Design and Application of a Machine Learning System for Prediction of Diabetes

Pilot-Study Proposal

Submitted as part of the assignment
In

CE802 Machine Learning

Submitted to

Dr. Vito De Feo

By

Tamanna
(PG21154802)

Pilot-Study Proposal (689)

School of Computer Science and Electronic Engineering University of Essex
January 2022

1. Pilot-Study Proposal

Diabetes is a disease caused by a high glucose level in the human body. Diabetes should not be neglected; if left untreated, it can lead to severe problems such as heart disease, renal disease, high blood pressure, eye impairment, and damage to other organs in the body. Diabetes may be managed if it is detected early. The task at hand specifies us to build a machine learning algorithm that can effectively provide predictions on whether the person will suffer from diabetes or not. Framing the question as, ’Will the given human being's habits lead him or her to be diabetic?’; the answer could be ’Yes’ or ’No’. Hence, we can categorize this scenario as a classification problem in a machine learning perspective. The dataset for prediction include 14 physical examination indexes:

Sr.no	Attribute Description
1.	Age
2.	Pulse rate
3.	BMI
4.	Height
5.	Weight
6.	Physique index
7.	Fasting glucose
8.	Waistline
9.	Left systolic pressure (LSP),
10.	Right systolic pressure (RSP),
11.	Left diastolic pressure (LDP),
12.	Right diastolic pressure (RDP),
13.	Low density lipoprotein (LDL)
14.	high density lipoprotein (HDL)

Table 1: Attribute Description for classification

The features should be included but not necessarily be limited to the ones compiled above. Once we obtain the features, we must apply the proper machine learning classification algorithm to this data. On this data set, we may test Decision Trees,, k-Nearest Neighbor Classifiers, Xg Boost and training the dataset.

Decision Trees is an algorithm which falls under Supervised Machine Learning in which the programmer describes what the input is and what the related output (in simple language if else type condition) is in the training data. The paths from root to leaf represent classification rules [1].

K-nearest neighbors (KNN) is a supervised learning approach for regression and classification. KNN attempts to predict the correct class for the test data by computing the distance between the test data and all of the training points. Then select the K locations that are closest to the test data.

The KNN technique examines the likelihood of test data belonging to the classes of 'K' training data and selects the class with the highest probability. In the case of regression, the value is the mean of the 'K' selected training points. It is a flexible and complex method that is also utilized for missing value imputing and resampling datasets [2]

Xgboost: The gradient boosting approach is used to develop ML algorithms using Xgboost. The gradient boosting decision tree (GBDT) is a fast and accurate parallel tree boosting algorithm that may be utilized in classification and regression applications. It is a popular machine learning method that is also quite efficient, since it wins half of the challenges on machine learning sites such as Kaggle. Xgboost is a gradient-boosted decision tree-based method. It includes a plethora of system and algorithmic improvements such as tree trimming, parallelization, and cross-validation [3].

Linear Regression is a machine learning approach that is based on supervised learning. It runs a regression test. Regression models a desired prediction value based on independent variables. Different regression models differ in terms of the sort of relationship they assess between dependent and independent variables, as well as the number of variables they evaluate[4].

Random Forest Regression is an ensemble approach that can do both regression and classification problems by combining several decision trees using a technique known as Bootstrap and Aggregation, sometimes known as bagging. The core idea is to use numerous decision trees to determine the final output rather than depending on individual decision trees[5].

Evaluation Metrics: In our proposed study three classification algorithms are going to be implemented and their performance will be evaluated on the basis of parameters like accuracy, F1 Score, Precision and Recall [6].

Evaluation Metrics	**Definitions**
Accuracy	Tells accuracy of implemented classifier in prediction
Precision	Accuracy of classifier will be measured by precision
Recall	Measures Completeness of classifier
F1 Score	Weighted average of precision and Recall

Table 2: Performance evaluation metrics

Deploying an appropriate machine learning model with maximum accuracy is somewhat a next step for clinicians from the NHS to predict whether a person is diabetic or not.

References

[1] M. Rout and A. Kaur, “Prediction of Diabetes Risk based on Machine Learning Techniques,” *Proc. Int. Conf. Intell. Eng. Manag. ICIEM 2020*, pp. 246–251, 2020, doi: 10.1109/ICIEM48762.2020.9160276.

[2] D. Shetty, K. Rit, S. Shaikh, and N. Patil, “Diabetes disease prediction using data mining,” *Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017*, vol. 2018-January, pp. 1–5, 2018, doi: 10.1109/ICIIECS.2017.8276012.

[3] L. Wang, X. Wang, A. Chen, X. Jin, and H. Che, “Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model,” *Healthc.*, vol. 8, no. 3, pp. 1–11, 2020, doi: 10.3390/healthcare8030247.

[4] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021, doi: 10.1016/j.icte.2021.02.004.

[5] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, “Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications,” *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/9930985.

[6] J. Xue, F. Min, and F. Ma, “Research on diabetes prediction method based on machine learning,” *J. Phys. Conf. Ser.*, vol. 1684, no. 1, 2020, doi: 10.1088/1742-6596/1684/1/012062.