Diabetes type Diagnoses utilizing machine learning

Salma El Shahawy

Mael Illien

The City University of New York
School of professional studies
CUNY SPS

Supervisor

Dr. Nasrin Khansari

The purpose of the study

Diabetes mellitus is suspected based on symptoms. It is crucial to diagnose the patient within the first 24 hours of admission to an Intensive Care Unit (ICU) to save the patient's life. However, current diagnosis tests commonly require a long time to determine the patient's diabetes condition. Therefore, the goal of this study was to build machine learning models that utilize ensemble algorithms to predict whether a patient suffers from diabetes mellitus. The study is based on patient information from a dataset compiled by MIT and GOSSIS which are part of a growing global effort and consortium spanning Argentina, Australia, New Zealand, Sri Lanka, Brazil, and more than 200 hospitals in the United States. Specifically, the datasets include more than 130,000 ICU visits from patients, spanning a one-year timeframe. The dataset consists of patients' health conditions and symptoms that will be used to train the ML model. Informed by this dataset, the ensemble algorithm will be able to classify whether a new patient is suspected to suffer from diabetes, which ultimately reduces diagnosis time.

Introduction

Diabetes Mellitus is a metabolic disturbance disease that is characterised by hyperglycaemia and a relative lack, or complete absence of insulin, which by virtue may affect all organ systems in the body [1]. Prevention, by rapid diagnosis, and treatment are important in patients that suffer from this disease. Many of the complications associated with diabetes, such as nephropathy, retinopathy, neuropathy, cardiovascular disease, stroke, and death, can be delayed or prevented by early diagnosis [2,3].

However, finding a way to diagnose the mellitus type within the first 24 hr of admission to the ICU unit is pretty challenging due to delays in the test results. With the advent of machine learning applications in the healthcare industry, early diagnosis of mellitus diabetes can be predicted by developing a machine learning algorithm that utilizes electronic patient's records. Information such as patients' age, BMI, previous blood test outcomes, and previous urine test outcomes will inform the algorithm to classify the diagnosis based on statistical methods.

Literature Review

There are a growing number of new statistical procedures Leo Breiman (2001b) has called ”algorithmic.” Coming from work primarily in statistics, applied mathematics, and computer science, these techniques are sometimes linked to “data mining,” “machine learning,” and “statistical learning.”. Among the great variety of algorithmic approaches, there is a group that depends on combining the fitted values from a number of fitting attempts; fitted values are said to be “combined” or “bundled” (Hothorn, 2003). For example, one might combine the fitted values from several regression analyses that differ in how nuisance parameters are handled.

Another example would be to average the fitted values from nonparametric regression applied to a large number of single-subject experimental trials (Faraway, 2004). The term “ensemble methods" is commonly reserved for bundled fits produced by a stochastic algorithm, the output of which is some combination of a large number of

passes through the data. Such methods are loosely related to iterative procedures on the one hand and to bootstrap procedures on the other. An example is the average of a large number of kernel smoothes of a given variable, each based on a bootstrap sample from the same data set. The idea is that a “weak” procedure can be strengthened if given an opportunity to operate “by committee.” Ensemble methods often perform extremely well and in many cases, can be shown to have desirable statistical properties (Breiman, 2001a; 2001c; Buehlmann and Yu, 2002; Mannor et al., 2002; Grandvelet, 2004)[17].

Ensemble learning is a powerful machine learning paradigm which has exhibited apparent advantages in many applications. An ensemble in the context of machine learning can be broadly defined as a machine learning system that is constructed with a set of individual models working in parallel and whose outputs are combined with a decision fusion strategy to produce a single answer for a given problem.

Most of the related literature on diabetes classification is based on the Pima Indians Diabetes Database (PIDD) data from the UCI machine learning repository. The approaches include Logistic Regression, Naives Bayes, Support Vector Machines but also tree-based methods like Decision Trees, Bagging and Random Forests. Several authors have used boosting techniques to improve performance over simple trees. Hasan et al. in [5] places particular importance on the pre-processing steps and uses an ensemble model to achieve better classification performance than similar works.

Summary

To sum up, our goal is to build a machine learning model that is capable of diagnosing diabetes. While logistic regression is the most common classification approach in the medical setting, the performance in other studies using the PIMA dataset was inferior to other methods. We include logistic regression in our study in order to have a ground truth model to reference and compare. The dataset is structurally different but the improved performance [5] justifies our approach in considering ensemble models built from the aggregation of weak learners with different parameters.

References

[1] A. S. Alanazi and M. A. Mezher, "Using Machine Learning Algorithms For Prediction Of Diabetes Mellitus," 2020 International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia, 2020, pp. 1-3, doi: 10.1109/ICCIT-144147971.2020.9213708.

[2] Diagnosis and Classification of Diabetes Mellitus, American Diabetes Association, Diabetes Care Jan 2010, 33 (Supplement 1) S62-S69; DOI: 10.2337/dc10-S062

[3] K. Driss, W. Boulila, A. Batool and J. Ahmad, "A Novel Approach for Classifying Diabetes' Patients Based on Imputation and Machine Learning," 2020 International Conference on UK-China Emerging Technologies (UCET), Glasgow, UK, 2020, pp. 1-4, doi: 10.1109/UCET51115.2020.9205378.

[4] D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2018, pp. 924-928, doi: 10.1109/IEMCON.2018.8614871.

[5] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[6] G. Luo, Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst 4, 2 (2016). https://doi.org/10.1186/s13755-016-0015-4

[7] N. Nai-arun, R. Moungmai, Comparison of Classifiers for the Risk of Diabetes Prediction, Procedia Computer Science, Volume 69, 2015, Pages 132-142, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.10.014.

[8] S. Perveen, M. Shahbaz, A. Guergachi, K. Keshavjee, Performance Analysis of Data Mining Classification Techniques to Predict Diabetes, Procedia Computer Science, Volume 82, 2016, Pages 115-121, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.04.016.

[9] A. M. Posonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 498-502, doi: 10.1109/ICISS49785.2020.9316001.

[10] D. Sisodia, D. S. Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science, Volume 132, 2018, Pages 1578-1585, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.122.

[11] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.

[12] V. V. Vijayan and C. Anjali, "Prediction and diagnosis of diabetes mellitus — A machine learning approach," 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Trivandrum, India, 2015, pp. 122-127, doi: 10.1109/RAICS.2015.7488400.

[13] Berk, R.A. (2003) Regression Analysis: A Constructive Critique. Sage Publications, Newbury Park, CA.

[14] Berk.R.A. (2005) “Data Mining within a Regression Framework,” in Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Oded Maimon and Lior Rokach (eds.), Kluwer Academic Publishers, Forthcoming.

[15] Berk, R.A., and J. Baek. (2003) “Ensemble Procedures for Finding High Risk Prison Inmates.” Department of Statistics, UCLA.

[16] Berk, R.A., Sorenson, S.B., and Y. He (2005a) “Developing a Practical Forecasting Screener for Domestic Violence Incidents,” Evaluation Review, forthcoming.

[17] Buehlmann, P. and Bin Yu (2002), “Analyzing Bagging.” The Annals of Statistics 30: 927-961.