Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 220 (2023) 862–867

Procedia
Computer Science

www.elsevier.com/locate/procedia

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

Identifying Diabetic Patient Profile Through Machine Learning-Based Clustering Analysis

João Gomes[a], João Lopes[a,]*, Tiago Guimarães[a], Manuel Filipe Santos[a]

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

Abstract

Given the rapid growth over the past 15 years, Diabetes is currently a key issue in medical science and healthcare administration. Considering the importance of the health sector in our society, it is critical to correctly diagnose and treat Diabetes in order to avoid immediate difficulties and reduce the chance of long-term issues. The analysis of vast amounts of data that are available in organizations is an important factor to describing their internal factors, predicting future trends, and prescribing the best course of action to improve their performance in light of the increasing technological evolution and the emergence of Artificial Intelligence (AI). The main objective of this project, which is being carried out in collaboration with the Unidade Local de Saúde do Alto Minho (ULSAM), is to define a typology of diabetic patients by building Machine Learning (ML) models from registered clinical information, medication, complementary diagnostic tools, therapeutic and monitoring data, and registered medication data.

Keywords: Clustering; Diabetes; Machine Learning;

1. Introduction

Diabetes is a group of metabolic diseases characterized by high blood sugar levels resulting from defects in insulin secretion, insulin action, or both. The International Diabetes Federation estimates that by 2017, diabetes affected 425 million people worldwide, 4 million of whom died in the same year [1]. The announced increase in the disease and the risks associated with it have led to the creation, by the monitoring committee of the National Diabetes Control

* Corresponding author. Tel.: +351 934373667.
E-mail address: lopesit@outlook.pt

10.1016/j.procs.2023.03.116

Program, of the Diabetic Guide. According to the commission, this guide is the responsible element for the achievement of the therapeutic objectives of the person with diabetes and several norms of good professional practice in the approach to this disease. Timely diagnosis, patient education in self-management, and ongoing medical care are necessary to prevent acute complications and minimize the risk of long-term complications. The role of Machine Learning (ML) techniques has been crucial to improve disease detection and improve medical decisions. This project was developed in collaboration with Unidade Local de Saúde do Alto Minho (ULSAM) and aims to demonstrate that, using Data Mining (DM) techniques, specifically clustering and classification algorithms, it is possible to identify patient typologies and, consequently, identify patients with the disease. The document is structured in the following way: The first section is the introduction where the main idea of this work is presented, the second is the background where the problem is defined as well as the theory behind the work, the third section is the description of the material and methods, where the tools used are described. The fourth chapter of this paper is the discussion where some views on the results are presented. Finally, the last section presents some conclusions and basic ideas about the work to be done in future.

2. Background

2.1. Diabetes

Diabetes mellitus is a serious worldwide health problem and one of the main chronic diseases affecting human beings. The problem occurs when there is a lack of insulin production and/or there is an inability of insulin to properly exert its effects, resulting in an impaired ability of the body to use the main source of energy, glucose, and consequent increase in glucose (sugar) levels in the blood. As sugar is essential for cell metabolism, it needs to be transported to the cells, and, to this end, the pancreas produces insulin, the hormone that will capture glucose from the bloodstream and take it to the cells throughout the body, where it will be used as energy [2].

2.2. Related Works

Due to the progress of Diabetes in recent years and the serious complications that result from it, the incorporation of ML tasks and models in healthcare has brought with it some previously unknown considerations. The study developed by Asha Gowda Karegowda et al., (2012) [3] proved that the cascade model developed with clustering (K-means) and K-Nearest Neighbors obtained the best results when compared to a collection of results from the literature analyzed. Initially, the sample data of diabetic patients were collected from the PIMA database, where they were divided into two categories: test positive and test negative, each with 8 characteristics, including the number of times pregnant, diastolic blood pressure, age, and Body Mass Index (BMI). The results demonstrated that, with the use of GA and CFS tools, it's possible to achieve 96.67% accuracy in the prediction of new diabetic patients.

