Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 220 (2023) 868–873

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Workshop on Hospital 4.0 (Hospital)
March 15-17, 2023, Leuven, Belgium

A Comparative Study of Classification Algorithms for Early Detection of Diabetes

César Carpinteiro[a], João Lopes[a,*], António Abelha[a], Manuel Filipe Santos[a]

[a]*ALGORITMI/LASI Research Center, University of Minho, Portugal*

Abstract

Diabetes is a chronic disease that affects millions of people worldwide. Early detection of diabetes is crucial for preventing or delaying the onset of its associated complications. In this study, in collaboration with Unidade Local de Saúde do Alto Minho (ULSAM), we conducted a comprehensive comparison of various classification algorithms for the early detection of diabetes. We collected and pre-processed a dataset of patient records, containing personal information, associated medical problems and drugs. The dataset was divided into training and testing sets and used to train and evaluate several popular classification algorithms. The results of our study revealed that the Multilayer Perception (MLP), Gradient Boost Machine (GBM) and Random Forest (RF) algorithms had the highest overall performance, closely followed by Support Vector Machines. These findings demonstrate the potential of these algorithms for use in the early detection of diabetes and suggest that further research is needed to refine and optimize these models for clinical use.

Keywords: Machine Learning; Classification Algorithms; Diabetes;

1. Introduction

Diabetes is a chronic disease that affects millions of people worldwide, and it has become a major public health concern [1]. According to the International Diabetes Federation, it is estimated that by 2017, diabetes affected 425 million people worldwide, of which 4 million died in the same year. The trend is for these numbers to increase

* Corresponding author. Tel.:+351934373667.
E-mail address: lopesit@outlook.pt

10.1016/j.procs.2023.03.117

dramatically in the coming decades, putting an ever-greater burden on healthcare systems [2]. Diabetes is characterized by high blood sugar levels, which can lead to a range of serious health complications if left untreated. Early detection and diagnosis of diabetes are crucial for preventing or delaying the onset of these complications. Diabetes can be classified into two main types: type 1 Diabetes and type 2 Diabetes. Type 1 is an autoimmune disorder that occurs when the body's immune system attacks and destroys the insulin-producing cells in the pancreas. When the patient develops insulin resistance or the pancreas is unable to produce sufficient insulin, Type 2 metabolic disorder occurs. [1]. Several screening methods have been developed to detect diabetes, including fasting plasma glucose, oral glucose tolerance test, and hemoglobin A1c test [3]. However, these methods can be time-consuming, costly, and may not be feasible for some populations. Therefore, there is a need for alternative methods for the early detection of diabetes. One potential solution is to use Machine Learning (ML) algorithms to analyze patient data and identify individuals at high risk of diabetes. ML algorithms, especially classification algorithms, can be trained to recognize patterns in patient data and make predictions about the risk of diabetes [4]. These algorithms have been successfully applied in various medical applications, such as cancer diagnosis and drug discovery.

In this study, in collaboration with Unidade Local de Saúde do Alto Minho (ULSAM), we conduct a comprehensive comparison of several popular classification algorithms for the early detection of diabetes. The study aims to evaluate the performance of different techniques, identifying the best-performing algorithm, and provide insights into the potential of ML for use in diabetes screening. The dataset used in this study contains personal information, health problems and prescription drugs of the patient and it will be used to train and evaluate various classification algorithms including Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Multilayer Perception (MLP), Gradient Boost Machine (GBM), Naïve Bayes (NB) AdaBoost (AB) and Linear Discriminant Analysis (LDA). The rest of the paper is organized as follows: In the next section, a small framework to the theme is performed, followed by a chapter on methodologies, techniques and technologies used. Next, we describe all phases of this study, including data collection, preprocessing, and evaluation. The results of the study are presented in the following section, followed by a discussion of the findings. The conclusion and future work are discussed in the final section.

