Design and Application of a Machine Learning System for a Practical Problem (Pilot Study)

University of Essex,School of Computer Science and Electronic Engineering

May 4, 2022

Registration Number: 2112123
Name: Prashant Raj
Subject: Machine Learning(CE802)

1 Introduction

This is a pilot study for Machine Learning Technique to classify the weather a person is going to suffer from diabetes yes or no .Based on the several features of the patients (e.g. age, average glucose level, ethnicity, etc.) which is provided by clinic where the clinician has access to historical data of past patients.Machine Learning technique have recently provided approachable decision during complex situation like diabetes,however classifying the stage of patient health is a challenging task but this can be outstanding handled by Machine learning classification technique.

2 Background

A machine learning model is will use in this classification problem .This Supervised learning comprises of labelled information feature and ideal target variable where data begin used to trained and predict categorical variable this is called learning classification which is allow to classified categorised whether they are diabetic or not.however this process we called binary classification.Whereas regression predictive model uses for predicting continuous values. Therefore, for predicting whether patient going to suffer from diabetes yes or no .fall under classification predictive task (True or False).

3 Proposal

The basic data to develop an algorithm is information from past patients connected to features in variables that may be used to classify them into a category. To train the model and produce the best potential result, this data should be of a fair quantity and free of errors and inaccuracies. This model will be built on classification methods because the challenge is primarily related to data classification. The following things must be considered when choosing the appropriate algorithm:

1. Training data size
2. Training Time
3. Number of features
4. Linearity

4 Learning procedure

4.1 Decision Tree

The Decision Tree algorithm is part of the supervised learning algorithms.The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.This is non-parametric and does not require any distribution. Whether or not the data sets can be divided linearly, a decision tree can effectively handle outliers.

4.2 Random Forest

It can be defined as the combination of multiple decision tress , which can be called as the supervised learning algorithms. The problem with the decision tree is of bias and over-fitting. This problem can be solved with the help of random forest which uses the hyper-parameters such as the node size , the number of the trees and and the number of features samples in the data set. It also uses the cross-validation to finalize the prediction. With the help of the random forest we can reduced the risk of the over fitting and we can also identify the importance of the features.

4.3 K-Nearest Neighbors

It can be termed as the supervised machine learning algorithms. It can be used for both the task be it the classification or the regression problems. In this algorithm choosing the right value. This algorithm classified the data based on the similarity measured among the data.

4.4 Linear Regression

Linear regression helps us to identify the linear relationship between the dependent and the independent variable. It helps us to predict the value of the target variable with the help of response variable. It is one of the oldest and simple models to use, understand and identify the relationship between different variables. If we are using only one explanatory variable, then it is called simple linear regression and if we are using more than one then it is called the multivariate linear regression.

5 Evaluation

Binary classification could be evaluated by prediction score which is indicate the model confidence on how the result divided in to the specific specific class .However to conclude a decision of whether the observation is classified as a positive and negative.here we can interpret the accuracy by selecting a classification threshold and comparing the probability against them.Observations with scores higher than the threshold are classified as positive and scores lower than the threshold are predicted to be negative.