On the other hand, Alghamdi et al., (2017) [4] used some classification models together with other approaches (Ensembling and SMOTE) to predict disease as accurately as possible. The study was incorporated into a project called FIT that included a dataset related to the cardiorespiratory fitness of 32,555 patients in the Detroit area - USA. The researchers used several ML models to predict the occurrence of diabetes and realized that even though these two models performed well, the Ensemble approach was even more accurate because it combines predictions from multiple models. In this study, the combination of three decision tree models achieved an AUC of 92%, using 13 features such as age, resting heart rate, obesity, hypertension, and others. Considering these results, the importance of clinical information for the diagnosis of disease should be emphasized, since it is an essential indicator in obtaining good predictive results.

3. Materials and Methods

3.1. DSR and CRISP-DM

For this research, two methodologies were followed: Design Science Research (DSR) and Cross Industry Standard Process for Data Mining (CRISP-DM). The first is a fundamental methodology for the success of a project in

Information Systems because, besides being consistent with previously established principles and guidelines, it aims to improve the production, presentation, and evaluation of the research. It is divided into six activities [5] which are: Understanding the Problem (1), Definition of solution objectives (2), Design and Development (3), Demonstration (4), Evaluation (5) and Communication (6). The second one was also adopted since it supports the life cycle of a DM project. This is composed of 6 stages [6]: Business Understanding (1), Data Understanding (2), Data Preparation (3), Modeling (4), Evaluation (5), and Implementation (6). For this study, a crossover between these two methodologies was used as represented in Table 1.

Table 1. Crossover of DSR and CRISP-DM.

	DSR Phases	CRISP-DM Phases
Phase 1	1,2	1,2
Phase 2	3	3,4
Phase 3	4,5	5
Phase 4	6	6

3.2. Tools and Technologies

For the development of the project, the Python programming language and, especially, the Sklearn library were used for phases 1 and 2 as well as for the third phase [7]. All models were developed using Jupyter Notebook.

4. Case Study

4.1. Business Understanding

The first phase focuses on understanding the project's objectives and requirements. The ULSAM aims to correct the non-adherence to treatment by diabetic patients by adopting new methods to define a typology appropriate to the population. Thus, the health service intends to work proactively to keep these patients under control and avoid the long-term complications of non-compliance with these treatments.

4.2. Data Understanding

The data was gathered from a single data source and divided into three datasets, with a time range of just one year. The 2864499 records across all datasets varied in number but shared the same characteristics, such as age, gender, and medication. The datasets were then combined, and although there were 14 accessible qualities, some of them were excluded from the analysis because they were deemed unimportant given the circumstances of this study.

4.3. Data Preparation

During this phase, several data processing methods were carried out, including normalizing the data type, using sampling techniques, and identifying individuals with diabetic disease. Additionally, 40 new attributes were added to the final dataset using the One-Hot Encoding technique to turn the categorical data from the top 20 issues and drugs into numerical values. As a result, the final dataset was correctly distributed in respect to the aim, diabetes, and had 9000 entries with various patients.

4.4. Modeling

Three distinct clustering models, based on K-Means, BIRCH Cluster, and Agglomerative Clustering, were implemented. On the other hand, various classification methods were applied, 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), in accordance with those discovered in earlier studies. The target corresponded to the value to be predicted while the other values referred to the features used to train the model. To comprehend the influence and/or behavior of the feature connected to the type of cluster in the classification models, six alternative scenarios had to be created throughout this phase.

4.5. Evaluation

The study of the modeling phase's output was included in the evaluation phase. Prior to examining the classification model findings, it was crucial to assess the clustering models using two metrics: the Davies-Bouldin Index and the Silhouette Coefficient. Then, to compare the derived predictions with the actual values, a set of appropriate measures such as Accuracy, Sensitivity, Precision, F1-Score, Kappa Index, and AUC were employed in order to calculate the performance of the developed ML algorithms.

5. Results and Discussion

The results obtained through the metrics for the evaluation of the clustering models are represented in Table 2.

Table 2. Clustering models evaluation.

