Intro

Who, why (context, roles, goal, interests, financing)

Overview of document

Methods (Python, Jupyter NTB, etc)

Columns, approach selection

Short description Uveitis (skippable)

Preprocessing

EDA

Visualization

Old preprocessing

Why, how, when

What we decided (formal)

New preprocessing

Linear integration

What we took

What we scraped/rewrote

Modelling

Targets

Testing

Automation

Results

Better approach/alternatives

Possible next steps

Conclusion



EDA

Exploratory data analysis

Missing values control

All columns

Column analysis  
ID  
Gender  
Race  
LOC  
CAT  
other  
ehr\_diagnosis  
specific\_diagnosis  
notes  
AC Abn Od Cells and AC Abn Os Cells  
VIT: AC Abn Od Cells and AC Abn Os Cells  
etc

Ranges of calcium, lactate\_dehydrogenase  
PCA analysis

The scope of exploratory data analysis was to evaluate and properly prepare the data for further elaboration while highlighting primary/principal insights.

The whole dataset was taken into consideration. Ascertaining and communicating a missing values strategy is paramount to ensure reliability, reproducibility and must be kept in consideration while analysing final results. For this, an overview of missing information was created [1] to allow to establish, during pre-processing, a satisfactory missing values approach.

Observations indicate that columns “\_others” and “notes” contain 79.07% missing values. Other columns have a similar issue; “anti-dnase\_b” is composed of 99.63% of missing values. Features “beta-2-microglobulin” and “lupus\_anticoagulant” contain approximately 65% missing values. This underlines the need for a highly flexible missing values strategy that is not limited to only imputing missing values but also to selectively remove features that score above a determined missing value percentage.

Next steps include controlling for data inconsistencies. Edge cases were found in the UOM columns, prompting an accurate evaluation and appropriate response during pre-processing. Then came formatting errors, where extensive work has to be invested to adapt non-standard missing values to machine readable information. Possible optimizations included collapsing variables. This includes the extreme where the target is strictly binary and less drastic measures, i.e., by removing or collapsing, low count occurrences in the “specific\_diagnosis” column. Totally removing features like “ehr\_diagnosis” and “notes” are also available options to be considered. These features are considered non-essentials.