sciendo

The Applicability of Some Machine Learning Algorithms in the Prediction of Type 2 Diabetes

Oana VÎRGOLICI*
Academy of Economic Studies (ASE), Bucharest, Romania
[*]Corresponding author, oanavirgolici2022@gmail.com

Laura Gabriela TĂNĂSESCU
Academy of Economic Studies (ASE), Bucharest, Romania
tanasesculaura15@stud.ase.ro

Abstract. *Type 2 diabetes is a metabolic disease that causes abnormal high levels of glucose in the blood. The pancreas is healthy, but the body doesn't respond properly to its own insulin. The principal culprit is obesity, too much high fat tissue. So, measuring the body mass index or the waist circumference is a step to estimate the risk for this disease. Many people have no symptoms and the disease develops silently, causing serious problems with eyes, feet, heart and nerves. The prediction of diabetes is a very topical problem. In addition to medical guides, more and more machine learning models appear, trained on different databases. The purpose of these models is to predict diabetes, based on different parameters, not all of them coming from medical analyses. In the paper we present four diabetes prediction models, respectively based on the decision tree, support vector machine, logistic regression and k-nearest neighbors' algorithms. All models are trained and tested on a database with approximately 65,000 records (divided into 70% for training and 30% for testing), which contains two blood markers (haemoglobin A1c and glucose), an anthropometric parameter (body mass index), age, gender and three categorical parameters (smoking status, hypertension, heart disease). We identify that Haemoglobin A1C and glucose are the most influential predictors. The models are evaluated in terms of accuracy score and confusion matrix and a ranking is presented at the end. The results obtained are very encouraging for all the presented models.*

Keywords: type 2 diabetes mellitus, Machine Learning, Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM), k Nearest Neighbors (kNN).

Introduction

Type 2 Diabetes Mellitus (T2DM) is a complex chronic disorder that requires continuous medical care, patient self-management for control of abnormal glucose levels, and multifactorial risk reduction strategies to normalize blood glucose levels, lipid profiles and blood pressure to prevent or minimize acute and long-term microvascular complications (including retinopathy, nephropathy and neuropathy) and macrovascular complications (such as a heart attack and stroke) (DeFronzo et al., 2015).

Table 1. Diagnostic reference values

Parameters	Normal	Prediabetes	T2DM
Hemoglobin A1c	<5.7% (ADA) <6.0% (WHO)	5.7–6.4% (ADA) 6.0–6.4% (WHO)	≥6,5%
Fasting plasma glucose	<100 mg /dl (ADA) <110 mg /dl (WHO)	100–125 mg /dl (ADA) 110–125 mg /dl (WHO)	≥126 mg/dl
Two-hour plasma Oral Glucose Tolerance Test	<140 mg /dl	140–199 mg /dl	≥200 mg/dl

ADA – American Diabetes Association; WHO – World Health Organization

Source: DeFronzo et al., 2015.

DOI: 10.2478/picbe-2024-0021

Diagnostic criteria for diabetes have traditionally relied on blood glucose levels. More recently, glycated hemoglobin (HbA1c) has been added as an integrated measure of long-term glycaemia. The lifespan of a red blood cell is ~120 days and when blood glucose is high, as happens in diabetes mellitus, a lot of glucose enters free in the red blood cell and transforms the normal adult hemoglobin, hemoglobin HbA1, in the glycosylated hemoglobin (HbA1c). So, one value of HbA1c gives a picture of the average blood glucose over the past 120 days (Table 1). As it is shown in Table 1, the value for HbA1c is used to identify prediabetes, to diagnose diabetes and it is a useful tool to monitor the diabetes treatment. However, haemolysis, which reduces red blood cell lifespan, can make HbA1c an invalid measure. Also, older age, non-white race, high dietary fat intake, alcohol consumption, cigarette smoking, liver disease, kidney disease and iron deficiency can affect HbA1c independently of glycaemia. The value of glycaemia associated with complications varies depending on the complication. Current criteria are based on maintaining fasting glucose, postprandial (after eating) glucose and HbA1c levels below the threshold that is associated with an increased risk of developing diabetic retinopathy (DeFronzo et al., 2015).

In recent years, several models have been proposed for the prediction of diabetes, based on machine learning and deep learning techniques. These models used different characteristics for training and testing (serum glucose being the main parameter resulting from the analyses, and body mass index, age, gender, blood pressure being often used). Each model applies, in the known form, with improvements or by combining them, specific machine learning algorithms (SVM, DT KNN and so on). The database can be a public one, such as the Pima Indian Dataset Database (PIDD) or local or national databases, with data obtained from medical or administrative organizations.

