Introduction to Data Science — Project – Phase 0

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh**

TA(s): Hamid Salemi, Mohammad Araghi

Deadline: Friday, Ordibehesht 7th, 11:59 PM

Introduction

In this project, we aim to have a complete data science project and implement all the principles we have learned theoretically in practice! So we will proceed step by step. In this phase, you should choose the data you want to analyze and investigate. This section constitutes 10% of your project. You must submit the specifications of the data you want to work on in the system by the deadline and wait for the approval of teaching assistants. For this, you need to enter information in two sections.

- In the first section, you must submit your information on the website under the specific section. Teaching assistants will give feedback about your work on the website.
- In the second section, your data information should be entered into this sheet to ensure that other groups are also informed. Please note that teaching assistants will check this sheet regularly, so that if a new group enters its data information, they check the website for the related info. As a result, if you upload your work on the website, but don't update the sheet, your work may not be checked!

Important Notes

1. In this project, the data of each group must be unique. Therefore, if two groups submit similar data, only the group that registers the specifications earlier can use the data. The criterion for registration is the website
2. Checking and confirming the data by teaching assistants may take time and may take several days, so please be patient. If your work isn't checked for more than 3 days, you may notify teaching assistants via Telegram or Email.
3. If your data is not approved by the teaching assistants, you'll have to resubmit the specifications and update the sheet as well. Therefore, don't delay registering the specifications until the last days.
4. You can use any dataset that meets the following conditions, but using unknown datasets will earn you a higher score!

5. **If you perform the data collection or crawling process yourself and do not use an existing dataset, your score for this phase will be doubled! (10% bonus score)**
6. To find datasets, you can use websites such as Kaggle and Hugging Face, or other similar websites.

Required Data Specifications

1. Your data must have at least 5 numerical features and 3 categorical features.
2. The dataset must contain over 2000 samples.
3. In the final phases of the project, you must be able to predict one of the data columns using other data features. So two essential points must be considered:
    - You must consider one variable as the target variable.
    - There must be a correlation and logical relationship between this variable and other variables.

Task

You need to submit the following information in the form of text or a PDF file with this info:

1. Dataset name(if data collection is done by yourself, mention this in the name)
2. Dataset link(if data collection is done by yourself, provide the link of the site you intend to crawl)
3. Target variable name
4. Data sample(sending just five rows is sufficient)

Notes

- Upload your work in this format on the website: DS_Project_P0_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- This phase will not be accepted after its deadline i.e. there will be no late and grace policy!

Good luck!