WIA1006/WID3006 MACHINE LEARNING

GROUP ASSIGNMENT (20%)

1) INTRODUCTION

This project is designed for undergraduate students who are taking the WIA1006 OR WID3006 in semester 2, session 2022/2023. The theme of the project is **HEALTHCARE HACKS**. In this project, you will be working in a group of 5 to explore and analyze a specific domain in healthcare using machine learning techniques. Your goal is to develop a machine learning model that can help to solve a real-world problem in healthcare.

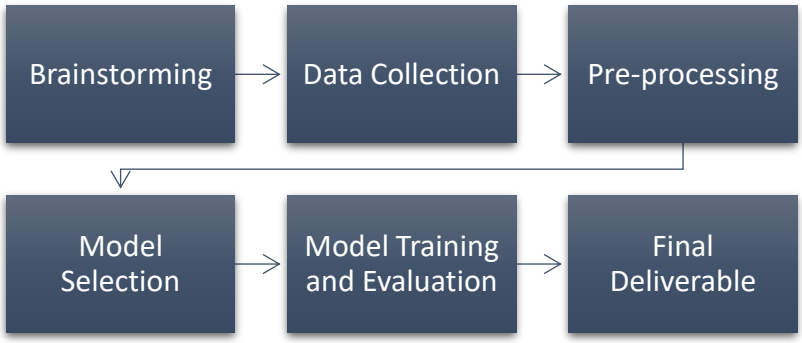
2.0 PROJECT DETAILS

The team is required to develop a machine learning model that could do a specific task such as classification, prediction or recommendation. Below are some of the examples of ML project that could be the kickstart of your innovation idea:

1. Predicting patient who are most likely to be readmitted to the hospital within a certain timeframe.
2. Classify type of cancer (e.g., benign or malignant)
3. Predicting patient outcomes such as mortality or length of hospital stay

You can explore more interesting project online and can take those examples as an inspiration to develop your own model.

The flow of the work could be as the following:



1. Brainstorming: As a group, you should first brainstorm and come up with a specific domain in healthcare that you are interested in exploring. This could be anything from predicting patient outcomes to diagnosing diseases.
2. Data Collection: Once you have decided on a domain, you will need to collect data that is relevant to your project. You can use publicly available datasets or collect data on your own. Ensure that your data is relevant, reliable, and sufficient for your project.
3. Pre-processing: After collecting your data, you will need to preprocess it to make it suitable for machine learning. This may include cleaning, normalization, feature extraction, and transformation.
4. Model Selection: Once your data is ready, you will need to choose an appropriate machine learning algorithm for your project. You may choose from supervised or unsupervised learning, and select the model based on the problem you are trying to solve.
5. Model Training and Evaluation: Once you have selected the model, you will need to train it on your preprocessed data and evaluate its performance using appropriate metrics. You may need to fine-tune your model to improve its performance.
6. Final Deliverable: Finally, you will need to present your findings and results in a clear and concise manner. This may include a project report, code implementation, and a presentation.

3.0 DATASET

Accessing public datasets in the healthcare domain can sometimes be challenging due to privacy concerns and restrictions on data sharing. However, there are several publicly available healthcare datasets that can be used for Machine Learning projects. Here are a few examples:

1) MIMIC-III: The Medical Information Mart for Intensive Care (MIMIC-III) is a large public dataset of de-identified electronic health records from patients admitted to the intensive care unit. It includes information such as vital signs, lab values, medications, and diagnoses.

2) SEER: The Surveillance, Epidemiology, and End Results (SEER) program is a public dataset of cancer incidence and survival data from the National Cancer Institute. It includes information on demographics, cancer site, stage, and treatment.

3) NHANES: The National Health and Nutrition Examination Survey (NHANES) is a public dataset of health and nutrition data from the Centers for Disease Control and Prevention. It includes information on demographics, health behaviors, and clinical measurements.

4) PhysioNet: The PhysioNet project provides access to a variety of publicly available datasets related to physiological signals, such as electrocardiograms (ECG), electroencephalograms (EEG), and blood pressure waveforms.

5) OpenFDA: The OpenFDA project provides access to a variety of datasets related to drugs, medical devices, and adverse events reported to the Food and Drug Administration (FDA).

6) COVID-19 Open Research Dataset (CORD-19): This dataset includes tens of thousands of scholarly articles about COVID-19 and related viruses. It is intended to facilitate research and analysis related to the pandemic.

7) Chest X-Ray Images (Pneumonia): This dataset includes thousands of chest X-ray images labeled as either normal, bacterial pneumonia, or viral pneumonia. It can be used to develop Machine Learning models for diagnosing pneumonia.

8) Skin Cancer MNIST: This dataset includes over 10,000 images of skin lesions labeled as either malignant or benign. It can be used to develop Machine Learning models for detecting skin cancer.

9) MURA (Musculoskeletal Radiographs): This dataset includes over 40,000 musculoskeletal radiographs labeled as either normal or abnormal. It can be used to develop Machine Learning models for diagnosing musculoskeletal conditions.

10) PhysioBank: This repository includes a variety of physiological signal datasets, such as ECGs, EEGs, and blood pressure waveforms. These datasets can be used to develop Machine Learning models for detecting abnormal physiological patterns.

Above are the examples of dataset that potentially can be used for your project. You can find it on GitHub and some on their own website. You can explore and use other dataset than whatever listed above. There is no restriction on exploring new dataset that is suitable for your specific project.

Note that, there are **limited** datasets for the certain problems in healthcare domain due to the privacy issue. Thus, you may need to develop or collect your own datasets for the problems that you have chosen. There is even a possibility of developing a **pseudo-dataset** to demonstrate your proposed problem and solution.

So be creative!

4.0 ASSESSMENT

The assessment of the project will be divided into two (2) sections:

1) TECHNICAL KNOWLEDGE

a) Relevance and Significance of the Problem (10 points)

b) Data Collection and Preprocessing (20 points)

c) Model Selection and Performance (30 points)

2) SOFT SKILL EVALUATION

a) Presentation Quality (20 points)

b) Creativity and Innovation (10 points)

c) Teamwork and Collaboration (10 points)

This assignment shall evaluate soft skills elements:

A. COMMUNICATION SKILLS (CS1, CS2, CS3) *

PRESENTATION (Flow of discussion, Presentation (language/fluency/idea coherency), Teamwork, Effort and Q&A skills).

B. CRITICAL THINKING AND PROBLEM SOLVING (CT1, CT2, CT3)

Creative and critical thinking (Innovation, connecting, synthesizing knowledge and transforming ideas into new forms/solutions) for Individual Project.

C. MORAL AND PROFESSIONAL ETHICS (EM1, EM2)

Moral & Professional Ethics – ethics in presentation of results in online/live presentations.

Total Points: 100 (20%)

Deadline: **10th June 2023 (Saturday) by 11.59 p.m.**

You are required to submit your recorded presentation including the project demo, coding and PowerPoint slides. The duration of the presentation is capped to 10 minutes. Each of the members needs to present their technical part contributing in this project.

We will announce the best 5 groups who will be in the running for cash prize by 17th June 2023. The selected groups will compete on the 20th June 2023 during lecture hour at DK1. The champion will be selected based on voting of the audience.

Good luck and have fun!