In this paper, we propose four models for predicting diabetes, based on the following Machine Learning algorithms: Decision Tree (DT), Logistic Regression (LR), Support Vector Machine (SVM) and k-Nearest Neighbors (kNN), trained and tested on a large database, obtained from Kaggle. We evaluate these models in terms of accuracy score and confusion matrix and a ranking is presented at the end.

The paper is structured as follows: a first chapter in which we briefly present the notable results in the domain of machine learning models (literature review), then a chapter in which we point out the methodology used in our research, a chapter with the main results and discussions (in text and, in certain cases, also in graphic format), and a final chapter of conclusions.

Literature review

Different models have been proposed for predicting diabetes, using various machine learning techniques. is structured in two parts, from the perspective of the database used in training and testing the models).

Models using PIDD

Anuja Kumari and Chitra had proposed a system using SVM for diabetes classification (Anuja Kumari & Chitra, 2013). The training dataset used was PIDD. In experiment RadialBasis Function (RBF) kernel of SVM was used and it examines the higher-dimensional data. The kernel output was dependent on the euclidean distance. The accuracy of 78% was achieved.

Soliman and AboElhamd used Least Squares-Support Vector Machine (LS-SVM) and Modified-Particle Swarm Optimization (MPSO) algorithms (Soliman & AboElhamd, 2014). LS-SVM was run to find the optimal hyperplan to separate the patients into two classes: live and die. They used data from PIDD and the accuracy of 97.833% was obtained. These algorithms were compared to other algorithms.

Sridar and Shanthi had proposed a medical diagnosis system for diabetes prediction using back propagation and Apriori algorithm (Sridar & Shanthi, 2014). Dataset was PIDD. The patients were classified into three classes: low risk, medium risk, and high-risk patients. The system was implemented using Java and DotNet programming languages. The accuracy of 83.5%, 71.2%, and 91.2% received from back propagation algorithm, Apriori algorithm and with combining these both algorithms, respectively.

In another study (Amour Diwani & Sam, 2014), all the patient's data are trained and tested using 10 cross-validations with NB and DT. The results predicted that the best algorithm is NB with an accuracy of 76.3021%.

Sen and Dash used Weka (machine learning algorithms for data mining) on PIDD. Different algorithms were used and best accuracy (78,6%) was obtained by CART (Classification and Regression Tree) (Sen & Dash, 2014).

Dewangan and Agrawal proposed a system for the diagnosis of diabetes using Bayesian classification and multilayer perceptron (Dewangan & Agrawal, 2015). Data was classified into diabetic and non-diabetic. PIDD was used. Accuracy of 81.89% was achieved.

Iyer et al. used DT and NB on PIDD, which was split in 70% for trainoing and 30% for testing. Various tests were performed. Highest accuracy was obtained by utilizing percentage split test (Iyer et al., 2015).

Sisodia & Sisodia found that, among SVM (Support-vector machine), NB (Naive Bayes), and DT (Decision Tree) on PIDD, the NB classifier shows better accuracy at 76.30% (Sisodia & Sisodia, 2018).

Zou et al. applied RF (Random Forest), DT, ANN (Artificial Neural Network) for classification algorithm on PIDD after the feature reduction using Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) methods. They found that best accuracy is 77.21% obtained from the RF with the mRMR (Zou et al., 2018).

Tigga and Garg applied LR on PIDD for diabetic prediction. They found the number of pregnancies, BMI, and glucose level are the most significant variables for diabetes prediction among all features in PIDD. Their model has an accuracy of 75.32% (Tigga & Garg, 2019).

Purnami et al. proposed Smooth SVM (SSVM) and MKS-SSVM (Multiple Knot Spline SSVM), using PIDD. In their results, they achieved high accuracy for MKS-SSVM than SSVM, with accuracy 93.2% (Purnami et al., 2019).

Alehegn et al. used different algorithms (RF, KNN, NB, and J48-DT) on PIDD. The authors built a hybrid model consisting of all of the above algorithms. For large data computations NB and J48 are good, and for smaller datasets the best score was obtained by using KNN (Alehegn et al., 2019).

