Design and Application of a Machine Learning System for a Practical Problem

Pilot Study Proposal

(Word count: 710)

MAY 3 2023

CE802 Machine Learning
Submitted by: Lokesh Palaniraj (2211543)

University of Essex

Overview

Risk Of Developing Diabetes

The purpose of this pilot study is to identify the patients who are at high risk of developing diabetes using machine learning procedures. Early identification of patients who might worsen the condition of diabetes enables healthcare providers to offer appropriate support and guidance proactively, to prevent complications. We might be considering multiple factors of the patients to predict the risk of diabetes. This report will provide an overall insight into healthcare providers and educate them more on the patients' current health form.

Type of Prediction

As this problem involves the identification of patients at high or low risk of developing diabetes, this is a "binary classification task", which involves many factors affecting patients' health condition.

It's a supervised technique in Machine Learning, where the prediction lies between two classes based on the instance of data, which can be 1 or 0.

Possible Informative Features

There are multiple features that might predict the risk of developing diabetes in patients; they are as follows:

- Age
- BMI
- Family's Medical History
- Blood Pressure
- Cholesterol
- Daily life routine, including the patient's smoking & drinking status, diet, and exercise.

Learning Procedures

The objective is to classify the input features to one of the output classes which is high or low risk of developing diabetes in the future. There are many numbers of learning procedures for binary classification tasks, they are:

Decision Tree Classifier: Decision Tress are perfect choice for the binary classification tasks as they are so easy to interpret and evaluate. The model will still perform very well even with less effort on the data cleaning and preprocessing. It can even handle both continuous and categorical features and still give out the best performance in predicting the target variable.

k-Nearest Neighbors: kNN is also another good choice for the binary classification task, but the performance depends upon the quality of the dataset provided. On the other hand, the model is so powerful in making predictions irrespective of the size of the dataset. In our case, it will perform well as we don't have any parameters and it'll predict the target using the nearest neighbors.

Logistic Regression: This model is a supervised learning technique which predicts the target (risk of diabetes or not) with the features (age, BMI, blood pressure) provided, using a logit function. This is an easy model to understand and just like DT, it can handle both types of features (categorical & continuous).

Support Vector Machines: Support Vector Machine is another significant machine learning model for classifying the classes by maximizing the margin between the two prediction classes. This model would be a better suit as we would be having multiple features of patients in higher amount to consider, which might be non-linear and high dimensional as well.

Naïve Bayes: This model, Naïve bayes is one of the best for filtering spam, but it can also be effective in our binary classification model with high dimensional medical records. This model is comparatively fast and simple to perform.

Evaluation of the System

There are many ways to evaluate the models before deploying it. We can train the dataset to a different ML model to predict the one which we are going to deploy, the process is known as cross-validation, in which the dataset would be split into training and testing and the accuracy will be evaluated by predicting the target. By performing the cross-validation, we can make sure that it'll perform well with the new data too. We can also investigate different metrics like Accuracy, F1_score, Precision, Recall, etc., to assess the performance of the models by deploying a confusion matrix, which is a good evaluation model with the tabular summary of true positives/negatives and false positives/negatives.

Conclusion

By culminating, we can assure that the machine learning models can be used efficiently for predicting patients at higher risk of developing diabetes, provided the models with the required features and the evaluation methods to assess the model's accuracy of prediction. These machine learning models would be very beneficial for the health care providers to identify the patients' diabetes status prior and act accordingly to avoid issues with the patients' health condition. For the healthcare industry to be proactive, these models can also be replicated for a variety of other diseases.