DATA698 Capstone Project - Literature Review

Salma Elshahawy, Mael Illien

Related Work

A literature review of machine learning methods for diabetes classification reveals a number of similar approaches, the most common being Decision Trees (DT), Random Forests (RF), Naive Bayes (NB), Support Vector Machines (SVM) and various ensemble models implementations such as boosting to increase performance. In general, RF and boosted DT seem to perform best. Most of the following papers use the Pima Indians Diabetes Database (PIDD) data sourced from the UCI machine learning repository. In some of these articles, we can also note performance enhancements obtained from data pre-processing steps.

Sisodia et al. in [10] use three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes to detect diabetes at an early stage. Naive Bayes performs a bit better than Decision Trees but both techniques greatly outperform SVM.

Perveen et al. in [8] classify the occurrence of diabetes in 3 ordinal age groups using C4.5 (J48) decision tree as a base learner. The base learner is improved using bagging and boosting ensemble techniques. In their model, Adaboost performed better than bagging or the standalone C4.5 (J48) decision tree.

The American Diabetes Association in [2] provides an in-depth look at diabetes mellitus from definition and description to diagnosis and classification. The article also provides categories for increased risk for diabetes which can be informative in feature selection and generation.

Nai-arun et al. in [7] apply Logistic Regression, Naive Bayes, Decision Trees, Artificial Neural Networks and Random Forests to the diabetes classification problem. Bagging and boosting methods are used to improve the robustness of these methods. In this paper, the best prediction accuracy was achieved by Random Forests.

Hasan et al. in [5] place particular importance on the data pre-processing step to improve robustness. Multiple machine learning techniques were employed including KNN, Decision Trees, Random Forests, Naive Bayes, AbaBoost and XGBoost. Multilayer Perceptron was also used. An ensemble of AdaBoost and XGBoost provided the best classification performance, exceeding the results of similar literature using the Pima Indians dataset.

Luo [6] explores the tradeoff between the performance enhancements that machine learning methods provide at the cost of interpretability, in comparison to statistical learning methods like logistic regression which is widespread in healthcare. The paper aims to automatically explain the results of any machine learning predictive model without degrading accuracy. It does so by building two models simultaneously, one for prediction and another for explanation. The first model maximizes accuracy, while the second is a rule-based associate classifier only used for explaining the first model's results without being concerned about its own accuracy.

Sonar et al in [11] compare Decision Trees, Artificial Neural Networks, Naive Bayes and Support Vector Machines to predict the diabetic risk of a patient. The best performance is obtained using Decision Trees.

Alanazi et al in [1] studies the efficiency of diagnosing and predicting diabetes using Random Forest and Support Vector Machines with data from the Security Force Primary Health Care in Tabuk, Saudi Arabia. Random Forests outperform SVM and achieve an AUC of 0.99 on this dataset.

Driss et al in [3] focus on data imputation and is limited to the K-nearest neighbors technique. In this paper, the best performance is achieved using k=11 on the imputed dataset.

Vijayan and Anjali in [12] use AdaBoost to increase the performance of Decision Stump, Support Vector Machines, Naives Bayes and Decision Trees as base classifiers. SVM performed best as a base classifier, but the boosted Decision Stump delivered the best performance overall.

Bibliography

[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.