Tasin et al. used the PIDD and collected additional samples from 203 individuals from a local factory in Bangladesh. They used a semi-supervised model with extreme gradient boosting has been used to predict the insulin features (for the private dataset). The authors used DT, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques. The best result (81% accuracy, 0.81 F1 coefficient, 0.84 AUC) was obtained for the XGBoost classifier with the ADASYN approach. The explainable AI approach with LIME and SHAP frameworks is implemented to understand how the model predicts the final results. Finally, the authors developed a website framework and an Android smartphone application in order to predict diabetes instantaneously (Tasin et al., 2023).

In their paper (Madhu et al., 2023), 768 PIMA Indians were used. Standardization, feature selection, missing value filling, and outlier rejection were all parts of the data preparation process.

Machine learning techniques such as logistic regression, decision trees, random forests, the KNN model, the AdaBoost classifier, the Naive Bayes model, and the XGBoost model were used in the study. Accuracy, precision, recall, and F1 score were the only metrics utilized to assess the models' efficacy. The best accuracy (86,61%) was obtained using kNN model.

Models using other databases than PIDD

In 2010 a system for the classification of diabetes patients using SVM was proposed (Yu et al., 2010). The training dataset for classification was taken from the year 1999 by the National Health and Nutrition Examination Survey. They used scheme I classification for predicting undiagnosed or diagnosed diabetes vs. no diabetes or prediabetes and scheme II used for prediabetes or undiagnosed diabetes vs. no diabetes. In the scheme I, family history, race, age, height, weight, hypertension and Basal metabolic index were included as variables. For scheme II they added physical activity and sex as variables.

Sanakal and Jayakumari proposed a system using data mining approach (in MATLAB) which was SVM and Fuzzy C-Means clustering for prediction of diabetes (Sanakal & Jayakumari, 2014). Training dataset was obtained from the UCI repository which comprises nine input attributes and 768 cases. The best outcome obtained by it is a positive predictive value of 88.57% and accuracy of 94.3%. SVM achieved an accuracy of 59.5%.

In 2016 a study used a dataset obtained from the Canadian Primary Care Sentinel Surveillance Network, which includes: systolic blood pressure (SBP), diastolic blood pressure (DBP), HDL, triglycerides (TG), BMI (Body Mass Index), fasting blood sugar (FBS), and gender. In the study Bootstrap aggregating, Adaptive Boosting, and DT model were used. Adaboost can be applied to predict diseases like diabetes, coronary heart disease, and hypertension (Perveen et al., 2016).

Olivera et al. proposed models using data from the Longitudinal Study of Adult Health (ELSA-Brasil). The authors used artificial neural network, logistic regression, naïve Bayes, K-nearest neighbor and random forest as machine-learning algorithms. The best models were those based on artificial neural networks and logistic regression (Olivera et al., 2017).

In their paper, Mujumdar and Vaidehi proposed a model for better classification of diabetes which includes few external factors, responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. The classification has been done using various algorithms (Logistic Regression with highest accuracy of 96%, AdaBoost classifier as best model with accuracy of 98.8%, Random Forest Classifier with accuracy of 98.1%, Extra Trees Classifier with accuracy of 96.3% and Linear Discriminant Analysis with accuracy of 95%) (Mujumdar & Vaidehi, 2019).

Daghistani and Alshammari performed comparison studies on RF (Random Forest) algorithm and LR (Logistic Regression) algorithm towards the prediction of diabetes (Daghistani & Alshammari, 2020). Dataset used for the study was from the Ministry of National Guard Health Affairs (MNGHA) hospital's database from three regions of Saudi Arabia. The accuracy of the RF algorithm was 88%, better than LR technique, whose accuracy was 70.3%.

Rhee conducted a study on the basis of the National Health Insurance Service-Health Screening (NHIS-HEALS) cohort of Korea. Overall, 335,302 subjects without T2DM at baseline were included. A model was trained based on 80% of the subjects, and tested in the remainder. Predictive models for T2DM were constructed using the recurrent neural network long short-term memory (RNN-LSTM) network and the Cox longitudinal summary model. The performance of both models over a 10-year period was compared using a time dependent area under the curve. The annual performance of the model created using the RNN-LSTM network was superior to that of the Cox model, and the risk factors for T2DM, derived using the two models were similar. (Rhee et al., 2019).

Islam et al. used the NB, LR and RF algorithm, after applying 10-fold cross-validation and percentage split evaluation techniques. The dataset contained 520 records. The best accuracy (99%) was obtained using the RF algorithm (Islam et al., 2020).

