University of Essex

CE802 Machine Learning

Name: Pinky Ramnath Mehta

Registration Number: 2211433

Date: 01/05/2023

Department: School of Computer Science and Electronic Engineering

Word count excluding Reference: 747

Introduction:

Diabetes is a fast-growing disease that affects millions of people worldwide and is rapidly increasing at alarming rate. Detecting the risk of diabetes at early stage reduces risk of complications and can improve patient outcomes. One way of predictions is by using Machine Learning Algorithms with the help previous medical records of patients. Finding a pattern and relationship which is difficult to detect by humans.

The report proposes a pilot study of investigating the usefulness of ML for

1. Patients who can have a high risk of developing diabetes with the help of certain Electronic Medical Records that is provided.
2. Patients whose blood glucose level exceeds diagnostic threshold if untreated.

Predictive Task:

The aim of this pilot study is to investigate the risk of developing diabetes and the predictive task for the study is Binary Classification. A prediction with only two outcomes whether or not the patient is positive towards developing the disease. The pilot study uses the electronic medical records of past patients and with the we will train and test the Machine Learning Algorithms on a labelled dataset.

Informative Features:

In order to develop an effective ML algorithm, identification of informative features needs to be extracted from the provide EMRs. The features that could have an effective impact include, BMI (Body Mass Index), age, gender, Blood Pressure and Cholesterol levels, family history and majorly lifestyle factors such as physical activities and status of smoking. With help of these records, the predictions can be made by doing laboratory tests which can also be included in analysis.

Learning Procedures and Reasons to use:

Various learning procedures can be made use of for our task. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-NN are some of the commonly used ML algorithms for a binary classification task. These models have been used in majority of healthcare data analyzation along with their performance validation in vast number of studies. References (Alghamdi & AlMallah, 2017; Kavakiotis et al., 2017).

The learning procedures used in study includes Logistic regression (LR), Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM).

LR is considered a popular ML method for binary classification as is simple and interpretable model which can handle both categorical and continuous input features.

DT is similar to LR but can also capture the non-linear relationships between features.

RF is the most flexible algorithm which handles both classification and regression task. It combines multiple DTs which will improve the accuracy rate and reduces overfitting. Advantage is that it can tackle with high-dimensional data and can identify important features for a classification model.

SVM can also handle high-dimensional data but it is best suited for problems with small sample sizes.

Performance Evaluation:

To evaluate the performances of the selected ML algorithms, we will first visualize the dataset by perform Exploratory Data Analysis that includes figuring out whether or not the dataset consist of null values, is it equally distributed, datatypes of the dataset, plotting a heat map to know the correlation between attributes.

Once EDA is completed, randomly split the dataset into training and testing sets as 80:20 ratio. The training set will be used to train the model on various ML algorithms and the testing set to evaluate the performance of the trained model.

Measuring performance of selected ML algorithms using following metrics:

Predict: Making use of trained models to make predictions on test set.

Accuracy Score: To calculate the accuracy of the predicted labels by comparing them with the actual target labels in the test set

Precision: Measures proportion of true positives among predicted positives

Recall: Recall measures proportion of true positives among the actual positives.

F1-score: Is the harmonic mean of precision and recall, which provides balance between the two.

Support: Is number of instances in each class.

Precision, Recall, F1-Score and Support altogether is represented by a confusion Matrix.

Conclusion:

In summary, pilot study aims to use ML methods predicting the risk of diabetes using EMRs. The study will use binary classification to classify patients as either a positive or negative of developing diabetes. Features like age, BMI, BP, family history, and lifestyle factors will be considered. LR, DT, RF, and SVMs will be used for predictions, and the performance of the system will be evaluated by accuracy score, prediction on test set, and confusion matrix. This study will provide valuable insights into the potential use of ML for diabetes risk prediction and could inform the development of largerscale studies in the future.

References:

Alghamdi, M., & Al-Mallah, M. (2017). Machine learning in cardiology: A review. Journal of Healthcare Engineering, 2017, 1-11.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., & Vlahavas, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.