Phase01 and Phase02

This project helps you gain hands-on experience of applying artificial intelligence/machine learning techniques on real dataset in an organized way. In the phases 01 and 02, we take the initial steps for creating a model to gain insight into the desired dataset. At the end of this phase, it is expected that you performed the tasks 1 to 12 described in Phase01 and Phase02 sections and document your outcome using the outline provided in pages 6-8. Please make sure your document contains the following sections:

Introduction
- **General Description**
- **Research question**
- **GitHub repository**

Dataset Description
- **URL of your Dataset**
- **Where, when, how the data is collected**
- **The name, definition, and characteristics of features**

Related Work
- **Advantages and disadvantages of 3 related works in compare to each other**
- **Please make sure you cite the GitHub or the source website of the work that you used in your project**

Project Plan

Data Exploration
- **Univariate analysis**
  - **Descriptive analysis**
  - **Distribution analysis**
  - **Outlier detection**
- **Bivariate analysis**
  - **Pair plot**

>>> Each group should have one documentation and one Jupyter Notebook. The head of team should upload them on the private GitHub repository.

>>> Each student must upload his/her team documentation and Jupyter notebook on Blackboard.

Phase01

Selecting your desired group, defining the proper research question, choosing a dataset, and review the related works

1. **Finding a group**

   a. Know each other better

   Please create a video of yourself with the length of 5 Min in Flipgrid platform, https://flipgrid.com/928c8993 . In this video, I expect you provide a short history of yourself including the following items:

   i. Your complete name
   ii. What are your interests?
   iii. What sort of person do you like to work with?
   iv. Do you have any computer programming skill or experience?
   v. How other students can reach you? **(Please write your SHU email on your video using the effect button)**
   vi. Please provide a short history of yourself.
   vii. How will you schedule your work in a team?

   b. Select the head of the team and team members

   It is expected that each team has around **8 members**. You must select a head for your team and a name to manage the communications and distinguish your group from the other groups. In your project progress report, please include the following items:

   i. The team name
   ii. The full name, student number, and SHU email of the head of your team
   iii. The full name, student number, and SHU email of the team members
   iv. One paragraph from each member to provide a short introduction about yourself, describe why did you want to work with the current team members, and how did you select the team head of the team

   c. Create a private GitHub repository

   The head of each team must create a private GitHub repository with the title of "AI _Your Project Title_Your Team Name". The head should invite other members of his/her group in addition to your instructor. You will upload any project-related documentations or their relevant codes on both Blackboard LMS and this GitHub repository. If you are not familiar with creating a GitHub

repository, please follow the instructions provided in the attached videos. Please remember to invite me, https://github.com/RezaSadeghiWSU (Links to an external site.), as a collaborator in your repository. You must get the permission of me before publicly publishing your private GitHub repository. Also, you must get the permission before publicly publishing any assignment answers or questions.

2. **Please select a research question for your AI project. Your research question can follow any AI task; however, regression or classification tasks are preferred.**

3. **Select a database to empirically address your research question. You can go with any database. The following repositories are good option to start with:**

    a. **https://archive.ics.uci.edu/ml/index.php**
    b. **https://keras.io/api/datasets/**
    c. **https://www.tensorflow.org/datasets/catalog/overview#all_datasets**
    d. **https://physionet.org/about/database/**
    e. **https://registry.opendata.aws/**
    f. **https://data.worldbank.org/**
    g. **https://data.cdc.gov/**

4. **Describe when, where, how these data are collected. The number of samples and the number of features. The description of each feature in one or two. This description should include the type of data.**

5. **Google the selected dataset and your research question and report 3 related works from three perspectives of research question, method, results, strength, and weakness. You can write a few paragraphs to compare these works together (not separately.)**

Phase 2

Exploring data

6. **Import** your dataset in Jupyter Notebook and store them in a data frame
7. Explore **the number of samples and features** in your dataset
8. As type of the features effect on your future analysis, please discover the **data type** of your features. E.g., integers, float, string.
9. **Missing values** are the common observations in all datasets. Unfortunately, most AI/ML algorithms could not handle missing values directly. As a result, it is a good idea to address these values in the initial steps. Please explore the number of samples with N/A value in your dataset.
10. It is probable to observe **duplicated samples** in your data set. As AI/ML algorithms work based on the inputs' samples, the duplicated samples can bias our algorithms to the repeated samples instead of learning the patterns among all samples. So, we should be aware of

duplicated samples in our dataset. Please investigate the number of duplicated samples in your dataset.

