CE-802-7-SP-Machine Learning

Pilot Study Proposal

Submitted by,

Karan Bhatt

(2112102)

Submitted to,

Dr. Vito de feo

Abstract:

Diabetes Mellitus is a disorder that is characterized by multiple organ failure caused by uncontrolled diabetes. Diabetes may be discovered and cured sooner and with more precisely than ever before. With the integration of machine learning, we will achieve a higher milestone which will let the doctor to generate an accurate report wherein the person would be spared of any concern and assault by diabetes. The purpose of this research is to build a program that can predict diabetes in a person depending on multiple variables, allowing for higher accuracy, precision, and recall in predicted results.

Introduction:

The major purpose is to explore the disease's roots and increasing number of diabetic sufferers. To solve this, we will develop a model to assist with prediction tasks. Diabetes Mellitus is one such condition [2].

We like to apply the models outlined above in our project because our problem is classification. The following essential features should be present to achieve

optimum accuracy: glucose levels, right and left systolic blood pressure, BMI, heart pulse rate, insulin pressure, breathing, food type consumed and weight.

Methodology (Classification)

We will leverage that understanding by applying those algorithms which will concentrate on the supervised learning aspect, especially classification, to check whether the individual would have to suffer from the disease or not by using classification techniques of machine learning. We will cross check our recommended technique after testing with numerous methods for enhancing precision, as well as performance and accuracy. In This project I used five disinterment techniques. The purpose is to forecast if the patient is diabetic or not based on the measured data. The algorithms employed include K nearest

neighbour, Random Forest, Logistic Regression, Decision Tree, and Support Vector Machine. To predict diabetes, the model with the best accuracy score is selected [2].

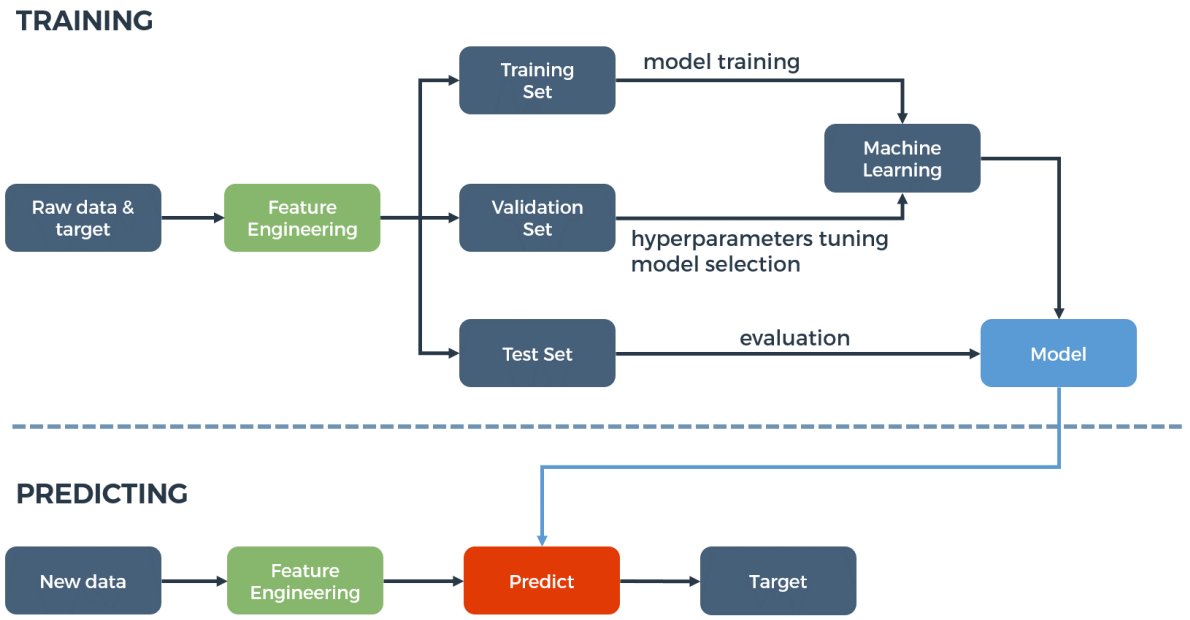

Figure 1 Indicates the Machine Learning Process

Model Evaluation:

We separate our data into two groups to attain the best level of prediction: validation and training. Our data must be divided into three types: training, testing and development. In our scenario, we shall additionally isolate certain data for cross-referencing. There are four possible results when doing classification predictions.

