***Move this to OUR specific dataset presentation (methods: dataset):***

*One of the possible labels to detect is the patients being malaria positive or negative. There are many missing values in the dataset and only a subset of features make sense to be included in the detection of malaria. One example is that of the geographical location of the clinic visit. In theory, this information would be greatly helpful to detect cases of disease spreading. However, the platform for now does not seek to detect evolution of diseases over time. Moreover, the number of missing values is too high to be able to recover a sub-dataset that is large enough for network training.*

\chapter{Methods}

\label{chap:methods}

\textbf{so can include here: the grid search values, the train tes tsplit, the features dropped, etc....}

% first thing: study design: what i'm trying to achieve, what the dataset looks like, what we are looking for

% this is ec.... design

% what is the overall approach: both data exploration and platform design

% should tell people: what did you do

% explore platform to understand pros and cons of interpretability, then I tested it out on data

% dataset can then be at the end of methods

This section is separated into two parts which correspond to the two axes of the projects. Each axis is standalone and can be separated from the others, although they interconnect in various ways. The two parts of the projects intertwine: the first part sets up tools and interpretability methods for ML models that are already put in place, the second part establishes a baseline for the results of these models in an effort to validate their use and provide a baseline for other projects on similar dataset. The dataset used in both parts is the same and will be presented now.

%=================================================================================================%%=================================================================================================%%=================================================================================================%%=================================================================================================%%=================================================================================================%

\subsection{Data Set}

% should include: level of healthcare, (hospital / clnic etc) and the patient departments, the people: tanzanians, pediatric (<5, range of age)

% 3192 patients recruited in 9 out patient departments in dar es salaam, patient were pediatrics 0 to 5 years, the recruitment period is

% 3 hospitals

% inclusion criteria age /

% december 2014 february 2016

\subsection{Data set}

The dataset used is an e-POCT collection dataset that collects data from children along a certain timeline.

The dataset we are using is a collection of patient information from different clinics. The different information are categorized as different features (organized per columns), and the different patients are organized per row.

The concatenation and collection of data was done on the 5th of December 2018.

The label is binary, indicating presence (1) or absence (0) of malaria for a given patient/row, making this a classification problem. The features are varied, some categorical, some numerical.

The data we are using is tabular. We have 166 total features and 3192 total data instances (rows). Some of the features are missing, at various degrees, for various data points.

We will select a subset of the available features in an effort to make the explanation easier to understand.

\subsection{Data Pre-Processing}

In order to pre-process the data, an investigation will be led on how much of the data is and can be used in the model training and evaluation tasks.

\subsubsection{Feature Selection}

\paragraph{Missing Values}

Some features contain too many missing values to be interesting. These features will be dropped. The remainder of the features will have their missing values substituted by the mean of the feature.

Features with more than 10\% of missing values are instantly discarded. This is a rather large cutoff point, however since the goal of the algorithm is to detect a patient's specificities (the specificities of the diagnosis) there is reason to have as little null values as possible so as to impute as little as possible.

\paragraph{Class imbalance}

Some of the features only have one type of value within the dataset, or have distributions that are exactly similar in both training and testing partitions.

This is the case for the feature detailing signs of respiratory distress, which contains only a total of 3 values that are not 0 in cases of malaria positivity (which is a total of 320 (ratio of 0.009375). These features could be kept, as they could improve the classifier's decision in marginal cases.

\begin{center}

\includegraphics[width=0.9\linewidth]{images/resp\_distress.png}

\end{center}

\paragraph{Selecting features of interest}

Some of the features within the dataset do not relate to the overall goal of the platform as of now. For example, even though the geographical location of the patient should in theory be important to predictions, this feature will not be taken into consideration. The reason for this is that while the geographical location of patients is useful in detecting an epidemic, with time this information will change. Since there is no have information on which data points were collected at which time, this feature will not be used.

Some other features are similar to the label we will be trying to predict. For example, there are features that pertain to the results of a laboratory diagnostic test for malaria - if the label is the diagnosis of malaria then this feature should not be considered. In addition to investigating the feature origins, a correlation coefficient should be computed in order to make sure that in the dataset used the data behaves as expected and that no paramount information is being discarded.. The goal of the task is not related to evaluating the efficiency of the test therefore the feature will be dropped.