11. We have several options to handle the samples with missing values or duplicated values. The commonly remove duplicated samples before further analysis. Although we can try the same strategy for missing values by removing the samples that contain any N/A features, we sometimes try to fill the missing values by our best estimation strategy. If we intend to provide an estimation for a value for our predictors, we are using an **imputation** strategy. If we try to provide an estimation for the value of our target feature, we will perform an **interpolation** strategy. You can find several strategies of imputation and interpolation using Scikit-learn in [1] and [2], respectively. For the sake of the simplicity, please remove samples with duplicated values and missing target value. Impute the missing values of predictor features by mean as described in reference [1].
12. We can explore our data from two perspectives of univariate and bivariate analysis. You can familiarize yourself with these analysis via [3].
    a. Univariate analysis is defined as analysis carried out on only one ("uni") variable ("variate") to summarize or describe the variable [4].
        i. Please using describe() method to get the outcomes of **descriptive analysis**, including min, max, Q1, and Q2.
        ii. You can extend your analysis by plotting **histogram** and **box-plot**. These two plots will help you to discover the underlaying pattern among **distribution** of your features. So, you can find weather you are working with normalized distributions or skewed ones. Also, you can find the outlier samples, which are deviated from the rest of the samples. We must remove outliers from our dataset for our further analysis.
    b. Bivariate analysis refers to the analysis of two variables to determine relationships between them [5].
        i. You can perform **Person correlation** and visualize the outcome matrix using a heatmap to explore the linear relationships among your features. Remember we are not interested to use features, which provide the same information for us. You can consider the features with the person values >0.85 and <-0.85 as linearly correlated features.
        ii. You can extend your exploration to **non-linear relationships** among features. Pair plot from seaborn module can visualize such relationships in the presence of target labels.
    c. If You are using image dataset, please follow the analysis provide in [6].

Reference:

1. **https://scikit-learn.org/stable/modules/impute.html**
2. **https://scikit-learn.org/stable/auto_examples/linear_model/plot_polynomial_interpolation.html**
3. **https://saedsayad.com/**
   https://purnasaigudikandula.medium.com/exploratory-data-analysis-beginner-univariate-bivariate-and-multivariate-habberman-dataset-2365264b751

4. **https://link.springer.com/referenceworkentry/10.1007/978-94-007-0753-5_3108**
5. **https://link.springer.com/referenceworkentry/10.1007/978-94-007-0753-5_222**
6. **https://towardsdatascience.com/image-data-analysis-using-python-edddfdf128f4**

Cover Page Sample

Project Name

Artificial Intelligence
Your course section

Your Name

Sacred Heart University
School of Computer Science & Engineering
The Jack Welch College of Business & Technology

Submitted To:
Dr. Reza Sadeghi

Spring 2022

The course Section Code_Project Progress Report_Phase #_Team Name

Title Page Sample

Project Progress Report # of Project Name

Your Name

Your SHU Email
.....@sacredheart.edu

Short Bio

The course Section Code_Project Progress Report_Phase #_Team Name

Table of Content Sample

Table of Contents

Notice:

- Please use the Writing Center facility to automatically get these points. Otherwise, **each typographical or grammatical error will cost -1 points**.

- If you failed to find a group, it is your responsibility to submit a project report alone.

- You can use any material or GitHub repository in this project, you must cite them. For each missing citation, you are subjected to 2 points penalty.

Questions and problem handling:

You can ask any questions regarding the project. You can ask your questions during class, or you can email your questions to your instructor sadeghir@sacredheart.edu.