2. Background

2.1. Portuguese Reality of Diabetes

Diabetes has evolved over the years, becoming a major concern worldwide. In 2015, this disease was responsible for 4% of deaths in Portugal. Moreover, "it is estimated to affect 13.3% of the population aged 20-79 years" [5]. In 2019, 4.2 million people died with Diabetes and 463 million were diagnosed with Diabetes between the ages of 20 and 79. To combat this problem, the monitoring committee of the National Diabetes Control Program created the Diabetic Guide, as a support element for health professionals, who care for patients with this pathology, with the aim of minimizing the effects through a set of standards of good practice, such as appropriate diagnoses, patient education in self-management and continuous medical care [6, 7].

2.2. Related Works

Due to the worsening of this disease worldwide, it has become a priority to understand how it is possible to anticipate diagnoses and control hospital costs inherent to any healthcare unit when receiving patients with this pathology, making it the target of several research studies and the development of intelligent solutions to improve its management.

A first study by Sneha & Gangil (2019) [8] focuses on detecting and preventing diabetes complications at an early stage through predictive analytics to be able to support people suffering from this disease immediately. An important analysis for this project focused on the fact that when blood glucose levels deviate from normal, it can lead to long term complications, so it is hoped that individuals alerted to the fast approaching changes in their blood glucose levels will enable them to take preventive measures. They studied and analyzed several datasets from the UCI Machine Learning Repository, and after conducting an analysis, resulted in a sample of 2500 rows and 15 attributes containing diabetic patient information such as their sexual orientation, age, weight chart, blood diagnosis

(glycated hemoglobin), hypertension, and smoking propensity. The application of various ML techniques such as NB, J48, RF, MLP and LR for comparison and prediction of diabetes proved the feasibility of these techniques in detecting the disease, and SVM was the model that generated the best results.

Another study, developed by Venkatesh et al. (2019) [9], attempts to understand the effect of antidiabetic drugs for a particular person's health through predictive Data Mining techniques. The authors point out that the type of medication that is appropriate depends a lot on the person to person and health factors, meaning that there is a high probability of the patient developing a series of side effects caused by unsuitable medications. With such a variety of medications, doctors themselves can have difficulty knowing which medications to prescribe, how they will affect the body, and to what extent it will help in treatment. So, they selected old records of hospital inpatients, which after pre-processing techniques such as removing outliers, cleaning data, normalizing data, and data reduction, were subjected to the application of some Data Mining algorithms, namely NB, DT, and J48. With this work, it was possible for physicians to analyze the results produced by these algorithms, and thus, based on attributes such as age, gender, and weight, treat the patient in the most appropriate way, knowing precisely which drugs are the most advisable, i.e., which ones will cause fewer side effects and at the same time help more in the treatment of the disease, allowing to reduce the number of drugs in the patient's test phase, reducing associated costs.

3. Methodologies and Materials

3.1. DSR and CRISP-DM

To understand whether it is possible to predict the clinical status of patients, two methodologies were followed: Design Science Research (DSR) as the research methodology, and Cross-Industry Standard Process for Data Mining (CRISP-DM). DSR structures the project in 6 phases: 1. Problem identification and motivation; 2. Definition of solution objectives; 3. Design and development; 4. Demonstration; 5. Evaluation; 6. Communication. These phases provide guidelines for the evaluation and development of research projects [10]. CRISP-DM is the process that presents the life cycle of a Data Mining project in a real-world context. This project includes a set of six phases, including: 1. Business understanding; 2. Data understanding; 3. Data preparation; 4. Modeling; 5. Evaluation; 6. Deployment [11].

3.2. Tools and Algorithms

Python programming language was used for the preparation of data and consequent analysis of it. A set of libraries were used to enable the analysis and consequent predictions of the data, highlighting:

1. Panda allows the use of the dataframe object to provide storage and manipulation [12];
2. Matplotlib for data visualization [13];
3. Scikit-learn to develop ML algorithms [14].

4. Case Study

4.1. Business Understanding

The first phase focuses on understanding the project’s objectives and requirements. The main purpose is predicting Diabetes, using a dataset with personal information, health problems and prescription drugs of the patients. Considering that this is a preliminary phase of the study, it will be attempted to understand which lines the project should follow to model future work.

