

Name: Emre Yesilyurt

Email: yesilyurttemre@gmail.com

Country: Turkey

College/Company: Dokuz Eylul University

Specialization: Data Science

GitHub repository link: https://github.com/emreyesilyurt/file_ingestion/tree/main/week7

Problem description & Business understanding:

A pharmaceutical company approached us and would like us to automate the process of determining whether each patient is persistent with drug use (these drugs are prescribed by doctors). We shall therefore use analytical techniques to investigate the relationships between the drug use and features in the dataset provided. We will build a classifier which should be useful for predicting whether patients are persistent in drug use or not. It may be better to account for false-negatives in this clinical context so we may wish to prioritise recall as a model evaluation metric.

Project Life Cycle:

Step 1:

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.

Step 2:

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

Step 3:

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.

Step 4:

Create a presentation which plainly describes and visualises the work from 3rd August – 6th August 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.

Step 5:

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

Data Intake Report

Name: Data Science: Healthcare — Persistency of a drug

Report date: 16/04/2022

Internship Batch: LISUM07

Version:1.0

Data intake by: Emre Yesilyurt

Data intake reviewer: Emre Yesilyurt

Data storage location:

https://github.com/emreyesilyurt/file_ingestion/blob/main/week7/Healthcare_dataset.xlsx

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).