Prediction of location and type of inflammation for patients with uveitis based on blood values and laboratory tests.

Studer Roman   
BSc Data Science Student, FHNWBrugg-Windisch, Switzerland  
roman.studer1@students.fhnw.ch

Rau Alexandre  
BSc Data Science Student, FHNWBrugg-Windisch, Switzerland  
alexandre.rau@students.fhnw.ch

*Abstract*—

Keywords—component, formatting, style, styling, insert (key words)

# Introduction

## Uveitis

Uveitis is a term which describes an inflammation of the uvea. Uveitis can be divided into anterior (lat. In front), posterior (lat. In back), intermediate or panuveitis (more than one segment affected). For example, an anterior uveitis involves the iris [1]. Panuveitis is an inflammation of the whole uvea tract as well as the retina and the vitreous humor (glass body) [2]. Uveitis can lead to the loss of eyesight among other things.

## Project Description

The aim of the project was to identify important features for the diagnosis of uveitis. For this purpose, a dataset with information on more than 1000 patients, collected by Dr. H. Nida Sen et al. from the National Eye Institute, Washington DC, was made available. After an initial exploratory data analysis, a preprocessing pipeline was developed, which can be used together with machine learning algorithms from sklearn, a Python machine learning library. Various algorithms were employed to classify the dataset. The results, especially the feature importance’s, were recorded and documented. Three features were identified as target variables:

Location: This marker describes the location of inflammation with the categories Anterior, Posterior, Intermediate, Panuveitis and Scleritis. The category "Scleritis" refers to inflammation of episcleral and scleral tissue [3]. Prediction of the location based on laboratory values allows the identification of a subset of laboratory tests (via feature importance) that are suitable for prediction. This would allow a small subset of tests to be used for diagnosis based on the location of the inflammation. In addition to faster diagnosis, the reduced number of tests required would lead to a reduction in costs.

A second target feature is “Category”. This marker describes the origin of the ocular inflammation. This can be, for example, systematic, infectious, or idiopathic. This feature is based on the results of laboratory tests and has been recorded retrospectively. Predicting the category, aka the origin of the inflammation, can aid the diagnosis further.

The third target feature is the specific diagnosis itself. In the dataset, 27 different diagnoses were recorded (some can be collapsed based on similarity). The direct prediction of a diagnosis based on laboratory tests could support the medical staff in their decision making.

## Data description

We received a total of 1075 samples from patients affected by certain types of ocular inflammatory diseases. Mostly subtypes of uveitis such as pars planitis but also other diseases that have inflammation in the eye as a symptom or consequence, e.g., white dot syndrome or sarcoidosis. We count 426 male patients and 649 female patients. The difference between male and female patients can be explained as women are disproportionately affected by ocular inflammation [4]. Each sample is described by a total of 64 attributes. The attributes can be divided into laboratory tests (blood values), meta-information about the patient (such as gender or race), and features describing the diagnosis. For the purpose of the analysis, the binary feature "uveitis" was introduced which determines whether the patient has a form of uveitis based on the specific diagnosis.

Information about the patient includes “Subject ID”, “Gender” and “Race”. The location of the inflammation is described in the marker “location” and in “AC Abn Od Cells”, “AC Abn Os Cells”, “Vit Abn Od Cells”, “Vit Abn Os Cells”, “Vit Abn Od Haze”, and “Vit Abn Os Haze”. These qualitative, ordinal features describe the severity of the inflammation of the Anterior Chamber Cells (AC) in either the left eye (OS) or the right eye (OD). The inflammation can be rated as 0, +0.5, +1, +2, +3, +4. The higher the value, the more severe the inflammation is. If a value of +0.5 or higher is present a patient can be considered as "Active", else as "Quiet". The diagnosis is described in the features “categorical”, “EHR Diagnosis” and “specific diagnosis”. The laboratory tests provide a variety of results (mostly blood values) and include: "Calcium”, "Lactate Dehydrogenase", "C-Reactive Protein, Normal and High Sensitivity”, “WWBC”, “RBC ”, “Hemoglobin”, “Hematocrit”, “MCV”, “MCH”, “MCHC”, “RDW”, “Platelet Count”, “Neutrophil %”, “Lymphocytes %”, “Angiotensin Conv#Enzyme”, “Beta-2-Microglobulin”, “Lupus Anticoagulant”, “Lysozyme (Plasma)”, “Anti-CCP Ab”, “Anti-Dnase B”, “Anti-ENA Screen”, “Antinuclear Antibody (ANA)”, “Complement C3”, “Complement C4”, “DNA Double-Stranded Ab”, “HLA-A\*”, “HLA\_A\_1”, “HLA\_A\_2”, “HLA-B\*”, “HLA\_B\_1”, “HLA\_B\_2”, “HLA-Cw\*”, “HLA\_C\_1”, “HLA\_C\_2”, “HLA-DRB1\*”, “HLA\_DRB1\_1”, “HLA\_DRB1\_2”, “HLA-DQB1\*/DQ\*” , “HLA\_DQ\_1”. “HLA\_DQ\_2”, “HLA-DRB\_\*”, “HLA\_DRB\*\_1”, “HLA\_DRB\*\_2”, “Myeloperoxidase Ab”, “Proteinase-3 Antibodies”, “Rheumatoid Factor”, “HBc (HepB core) Ab”, “HBs (HepB surface) Ag”, and “HCV (HepC) Ab”. Features containing the substring “HLA”, which stands for “Human Leukocyte Antigen” represent different haplotypes.

# Exploatory data 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.

# Preprocessing

# Modelling

## Location

## Category

## Specific Diagnosis

## 

# Results

##### Discussion

##### Conclusion

##### Acknowledgment

##### References