Group name: Cheetah

Name: Obiri Vincent

Email: vince23.ac@gmail.com

Country: Kenya

College/Company: KCA University

Specialization: Data Science

GitHub repository link:
https://github.com/Vince689/Data_Glacier_Final_Project_Week_7_to_13.git

Problem description & Business understanding:

A pharmaceutical company approached us and asked whether we could automate the process of detecting whether a patient's drug use is consistent (these drugs are prescribed by doctors). As a result, we'll employ analytical approaches to look into the relationships between drug use and the attributes in the dataset. We'll create a classifier that can predict whether or not patients will stick to their treatment regimen. In this clinical environment, we may want to prioritize recall as a model evaluation parameter to account for false-negatives.

Project Life Cycle:

30th January – 5th February:

Describe the business problem. Understand the data, provide an overview of the data and investigate it, and note any problems. Write a report (in PDF format) pertaining to the aforementioned tasks.

6th February – 12th February:

Understanding the dataset, decide on the data to use and identify any problems in the data such as missing values, outliers, skewedness, etc. Determine the approaches to apply to overcome this problem.

13th February – 19th February:

Wrangle / clean the data and ensure it's ready for analysis. Determined the best method for working with missing values. Write a report (in PDF format) pertaining to the aforementioned tasks.

20th February – 26th February:

Perform exploratory data analysis on the dataset, use the work from Week 2 of the internship as inspiration. Make recommendations to help the pharmaceutical company.

27th February – 5th March:

Create a presentation which plainly describes and visualises the work from 20th February – 26th February on EDA for non-technical business users and also present the final recommendations. On the final slide, include the recommended model for this dataset, which will be useful for technical users.

6th March – 12th March:

Explore a classification model from each family in addition to the base model (e.g., gradient boosting, neural network, logistic regression, KNN, etc). Ensure selected model fits business requirements.

12th March – 18th March:

Write a report for the project and also include a PowerPoint presentation.

Data Intake Report

Name: Data Science: Healthcare —Persistency of a drug: Group Project

Report date: 5th February, 2022

Internship Batch: LISUM05

Version:1.0

Data intake by: Obiri Vincent

Data intake reviewer: Obiri Vincent

Data storage location: Local disc

Tabular data details:

Total number of observations	3424
Total number of files	1 (2 excel sheets in 1 file)
Total number of features	69
Base format of the file	.xlsx
Size of the data	898.01KB

Proposed Approach:

- The duplicated() command will be used in Python to identify duplicate entries.
- Assumptions: No assumptions have been made so far, but as we investigate the data we may find that some assumptions are required. This will be mentioned in the notebook and/or other report(s).