Pilot Study Proposal: Design and Application of a Machine Learning System for a Practical Problem

Rahul Kithalamane Basavaraj - 2212064
Word Count - 737

May 2, 2023

1 Introduction

In this pilot study proposal, we are identifying patients who are at high risk to diabetes using machine learning techniques. The objective here is to identify patients at high risk and provide healthcare who can proactively intervene with preventive measures such as lifestyle interventions, medication, or referral to a specialist. Further in this study we identify the type of predictive task that must be performed, possible informative features, procedure to choose and evaluation of the system.

2 Type of Predictive Task

Here in this study, we are doing the predictive task using classification algorithms where the target column is predicted for the given set using the historical data. The historical data extracted from the medical records have been given certain labels to indicate whether the patient has been diagnosed with diabetes or not and to predict this output using machine learning algorithms we classify them into 2 groups, those at high risk and those without the risk of getting diabetes.

3 Informative Features

There are several features which is having information to predict the developing diabetes in patients. Some of the general lifestyle conditions

and clinical measurements gives rise to a risk in diabetes, and these features include the age, gender, exercise, family history, high blood pressure (HBP) , high waist measurement, overweight, depression, bipolar disorder, schizophrenia, alcohol, sedentary lifestyle, disturbed sleep, Body Mass Index (BMI) , cholesterol, and sugar levels can be used as an Informative feature.

4 Learning Procedures

Some of the machine learning procedures which can be considered for the given data are as follows:

1) Decision Tree algorithms (DTs) is a very popular algorithm to conduct the classification study. Some of the important and the common features they can handle in this medical dataset are the categorical, noisy, continuous, and missing variables which are very commonly found in this dataset. The decision tree algorithm can be a very good tool to interpret and visualize the data.

2) Support Vector Machines (SVM) are very good if the data contains lot of noise and the values are very huge. Even these algorithms can handle categorical and missing values in the data set. The performance of this algorithm improves the when the feature is selected and the data is pre-processed with techniques such as normalization and scaling.

3) Random forest classifier (RFC's) is a very powerful algorithm in dealing with data that contains lots of noise and are having high dimensional feature spaces. This classifier does not require normalization of data as it is a rules-based approach. The performance of this algorithm can be further improved by data pre-processing and hyper parameter tuning.

4) K-Nearest Neighbors (KNN) classifier algorithm is very effective if the datasets are very small. These algorithms can also handle categorical and missing values in the dataset but are very sensitive to noise in the data. To reduce this, we can perform various techniques like feature selection, dimensionality reduction and hyperparameter tuning. This algorithm is useful for classification problems.

5 Evaluation Method

The performance of the system can be evaluated before the deployment

by using cross validation where the data is divided into k equal sized folds 80% and 20% respectively. Also, if necessary hyper parameter tuning is also done. It is trained on the 1st fold and tested on the remaining fold. This process is repeated k times to overcome the problem of over fitting. We can also evaluate by measuring the accuracy of the model. The other method we could use is by creating a confusion matrix table which provides information on true positive, false positive, true negative and false negative numbers, which can be used to evaluate metrics such as precision, accuracy, recall and F1-score.

6 Conclusion

In conclusion this pilot study shows how to identify patients with diabetes based on the existing data set. We also have proposed classification tasks which can be performed, informative feature which can be identified in the data set and the evaluation method which can be used for this study. We recommend using classification models which can handle both categorical and continuous features and study the performance of a model using cross-validation technique with precision, accuracy, recall and F1-score as the evaluation metric. With this study we try to predict patients who are at high risk to diabetes and help prevent them from further sufferings in future by providing them proper healthcare.

References

[1] Scikit-learn Machine Learning in Python
`https://scikit-learn.org/stable/index.html`