Malik et al. developed a framework, implementing NB, BayesNet, DT, RF, AdaBoost, Bagging, kNN, SVM, LR, and Multi-Layer Perceptron. Experimental results procured for the Frankfurt Hospital (Germany) dataset shows that kNN, RF, and DT were the best algorithms in terms of all metrics (Malik et al., 2021).

Beghriche et al. proposed a model based on Deep Neural Network (DNN). The patients were selected from the Hospital of Frankfurt, Germany, with the following features: Pregnancies, Glucose, Diastolic blood pressure, Triceps skinfold thickness, Patient insulin in the blood, Body mass index, Diabetes Pedigree Function, Patient age, Outcome (Presence or absence of diabetes). The obtained results provide promising performances with an accuracy of 99.75% and an F1-score of 99.66% (Beghriche et al., 2021).

A study (Al-Gharabawi & Abu-Naser, 2023) employs machine learning to predict diabetes using a Kaggle dataset with 13 features. Their three-layer model achieved an accuracy of 98.73% and an average error of 0.01%. Feature analysis identifies age, gender, polyuria, polydipsia, visual blurring, sudden weight loss, partial paresis, delayed healing, irritability, muscle stiffness, alopecia, Genital thrush, weakness, and obesity as influential predictors.

Methodology

In our research method, we used a dataset, taken from Kaggle, men and women, with overall different features from those of PIDD (some features of PIDD can also be found in the database used in our study): gender (1 for men and 2 for women), age (integer), hypertension (0 for no hypertension and 1 for hypertension), heart_disease (1 for heart disease and 0 otherwise), smoker_status (0 for no smoker, 1 for former smoker, 2 for smoker and 3 for no current smoker), bmi (body mass index) (numeric), HbA1c_level (numeric) and blood_glucose_level (numeric). We kept only the records with no missing values, so in the final database contains 64184 records.

Each patient is classified with diabetes or not (feature named diabetes, whom values are 1 for diabetes – 7046 cases and 0 for the rest). The first 5 records from the database are shown in Figure 1.

gender	age	hypertension	heart_disease	smoker_status	bmi	HbA1c_level	blood_glucose_level	diabetes
2	36	0	0	1	23.45	5	155	0
1	76	1	1	1	20.14	4.8	155	0
1	40	0	0	1	36.38	6	90	0
2	41	0	0	1	22.01	6.2	126	0
1	50	1	0	1	27.32	5.7	260	1

Figure 1. First five records of dataset

Source: Authors' research (from csv file).

We write the application using Python, in which the following modules were imported: pandas (for data analysis), numpy (for performing the numerical calculation) and other modules required for each proposed algorithm.

Dataset was split, in all the models, into train set (70% from the whole dataset) and test set (30% from the whole dataset), by using train_test_set() function from sklearn.model_selection module.

For all the algorithms presented above, we evaluate the performance by showing confusion matrix() and accuracy_score() (imported from sklearn.metrics module).

Results and discussions

Decision Tree (DT) Algorithm

The DT model is based on DecisionTreeClassifier() function (imported from sklearn.tree), using "entropy" as criterion parameter.

In Figure 2 we visualize the decision tree in order to better understand how the model learned to classify patients as diabetic or not. The most influential predictors are hemoglobin A1C level and glucose level, but also age has its role. These findings are consistent with the diagnostic criteria illustrated in Table 1.

We obtain an accuracy score of 0.96, and the confusion matrix has the following form:

(17122 0 721 1413)

The results are very satisfactory, with no false positive values. This model and the SVM model are the only one, from all the proposed models, which have no false positive values.