A true positive TP occurs when an observation is predicted. A true negative occurs when an observation which does not correspond to a given class is predicted. False positives happen if a feature is expected to correspond to a class when it actually doesn't. When a feature doesn't really correspond to a class, it generates false negatives. but does after prediction. When analysing classifier

models, the dataset is separated into two sets, and the decision boundaries and observations are shown. This can help in double-checking the rate error acquired after plotting a confusion matrix to examine the true-false negatives and positives. Furthermore, the use of a ROC curve, often called as the area under the curve, would help in determining accuracy. So, it's concluded that model would be focused on two types, Precision and recall [1].

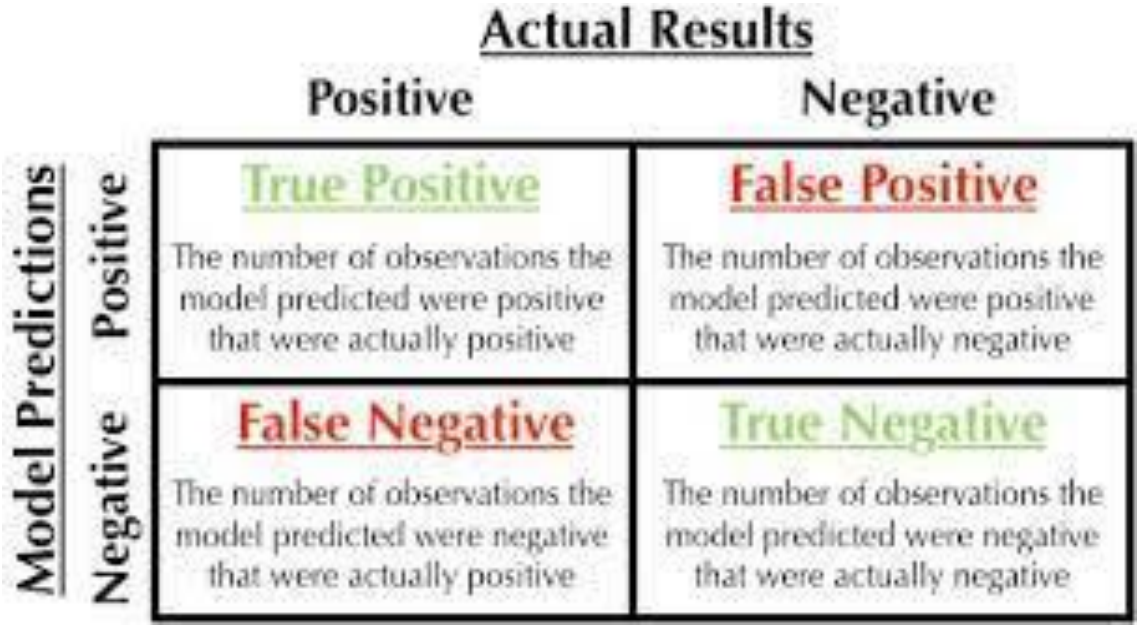

Figure 2 indicates the Confusion Matrix

References:

[1] J. a. M. G. Davis, "The relationship between Precision-Recall and ROC curves," *The relationship between Precision-Recall and ROC curves,* pp. 233-240, 2006.

[2] M. K. e. a. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers.," *IEEE Access,* pp. 76516-76531, 2020.

[3] N. a. T. G. Sneha, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big data,* vol. 6, no. 1, pp. 1-19, 2019.