CS3121 Introduction to Data Science

Data Pre-processing Lab Exercise

The National Institute of Diabetes and Digestive and Kidney Diseases, United States, has collected a dataset regarding diabetes patients in Pima Woman's Hospital of Arizona. The objective of the dataset is to diagnostically predict whether a patient has diabetes based on specific diagnostic measurements included in the dataset. All the patients recorded in this dataset are females of the Pima Indian heritage. The dataset consists of 8 medical predictor variables and one target variable, Outcome. Table 1 shows the description of each variable.

Patient_ID	Identification number of the patient
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (muU/ml)
BMI	Body mass index (weight in kg / (height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Class variable (0 or 1)

What to do

1. Set up the environment

In this lab we are using Python 3.6 and it is recommended to set up the programming environment using a virtual environment package. Read this article to learn more about virtual environments. This article describes how to instal virtual environments in windows and Linux.

You can also use following two Python libraries.

- Pandas - 0.24
- NumPy - 1.17

You are provided with the `Diabetes_dataset.py` file in which you will add codes for pre-processing the given dataset. The above libraries are imported at the beginning of the Diabetes_dataset.py file.

The code to load the dataset and write the pre-processed dataset into a CSV file are provided. Do not change those codes. Change the value of `'student_id'` variable to be your student identification number.

Write required code to do the following

- Remove duplicates if any.
- Fill missing values (if any) without deleting any record.
- Resolve out-of-range values if any.

Submission

Rename the script as <<Your_Student_Id.py> and submit it to Moodle by **16th April 2023 11.59 p.m.**

This lab will be evaluated automatically. When your submitted Python script is run, it should read the `diabetes.csv` located in the same folder and write the pre-processed dataset as a CSV into the same folder with the name instructed above. This lab will be evaluated based on the generated results. Therefore, prior to submission, run and test your script properly and make sure it is free of bugs.

If you have any questions email *sandarekaw@cse.mrt.ac.lk*.