# 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:

## 任务 1：系统开发 （30%）

您的小组将获得一个业务文本数据集，其中包含标准化的数据拆分和评估指标。您的任务是开发一个对文本进行分类的“*性能良好*”的决策支持系统。为此，您需要：

1. 构建基线 模型（即具有基本设置且无需微调的模型，例如 Logistic 回归）
2. 与**一系列不同的数据分析模型**（例如，梯度提升树、决策树等）和**不同的词嵌入模型**（BERT、GloVE、Word2vec 等）进行比较
3. 为上述每个模型微调不同的**超参数**并报告最佳分数
4. **使用一系列不同的数据统计方法进行可视化**。描述数据中模式的方法有很多种，包括图表和数字（例如，条形图、折线图、词云、词分布等），以及统计指标，如集中趋势（平均值、众数和中位数）和离散度（范围、方差、最大值、最小值、四分位数）、相关性和标准差。
5. 探索不同的方法来改进您的基线模型。 将根据以下因素进行评分：***方法的新颖性***（例如，如果类/项目说明中没有涉及这些方法）、***方法的多样性***（例如，开发多种可行的方法）、***实施的难度***以及  ***方法带来的***改进。

在您的提交（即 Jupyter Notebook 文件）中，您需要包含第 1-5 点的代码实现**;** 以及**描述第 1-5 点的简短表格和/或文本（少于 300 字）（**例如，您尝试过哪些模型、超参数、可视化技术和/或它们的分数和详细信息）。

您可以使用在课程中学到的任何可视化和模型改进技术和/或包含课程中未涵盖的其他技术。但是，您必须 **在整个项目中使用相同的数据拆分和评估指标** （如果发现这三个条件在提交中发生了变化，则会扣分）。

## Task 2: Case Study (10%)

Each group will be assigned following the company case:

* Fitbit

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.**