Model	Silhouette Coefficient	Davies-Bouldin Index
K-Means (K=2)	0.17	0.95
Agg. Clustering (K=4)	0.03	4.89
Agg. Clustering (K=10)	0.007	4.48
BIRCH Cluster (K=4)	0.11	2.47
BIRCH Cluster (K=10)	0.15	2.41

It is feasible to see that the K-means model has the best distribution between generated clusters and the best separation of values. Given that values near to 0 show that the distance between them is not important, this model displays relatively low values for the Silhouette metric. The data obtained show good results for the Davies-Bouldin Index metric, which measures the average similarity of each cluster with its most comparable cluster. For all of the previous models, it is simple to understand how the data's similarity and density prevent the formation of distinct clusters that would allow for the identification of traits that could potentially be involved in the emergence of potential diseases. The metrics used to evaluate the classification models in predicting diabetes, considering 6 different scenarios, are depicted in the following results in Table 3. Only the metrics that have a more important weight in the decision of the best models, such as Accuracy, Sensitivity and Kappa index, are illustrated.

Table 3. Classification models evaluation.

Model	Metric	S1	S2	S3	S4	S5	S6
NB	Accuracy	74%	74%	74%	74%	74%	74%
	Sensitivity	74%	74%	74%	74%	74%	74%
	Kappa index	47%	48%	48%	48%	49%	47%
DT	Accuracy	75%	75%	75%	75%	75%	74%
	Sensitivity	75%	75%	75%	75%	75%	74%
	Kappa index	49%	50%	49%	50%	50%	48%
RF	Accuracy	79%	79%	79%	79%	79%	79%
	Sensitivity	79%	79%	79%	79%	79%	79%
	Kappa index	58%	58%	58%	58%	58%	57%

LR	Accuracy	79%	80%	79%	80%	79%	79%
	Sensitivity	79%	80%	79%	80%	79%	79%
	Kappa index	58%	59%	58%	59%	58%	58%
SVM	Accuracy	79%	80%	80%	80%	79%	79%
	Sensitivity	79%	80%	80%	80%	79%	79%
	Kappa index	59%	60%	59%	59%	59%	58%
GBM	Accuracy	79%	80%	80%	79%	79%	79%
	Sensitivity	79%	80%	80%	79%	79%	79%
	Kappa index	59%	60%	60%	59%	59%	58%
KNN	Accuracy	69%	64%	69%	65%	70%	65%
	Sensitivity	69%	64%	69%	65%	70%	65%
	Kappa index	38%	28%	37%	30%	40%	30%
MLP	Accuracy	80%	81%	80%	81%	80%	80%
	Sensitivity	80%	81%	80%	81%	80%	80%
	Kappa index	60%	61%	61%	62%	59%	61%

During an analysis of the data, it is evident that a group of models, including the MLP, GBM, SVM, LR, and RF, provide adequate outcomes. The accuracy ranges from 79% to 80% in all these models, indicating that the values in relation to the predicted diabetic patients who have the issue are satisfactory. The sensitivity, which relates to the percentage of correctly identified diabetic patients, also has average values of around 79%. All the models examined have relatively low values for the Kappa index, one of the key indicators for model evaluation in health-related initiatives. Since no index higher than 62% was reached and this metric shows that the data in the model are random, it is consistent with the findings and recommendations drawn from clustering model analysis.

6. Conclusions and Future Work

After the evaluation phase is through, it's important to interpret the results, paying particular attention to the detection of Diabetes to determine whether it's able to spot clinical patterns. Through the correlation of the factors in the distribution of the clusters, any clinical pattern might be identified. Heat maps are visualized in Figures 1 and 2 to show the significance of various factors in clustering models.