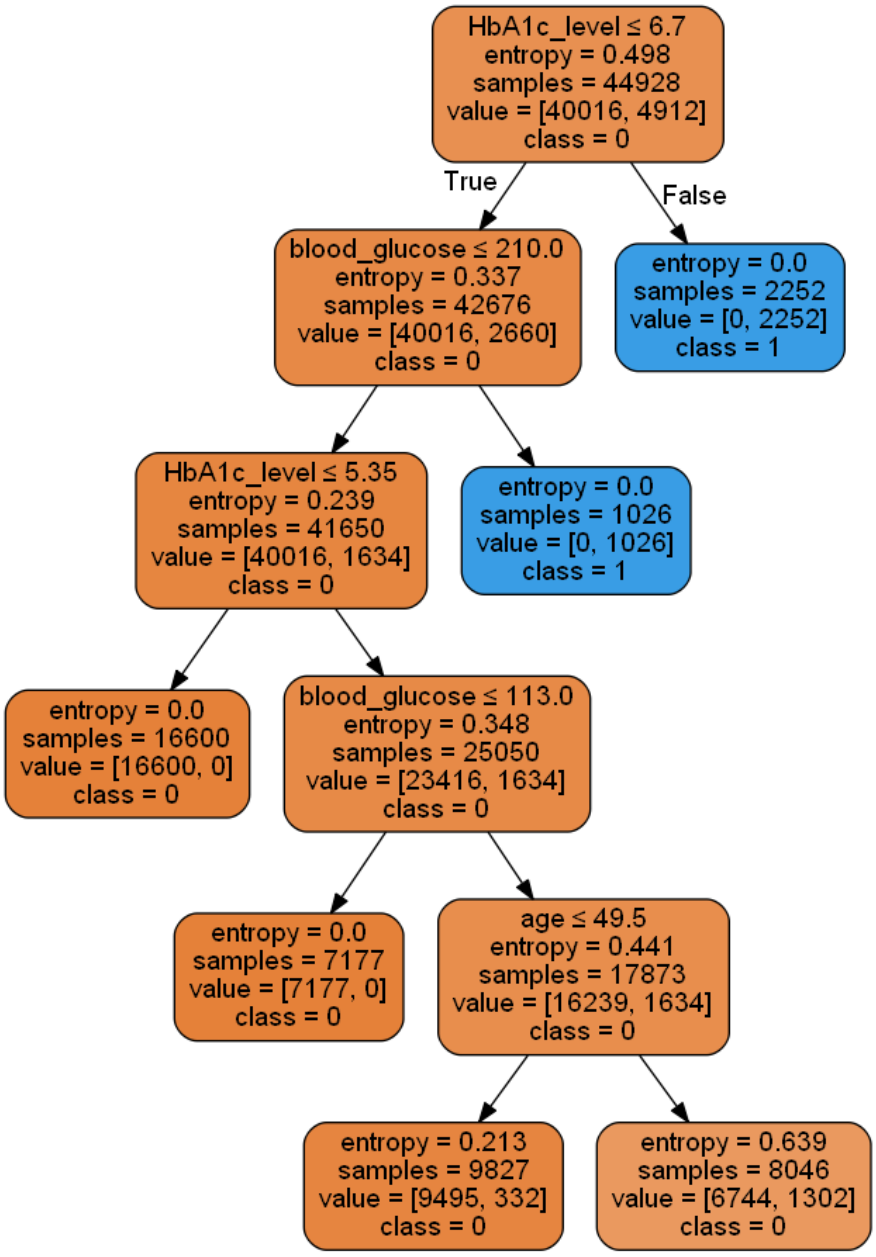

Figure 2. Decision tree chart

Source: Authors' program in Python.

Logistic Regression (LR) Algorithm

We imported the Logistic Regression module from sklearn.linear_model. The coefficients were computed by using coef_ property.

The model obtained an accuracy score of 0.94, and the confusion matrix is:

(16940 182 813 1321)

The contributions of each parameter to the logistic regression equation are shown in Figure 2, obtained by using pyplot module in Python.

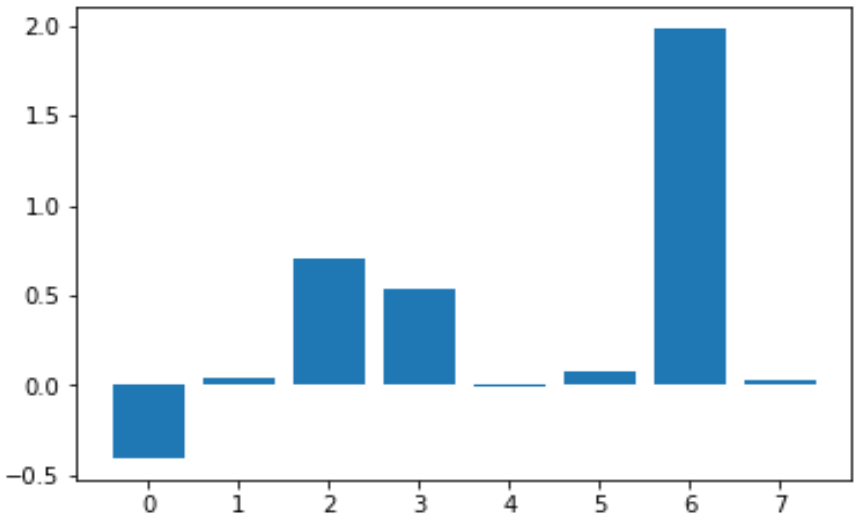

Figure 2. The coefficients of the logistical regression equation

Source: Authors' program in Python.