Features dropped:

\begin{itemize}

\item lab tests that are not malaria related (contain "lab or dx" but also "malaria") in their name and are not the label.

\item chief complaints because it is too hard to evaluate whether a patient's indications are enough

\item demographic ward locations - even though they could help, the model is to be used only based on symptoms not based on different populations; for the same reason "dem\\_district" containing features are also dropped

\item history based : "hist" which are records of the patients' history. same as before, a patient's diagnosis should be able to be computed without prior knowledge of their state

\end{itemize}

\paragraph{Correlated Features}

With the remaining features, Pearson correlation coefficients will be investigated and features that are highly correlated to each other will be dropped. The correlations might arise from the fact that some features are measuring the same thing in different ways.

\begin{center}

\includegraphics[width=0.9\linewidth]{images/correlation.png}

\end{center}

From these features, a threshold of an absolute value of 0.5 can be set and features that are higher than this threshold should be investigated. This method yields 6 pairs of features with high correlations between 9 different feature combinations. The choice is made individually upon further investigation of the feature meaning. An example of this research is that two features that are highly correlated are MUAC and signv\\_acz. The MUAC (mid upper arm circumference) is relative to a person's health (more specifically it is used as an indicator of malnurishment) while the ACZ score is the z-score of that measure. \textbf{[https://www.msf.org/muac-measure-and-definition]}. This ensures investigation allows to backup the manual feature investigation done previously and make sure that the values that are thought to be similar measures are indeed highly correlated.

\paragraph{Boruta Selection of Features}

A Boruta algorithm can also be used in order to determine which features are important relative to the prediction. There are two corrections implemented in the package used:

The features are rejected if the corrected p-values are rejected at an alpha confidence level of 0.05.

\textbf{include list of selected features here}

\paragraph{Selected Features}

The features selected after this process were ... \textbf{and also add in the total number + how many data points are left after removing nulls}

\subsubsection{Data Transformation}

\paragraph{Feature Transforms} Some features are categorical, and some categorical features are non-binary. Therefore one-hot encoding and label-encoding will be performed. For numerical features, a min-max scaler will be used.

\paragraph{Data imputation}

Before imputation, any data point for which the label is missing will be removed: meaning that any patient with data `lab\\_malaria\\_any\\_d0` missing needs to be discarded from the dataset.

Missing data-points can be imputed feature-wise, with various methods. However these methods are robust only if there are enough data points to evaluate what would be a good substitute for missing values. In this case, missing values are substituted by the median of that feature. Since any feature with "too many" missing values have been removed, the median should be a "good enough" intuitive imputation of missing data. However, it is not perfect, and more likely it is best to use intelligent imputation methods. Such methods notably allow data to be clustered in an unsupervised way, and imputation to be done on these clusters separately, so as to give a more tailored imputation per cluster.

\paragraph{Class imbalance}

The ratio of positive to negative cases in the label class is imbalanced: 8.375 of negative over positive (2680 to 320 respectively).

Over sampling will be used, Since the dataset is small, so as to not reduce the overall number of data points.

There are various re-sampling techniques, the most straight-forward one being random sampling.

For over sampling,a random over-sampling technique will be used: a data point randomly select a data point from the available data points (within the "smaller" class), and add it as a new sample to the overall dataset (with replacement). This means that there will be duplicate data points. This technique obviously has some drawbacks, the first of which is that duplicating data points biases the way the model works.

To deal with this issue, synthetic samples can be created, such as in SMOTE technique.

There are various other sampling techniques, including some that allow to introduce noise / perturbations in the newly added data points (so that it becomes highly unlikely that there are duplicate data points), and others which combine both under and over sampling. Some techniques are based on reducing the data points to a low-dimensional space, and then removing those data points which are close in this new dimension space but which have different labels (which amounts to trying to reduce the complexity of the classification task of the ML model). Each sampling techniques has advantages and drawbacks, and it is important to understand exactly what these are in order to be able to explain any abnormal behavior in our machine learning model.

The method first used is oversampling to obtain a ratio of minority/majority class of 0.3. This yields a total number of samples of 3484.

SMOTE will be used next in order to see if it can yield better results.

\paragraph{Train-Test Split}

The train-test split is done with a 0.3 ratio on the pre-processed dataset.