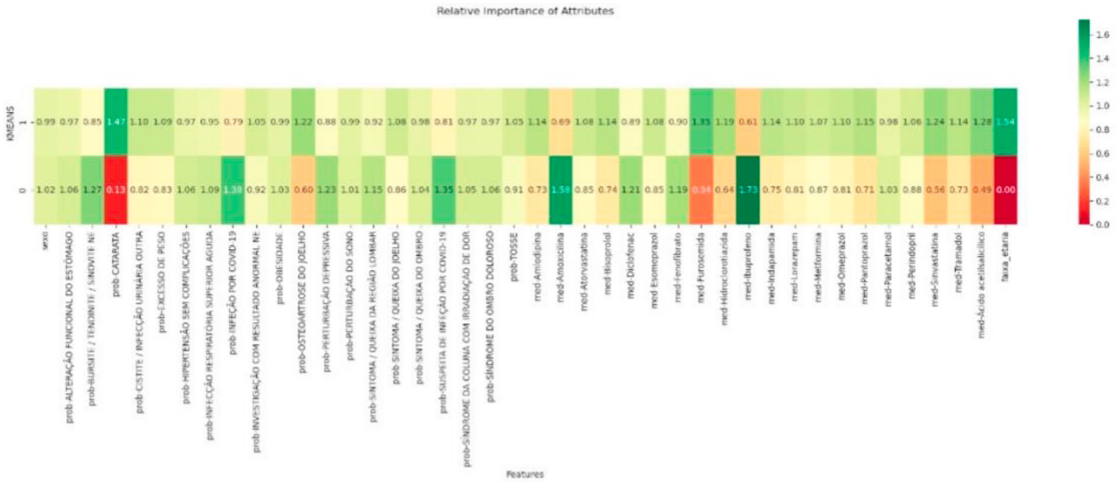

Fig. 1. Relative Importance of Features in K-Means (K=2).

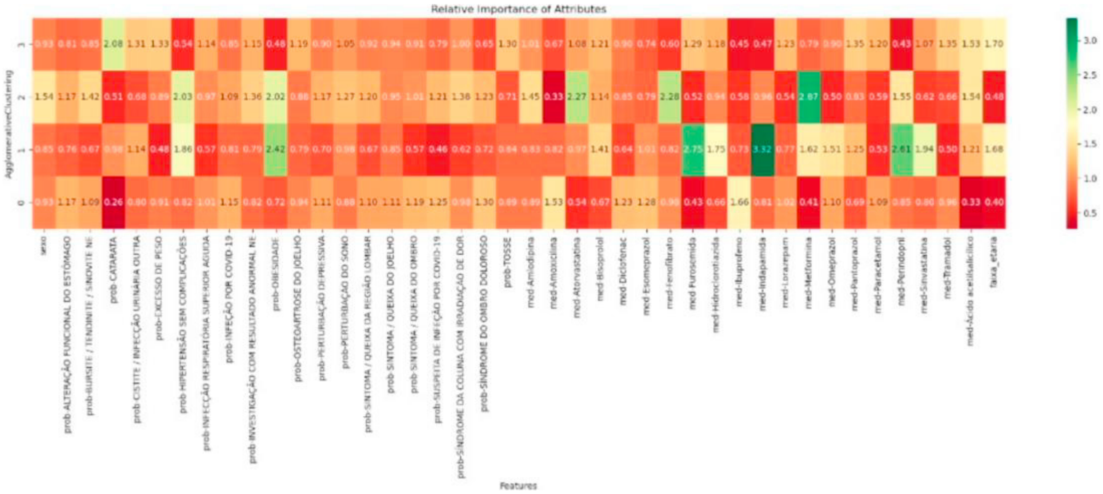

Fig. 2. Relative Importance of Features in Agglomerative Clustering (K=4).

It was possible to verify the existence of common variables in all models with a certain degree of influence, these being: Cataract pathology, as well as the drugs Amoxicillin, Tramadol, Perindopril, Furosemide and Indapamide, along with the patient's age group.

The results obtained from the classification models also point to the development of Deep Learning algorithms. The performance of the MLP model, the amount of existing data and the future focus on a timeline for identification of diabetic patients support the future phases of the work to be developed.

References

[1] Meuleneire, F. (2008). Management of diabetic foot ulcers using dressings with Safetac®: A review of case studies. Em Wounds UK (Vol. 4, Número 4).

[2] Maria, A., & Gadelha, J. (2009). Global burden of disease attributable to diabetes mellitus in Brazil Carga global de doença devida e atribuível ao diabetes mellitus no Brasil. 25(6), 1234–1244.

[3] Asha Gowda Karegowda, M.A. Jayaram, & A.S. Manjunath. (2012). Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients 148. International Journal of Engineering and Advanced Technology , 1(3), 147–151.

[4] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS ONE, 12(7), 1–15.

[5] V. Vijay, B. Kuechler, e S. Petter, «Design Science Research in Information Systems», n. 1, pp. 1–66, 2012, doi: 1756-0500-5-79[pii]\r10.1186/1756-0500-5-79.

[6] C. Pete et al., «Crisp-Dm 1.0», CRISP-DM Consortium, p. 76, 2000.

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