The values for each coefficient are:
Feature 0 (gender), Score: -0.40354
Feature 1 (age), Score: 0.04431
Feature 2 (hipertension), Score: 0.70356
Feature 3 (heart_disease), Score: 0.53448
Feature 4 (smoker_status), Score: -0.00104
Feature 5 (bmi), Score: 0.07962
Feature 6 (HbA1c_level), Score: 1.98155
Feature 7 (blood_glucose_level), Score: 0.03044

The very low value of the coefficient for blood glucose level, which is one of the determining parameters of diabetes, according to medical guidelines (Table 1), as well as the importance of the coefficients of hypertension and heart disease in the logistic regression equation, it is apparently surprising, but the models learn by trying to find connections between the various characteristics, these connections being able to lead, if they are validated in time, even to their standardization in medical guidelines.

Support Vectors Machines (SVM) Algorithm

This model imports svm module and it's based on SVC() function, whose parameter 'kernel' can have a value from several options. There are various types of functions: linear ('linear'), polynomial ('poly'), sigmoid ('sigmoid') and radial basis function ('rbf'). We ran the program that implements the SVM algorithm by giving, one by one, the following values to the kernel parameter: 'poly', 'sigmoid' and 'rbf'.

The model using 'poly' kernel obtained an accuracy score of 0.93 and the confusion matrix has the following values:

(17067 55 1104 1030)

The model using 'sigmoid' kernel obtained an accuracy score of 0.78 and the confusion matrix has the following values:

(15155 1967 2134 0)

The model using ‘rbf’ kernel obtained an accuracy score of 0.93 and the confusion matrix has the following values:

(17122 0 1337 797)

The best accuracy score of 0.93 is obtained both for ‘poly’ kernel and ‘rbf’ kernel, but in the case of 'rbf' kernel there are no false positive values, but there are more false negative values and the true negative values are less than in the case of the 'poly' kernel.

k-Nearest Neighbors (kNN)

The algorithm is based on the KNeighborsClassifier() function. The ‘euclidean’ distance was chosen as a metric. Choosing ‘K’ is crucial. It represents the number of neighbors considered. KNN is a lazy learning algorithm, meaning it doesn’t update distances with every calculation to save computational resources. We programmatically searched for the optimal value of K. The variation of the accuracy score in relation to K (the number of neighbors) is shown in Figure 3. The optimum value for accuracy score (0.9391) is obtained for K=4, in which case the confusion matrix has the following form:

(17028 94 1077 1057)

For the default value of K=5, the accuracy score is 0.9378 and the confusion matrix has the following form:

(16932 190 1007 1127)

For K=137, which represents the square root of total numbers of records used for test, the accuracy score is 0.9310 and the confusion matrix has the following form:

(17119 3 1325 809)

The variations of the accuracy score values are still small for different values of K and we can estimate that the kNN model has an accuracy score of 0.93, which is a very good one.

