Project Phase03

You identified your research question and explored the characteristics of your desired dataset as the outcome of the project phases 01 and 02. It is the time to complete the previous phases by providing a prediction strategy for unseen data and provide an educated guess regarding the probable successfulness of our suggested model. To provide a successful predictive model, please follow the steps below:

1. Most AI/ML algorithms accept the input data in a specific form. As a result, preprocessing the input samples shape an important step of working with data. scikit-learn provide an enriched library of preprocessing [1]. Please apply the required transformations on your data. The following steps describe some of the required transformations:
    a. If you want to perform linear regression, your target/dependent feature must be normalized. If it is not normalized, please apply second square root, power, or logarithm functions to **normalized your target feature**. You can use the provide transformations in Scikit-learn, such as power transformation and quantile transformation [2].
    b. It is a good idea to transform your categorical features to numerical ones using **encoding methods**. You can use the methods provided in [3].
    c. AI/ML algorithms are interested to weight more to the features that better separate samples in respect to the target feature. So, it is a common pitfall for these algorithms to tend to provide more weight to features with higher diversity. Also, non-scaled features can disrupt the optimization process. Say, the zig-zag behavior of gradient decent optimization To avoid such behaviors, you can transform features to a **standard scale** using StandardScalar() function from sklearn [4].
    d. Form the previous project phase, you must be able to recognize the high related features. Please remember to just use one of **the related features** in further analysis.
    e. If you are using an image dataset, usually your algorithm requires to use images in the same size [5].

2. To have a fare evaluation, you must **break your data to two parts of train and test**. You can use train_test_split() from sklearn to [6].
3. **Select 3 regression models or classifiers** from sklearn then fit your model and save it in a specific object
4. **Measure the performance** of your model by calling predict method from you created model. Following this pass the predicted values and the real values to your performance measure metric. Say, accuracy for the classification task and MES for regression.
5. Investigate the performance of your modules using different metrics including Accuracy/MAE, Confusion Matrix, Receiver Operating Characteristic plot using roc_curve() method. (Hint: Assignment03 code contains ROC curve)

References:
[1]. https://scikit-learn.org/stable/modules/preprocessing.html
[2].https://scikit-learn.org/stable/modules/preprocessing.html#mapping-to-a-gaussian-distribution
[3]. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
[4]. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
[5]. https://www.tensorflow.org/api_docs/python/tf/image/resize
[6].https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Notice:

>>>>Please complete your previous report by adding the section of Data Modeling. Please report your project progress using the following suggested structure:

Introduction
- **General Description**
- **Research Question**
- **GitHub Repository Address**

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

Related Work
- **Advantages and disadvantages of 3 related works in comparison to each other**

Project Plan

Data Exploration
- **Univariate Analysis**
  - **Descriptive Analysis**
  - **Distribution Analysis**
  - **Outlier Detection**
- **Bivariate Analysis**
  - **Person correlation**
  - **Pair plot**

Data Modeling
- **Preprocessing**
- **Data Splitting**
- **Fitting the model**
- **Measuring Performance**
  - **Accuracy/MAE**
  - **Confusion matrix**
  - **ROC curve**

References

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

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

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

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

Title Page Sample

Project Progress Report # of Project Name

Your Name

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

Short Bio

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**.

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.