

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:

Use Python to clean and transform the data in the healthcare (persistence of a drug) dataset so that it will be ready for use in future analyses / modelling.

Transformations / wrangling / cleaning done on the dataset:

Categorical variables were converted to dummy variables with `pd.get_dummies`, this will make it possible for machine learning algorithms to handle the data.

For the `Age_Bucket` feature, the ranges are replaced with 0s (<55) and 1s (55+), this allows us to compare extremes of age (considerably younger against considerably older).

For the `Ptid` feature, the 'P' were replaced for each entry, so this column can be used as an index.

We convert every variable to a float so machine learning algorithms can use the data.

Some algorithms require scaled data, so we scaled the numerical variables. The numerical variables were not normally distributed so `MinMaxScaler` was used.

Each row containing at least one column with an outlier (a value exceeding 3 standard deviations from the mean) was removed, and seemed to make the data more high-quality (from a simple logistic regression model and the accuracy and f1 scores).

The dataframe was split into features (X) and target (y) dataframes / arrays respectively with `df.loc`.

We wanted to reduce the number of features, since we had 115, which was very high. We used `SelectKBest` and with `k=6`. Going beyond 6 features didn't improve the performance in the simple logistic regression model we tested.

The X and y objects were split into training and testing arrays with `train_test_split`.