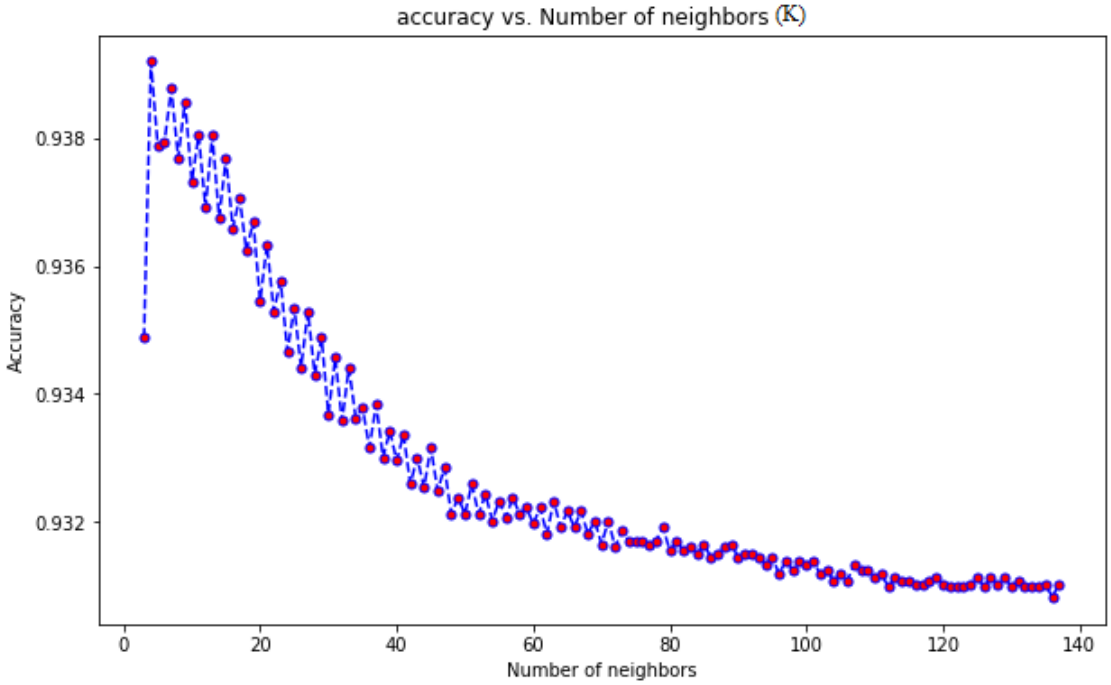

Figure 3. The accuracy score related to K between 3 and 140

Source: Authors’ program in Python.

As can be seen from the performance evaluation of the presented models, the best accuracy score is obtained by the model based on the Decision Tree, but, overall, all models are very satisfactory. In Table 2, the best performances of each model are systematized.

Table 2. Accuracy score for each model

	Decision Tree (DT)	Logistic Regression (LR)	Support Vector Machine (SVM)	k Nearest Neighbors (kNN)
Accuracy score	0.96	0.94	0.93	0.93

Source: Authors' own research.

The decision tree (DT) model apply the criteria from the medical guidelines shown in Table 1 in the most appropriate form, but the purpose of a model is, on the one hand, to validate a medical theory (in the present case) and, on the other hand, to propose new approaches between the various characteristics used as input data.

Conclusion

The models presented in the paper (based on Decision Tree, Logistic Regression, Support Vectors Machine and k-Nearest Neighbors algorithms, respectively) give very satisfactory results (accuracy scores are, for all the models, above 0.93).

The features in the dataset are based on only two parameters determined by usual and cheap medical analyses (Haemoglobin A1C and serum glucose). Other numerical parameters involved in the presented models are age (in years) and body mass index, which are easily determined by measuring the patient's height and weight. The other parameters required for training and testing the models are categorical variables (hypertension, heart disease, smoking status).

Each proposed model tries to capture connections between the output variable, which achieves the required classification, and the features of the database. These connections, which are sometimes not even suspected by doctors, can lead in time to the standardization of such models. The purpose of any proposed model is to give doctors the opportunity to test the input data for a patient, in order to obtain the classification of diabetes or not, but not always the patient comes to the doctor with the same set of characteristics, hence the usefulness of the diversity of models.

In medicine, the aim of a screening method (a strategy used in a population to identify an unrecognized disease in individuals without signs or symptoms) is to prevent a disease. Measuring the value of HbA1c in the general population, once a year can be a screening test for diabetes mellitus. Therefore, such models, as well as other types of medical alerts, can help medical management.

References

Al-Gharabawi, F.W. & Abu-Naser, S.S. (2023). Machine Learning-Based Diabetes Prediction: Feature Analysis and Model Assessment. *International Journal of Academic Engineering Research (IJAER)* 7 (9), 10-17.

Alehegn, M., Joshi, R.R. & Mulay, P. (2019). Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: an ensemble approach. *Int J Sci Technol Res.*, 8(9), 1346–1354.

Amour Diwani, S. & Sam, A. (2014). Diabetes forecasting using supervised learning techniques. *Adv. Comput. Sci.: Int. J.*, 3(5), 10–18, Retrieved from: http://www.acsij.org/acsij/article/view/156.

Anuja Kumari, V. & Chitra, R. (2013). Classification of diabetes disease using support vector machine", *Int. J. Eng. Res. Appl.*, 3, 1797–1801.

Beghriche, T., Djerioui, M., Brik, Y., Attallah, B. & Belhaouari, S.B. (2021) An Efficient Prediction System for Diabetes Disease Based on Deep Neural Network. *Hindawi Complexity*. Retrieved at: https://doi.org/10.1155/2021/6053824.