4.2. Data Understanding

Initially, a list of several diabetic and non-diabetic patients was obtained. The diabetic cases were isolated to identify some important characteristics, such as correlations, classes, attribute descriptions, and value variations. It was concluded that the information obtained presents various records for each patient, such as the identified health problems, prescribed medication, and personal information such as their age and gender. It was possible to understand a set of clinical data describing the existence of cardiovascular, respiratory and other pathologies on a given day, as well as other relevant ones, such as the suspicion of infection to COVID-19 or other relevant criteria (such as anxiety states, obesity, among others). Drugs already prescribed are normally associated to their codes, allowing the codification of this attribute.

4.3. Data Preparation

The data preparation stage encompasses all the changes that were made to the data before any kind of modeling process was applied.

A set of sequentially developed steps follows:

1. Creation of a target variable that did not exist in the data set, through a binary classification of an attribute of which contained the codes associated with the diabetes health problem (number 1 assigned to diabetics, and the others with the number 0);
2. Due to a large disparity of values in the age of the patients, the granularity of the patients was decreased, fitting them into partially balanced age groups. Three sets were then created, inserted in a new column, group 0 (from 0 to 63 years) with 1009686 records, group 1 (from 64 to 75 years) with 809817 records, and group 3 (from 76 years upwards) with 1009686 records. The separation points chosen were those that, given the data set, best separated the ages into equitable sets;
3. The columns composed of categorical values, considering that most ML algorithms only work with variables in numerical format, were subjected to encoding techniques, to be new representations of these data, so that they could be useful and contribute to the modeling process;
4. The attributes referring to problems and medications were subjected to the application of the One-Hot Encoding technique, which consists of creating a new column for each unique value of an existing column;
5. Execution of a sampling technique, called undersampling, to reduce the number of non-diabetic patients, resulting in a data set where the target values are balanced, minimizing overfitting problems. Both sets now contain 355496 records each;
6. The last step refers to data cleaning and removal tasks, to contain one row per identified patient, based on their identification. Thus, duplicate examples and information were avoided, which would lead to generalization of individual patient conditions, which contained many records and could wrongly influence the predictive ability.

4.4. Modeling

Different classification techniques were applied, according to those identified in previous studies, such as: Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Multilayer Perception (MLP), Gradient Boost Machine (GBM), Naïve Bayes (NB) AdaBoost (AB) and Linear Discriminant Analysis (LDA). All models were performed with a set of operations used to optimize results, which the most relevant to mention is hyperparameter tuning, a technique capable to find the best combination of model parameters. Since model performance is usually very sensitive to such parameters and tuning those based on a predefined split of the dataset should be avoided, we applied this technique with a grid search, so that each model picks a set of values within a grid of hyperparametric values, iteratively evaluating the model's performance and choosing the one that translates optimal values within the learning process. It was also applied the cross-validation technique with 5 folds.

4.5. Evaluation

The developed models are evaluated to understand if they meet the defined objectives, enabling a global analysis of this research with a review of the entire process and determination of future steps. Metrics are used to understand which algorithm has the best results. Considering the context of this project, the following values were defined for the metrics Area under the ROC Curve (AUC), Accuracy, Precision, Sensitivity and F1-Score:

- 80 % < AUC, Precision, Sensibility, Accuracy and F1-Score < 90%

5. Results and Discussion

The best result obtained by the algorithm are shown for the target under study, presented by Table 1.

For each model, the evaluation values are shown. For this purpose, Accuracy is used, to measure the closeness of the predictive values, and AUC, which allows one to understand the overall performance of the developed models. The measures Precision (ratio of True Positives to all Positives) and Sensitivity (evaluating the True Positive Rate) allow one to evaluate the relevance of the results obtained [16]. It is evident that SVM, MLP and GBM produce the best predictive results.

