Théo HENAFF December 14, 2020

Problem definition of the software project

I. Description of the project (what will be solved)

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases [1].

The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. All patients here are females at least 21 years old of Pima Indian heritage.

Therefore, I will try to determine the most effective way to classify diabetes's positive and negative patients.

II. Problem specification (input data, output results)

The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. There are 768 rows with 8 features each.

The output data is whether it is positive or not to diabetes. It is a binary value.

III. Specification of the ML task

A. Performance measurement

Because it is a classification problem the performance measurement is done by comparing in a test dataset the "true" value and the one predicted.
In this dataset there is not the same number of Outcome positive and negative rows. Positive diabetic patient represents only 35% of the whole dataset. Therefore, we cannot use the accuracy as the performance index because we would have very high scores which are totally meaningless.
My solution is to use the f1_score which take both accuracy and recall being calculated.

B. Experience provided

The training experience is done by reading patients medical information's and it is given diagnosis in a training dataset. Thus, the experience is a record of the dataset.

[1] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.