Daghistani, T. & Alshammari, R. (2020). Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. *J. Adv. Inf.Technol.*, 11(2), 78-83.

DeFronzo, R.A., Ferrannini, E., Groop, L, Henry, R.R., Herman, W.H., Holst, J.J., Hu, F.B., Kahn, C.R., Raz, I., Shulman, G.I., Simonson, D.C., Testa, M.A. & Weiss, R. (2015). Type 2 diabetes mellitus. *Nat Rev Dis Primers*, 1,15019. doi:10.1038/nrdp.2015.19.

Dewangan, A.K. & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. *Int. J. Eng. Appl. Sci.*, 2(5), 145-148.

Islam, M.M.F., Ferdousi, R., Rahman, S. & Bushra, H.Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques, in *Gupta M, Konar D, Bhattacharyya S, Biswas S (eds) Computer vision and machine intelligence in medical image analysis. Advances in intelligent systems and computing*", 992, Springer, Singapore, 113–125. Retrieved from: 10.1007/978-981-13-8798-2_12.

Iyer, A., Jeyalatha, S. & Sumbaly, R. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process (IJDKP),* 5, 1-14. Retrieved from: https://doi.org/10.5121/ijdkp.2015.5101.

Madhu, B., Aerranagula, V., Mahomad, R., Ravindernaik, V., Madhavi, K. & Krishna, G. (2023) Techniques of Machine Learning for the Purpose of Predicting Diabetes Risk in PIMA Indians. E3S Web of Conferences, 011. Retrieved at: https://doi.org/10.1051/e3sconf/ 202343001151.

Malik, S., Harous, S. & El-Sayed, H. (2021). Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women. *Modelling and Implementation of Complex Systems*", Springer International Publishing. Retrieved from: https://www.springerprofessional.de/en/comparative-analysis-of-machine-learning-algorithms-for-early-pr/18351326.

Mujumdar, A. & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, 165, 292–299.

Olivera, A.R., Roesler, V., Iochpe, C., Schmidt, M.I., Vigo. Á., Barreto S.M., Duncan, B.B. (2017). *Sao Paulo Med J.*,135 (3), 234-46.

Perveen, S., Shahbaz, M., Guergachi, A. & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Comput. Sci.*, 82, 115–121. Retrieved from https://dspace.library.uvic.ca/bitstream/handle/1828/9390/Keshavjee_Karim_ProcediaComputSci_2016.pdf?sequence=1&isAllowed=y

Purnami, S.W., Embong, A., Zainand, J.M. & Rahayu, S.P. (2019). A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis/ *Journal of Computer Science*. 5(12), 1003-1008.

Rhee, S.Y., Sung, J.M., Kim, S., Cho, I.J., Lee, S.E. & Chang, H.J. (2019). Development and Validation of a Deep Learning Based Diabetes Prediction System Using a Nationwide Population-Based Cohort. *Diabetes Metab J.*, 45, 515-525. Retrieved at: https://doi.org/10.4093/dmj.2020.0081.

Sanakal, R. & Jayakumari, S.T. (2014). Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. *Int. J. Comput. TrendsTechnol.*, 11, 94–98.

Sen, S.K. & Dash, S. (2014). Application of Meta Learning Algorithms for the Prediction of Diabetes Disease. *International Journal of Advance Research in Computer Science and Management Studies*, 2, 396-401.

Sisodia, D. & Sisodia, D.S. (2018). Prediction of diabetes using classification algorithms. *Procedia Comput. Sci.* 132, 1578–1585.

Soliman, O.S. & AboElhamd, E. (2014). *Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine*. Retrieved from arXiv:1405.0549.

Sridar, K. & Shanthi, D. (2014). Medical diagnosis system for the diabetes mellitus by using back propagation-Apriori algorithms. *J. Theor. Appl. Inf. Technol.*, 68(1), 36-43.

Tasin, I., Ullah, T., Sanjida, N. & Khan, I.R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthc. Technol. Lett.*, 10, 1–10. DOI: 10.1049/htl2.12039.

Tigga, N.P. & Garg, S. (2019). Predicting type 2 Diabetes using Logistic Regression. *Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems MCCS*, Lecture Notes of Electrical Engineering, Springer.

Yu, W., Liu, T., Valdez, R., Gwinn, M. & Khoury, M.J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med. Inform. Decis. Mak.* 10(16). Retrieved from doi:10.1186/1472-6947-10-16.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. & Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontiers in genetics*, 9, 515. Retrieved from https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full.