Table 1 – Evaluation of the implemented ML techniques

Model	AUC	Accuracy	Precision	Sensibility	F1-Score
LR	87.0%	79.0%	79.0%	79.0%	79.1%
DT	74.2%	75.0%	79.0%	75.5%	75.6%
RF	86.8%	79.1%	79.0%	79.0%	79.5%
KNN	74.3%	69.2%	72.2%	69.4%	67.7%
SVM	87.6%	80.0%	80.3%	80.0%	80.1%
MLP	88.0%	80.2%	80.3%	80.5%	80.0%
GBM	87.2%	80.4%	80.0%	80.1%	80.1%
NB	81.2%	74.0%	74.1%	74.0%	73.0%
AB	86.9%	79.1%	79.0%	79.3%	79.0%
LDA	86.9%	79.2%	79.4%	79.2%	79.0%

6. Conclusions and Future Work

This research highlights the ability to predict a diabetic patient using ML techniques with information from the patient, along with their presenting pathologies and medications, at the time of the medical appointment. The predictive results show a clear distinction of three models according to the metrics used (SVM, MLP and GBM). The implementation of these techniques and the results obtained were reasonably satisfactory, given the existing expectations, since all the measures generally meet the initial objectives. Even so, it is also possible to understand that there is room for improvement in these values.

In terms of future work, it is defined to optimize these results, with new implementations focused on Deep Learning. We believe that the fact that there is a significant amount of data, namely the ability to reorganize these patients in time, as well as the union of data regarding Complementary Diagnostic and Therapeutic Means (CDMT) and clinical history, may justify these new developments and produce better results.

References

[1] Contreras, I., & Vehi, J. (2018). Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. Journal of Medical Internet Research.

[2] Toniolo, A., Cassani, G., Puggioni, A., Rossi, A., Colombo, A., Onodera, T., & Ferrannini, E. (2019). The diabetes pandemic and associated infections: Suggestions for clinical microbiology. Reviews and Research in Medical Microbiology, 30(1), 1–17. https://doi.org/10.1097/MRM.0000000000000155.

[3] Mayega, R. W., Guwatudde, D., Makumbi, F. E., Nakwagala, F. N., Peterson, S., Tomson, G., & Östenson, C.-G. (2014). Comparison of fasting plasma glucose and haemoglobin A1c point-of-care tests in screening for diabetes and abnormal glucose regulation in a rural low income setting. Elsevier BV. https://doi.org/10.1016/j.diabres.2013.12.030.

[4] Adaptive Business Intelligence: Predictive and Optimization Models in Healthcare. Master's Thesis, University of Minho, Guimarães, Portugal, 2020.

[5] Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. Journal of Medical Internet Research. Diabetes facts & figures. Obtido de International Diabetes Federation - Diabetes facts & figures: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.

[6] Saúde, S. -S. Acordo de cooperação entre a APDP e o Serviço Nacional de Saúde. Obtido de SNS - Serviço Nacional de Saúde: https://www.sns.gov.pt/noticias/2018/05/14/diabetes/.

[7] DGS. (2008). Programa Nacional De Prevenção e Controlo da Diabetes. Diabetes, 1(1), 1–6. https://www.dgs.pt/programa-nacional-para-a-diabetes/programa-nacional-para-a-diabetes.aspx.

[8] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0175-6.

[9] Venkatesh, G., Lawanyashri, M., & Sai Saraswathi, V. (2019). DATA mining application towards adverse effects of anti-diabetic drugs. International Journal of Innovative Technology and Exploring Engineering, 8(7), 2544–2546.

[10] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger & Samir Chatterjee (2007) A Design Science Research Methodology for Information Systems Research, Journal of Management Information Systems, 24:3, 45-77, DOI: 10.2753/MIS0742-1222240302.

[11] Azevedo, Ana & Santos, Manuel. (2008). KDD, semma and CRISP-DM: A parallel overview. 182-185.

[12] Pandas Package. pandas. Retrieved December 2022, from https://pandas.pydata.org/.

[13] Matplotlib Package. matplotib. Retrieved December 2022, from https://matplotlib.org/.

[14] Scikit-learn Package. Machine Learning in Python. Retrieved December 2022, from https://scikit-learn.org/stable/.

[15] Cortez, P. (2014). [BOOK] Modern Optimization with R. In Springer.

[16] Santos, M. F., & Azevedo, C. (2005). DATA MINING - Descoberta de Conhecimento em Bases de Dados. FCA - Editora de Informática.