

A Case of Study for Classifications Algorithm to a Tabular Vector Borne Disease Dataset

Alberto Arath Figueroa Salomon

December 11, 2024

1 Introduction

Vector-borne diseases have become relevant because human activity creates imbalance in ecosystems, and this family of diseases has become a recurring problem in warm climates. These diseases have a common set features, in addition to the fact that they are transmitted bites of a blood-sucking arthropod.

Such common features are the prognosis, where the symptoms between each other could be almost indistinguishable

The purpose of the study is to build a machine learning model that can discern classification patterns for each set of shared symptoms.

2 Data

2.1 Instance composition

Data set is composed by labeled data, 707 rows and 64 features. Training data has a synthetic origin so there is really no need for data preparation which is good as we can focus solely in the application of Machine Learning techniques

Muscle Pain	Fatigue	Weakness	Prognosis
Present	Not Present	Present	Present
Not Present	Not Present	Present	Present

Table 1: Sample training set with a subset of features.

2.2 Predictor Set

Is composed of binary variables specifying if given symptom is present in this prognosis instance it represents the presence or absence of a specific condition, The variable X takes values from the set:

$$X \in \{0, 1\},$$

where:

$$X = \begin{cases} 1 & \text{if the condition is \textbf{present},} \\ 0 & \text{if the condition is \textbf{not present}.} \end{cases}$$

2.3 Target Variable

The target variable is called Prognosis containing the Vector-borne disease classifier

The target variable, denoted as Y , is a categorical variable representing the prognosis outcome it can take values from the following finite set of classes:

$$Y \in \{C_1, C_2, \dots, C_k\},$$

where C_i represents the i^{th} class label, for $i = 1, 2, \dots, k$.

In this study, $k = 11$, and the classes are defined as follows:

- C_1 : Dengue,
- C_2 : Zika,
- C_3 : Malaria.

To allow the application of machine learning algorithms label encoder variables as in following example

$$Y = \begin{cases} 0 & \text{if the class is } C_1 \text{ (Malaria),} \\ 1 & \text{if the class is } C_2 \text{ (Dengue),} \\ 2 & \text{if the class is } C_3 \text{ (Zika).} \end{cases}$$

This encoding To allows the application of machine learning algorithms such as softmax regression, decision trees, or neural networks to predict the likelihood of each class based on the input features.

3 Exploratory Data Analysis (EDA)

3.1 Class distribution

Class distribution shows there is no significant class imbalance, Figure 1 show West Nile Fever with 80 instances and Malaria falls short with 50 as part of study, class imbalance could have an impact but we expect not to be significant. Just for the sake of the study we will apply oversampling techniques to improve model.

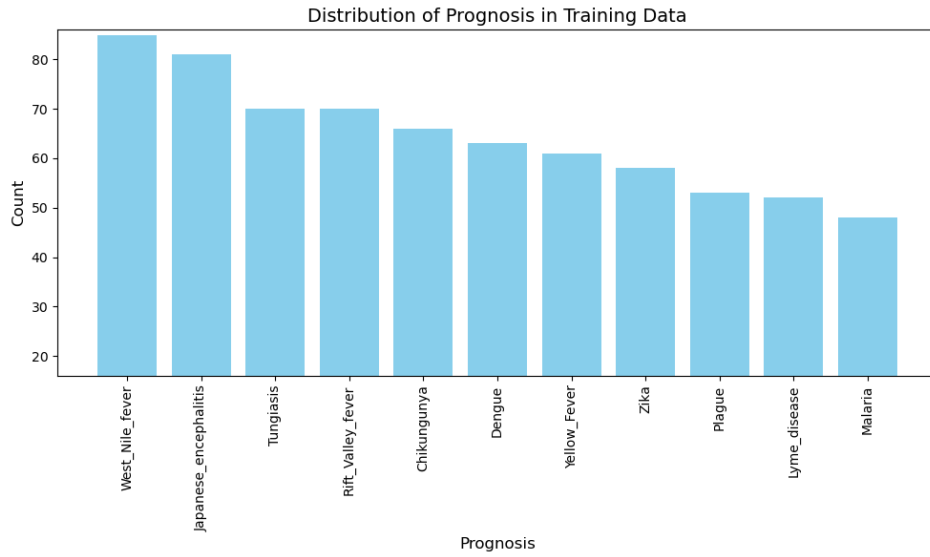


Figure 1: Distribution of Prognosis in Training Data.

3.2 Features correaltion matrix

Correlation matrix shows there is little information between features. So dimensionality reduction techniques are not likely to be effective, in Model section we will place this hypothesis to test.

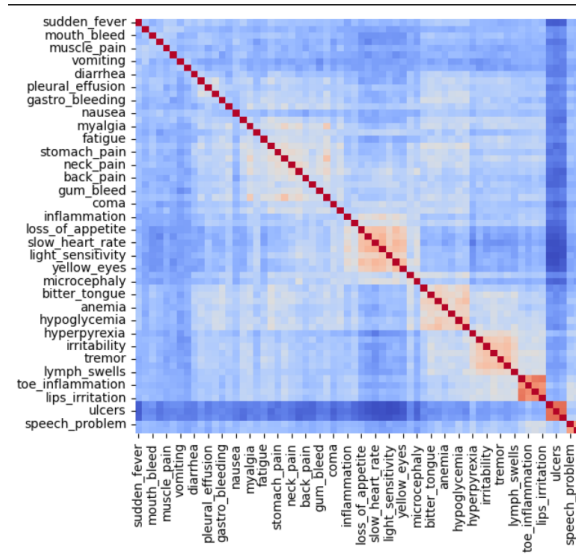


Figure 2: Features Correlation Heat Map. Blue areas indicate low correlation.

4 Methodology

4.1 Grid Search approach to find best cross validation results

4.2 Oversampling

4.3 Dimenstionality Reduction

5 Results

5.1 Score and prediction

5.2 Confusion Matrix

6 Conclusions

References