Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2835-2841 | ISSN:2251-6727

Prediction of Diabetes Gene data Using Machine Learning Techniques

K.Reka[1] , Raja G[2] , U.Srinivasulu Reddy[3]

[1]Department of Computer Science, Cauvery College for Women, Trichy India

[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

[3]Department of Computer Applications, National Institute of Technology, Trichy India

(Received: 07 January 2024 Revised: 12 February 2024 Accepted: 06 March 2024)

KEYWORDS

T2D, Gene, Diabetes, Machine learning, Classifiers

ABSTRACT: Diabetes complications significantly affect patients' quality of life. Diabetes and the health problems associated with it can be prevented by early diagnosis and treatment. The purpose of this study is to determine the risk of diabetes among those who need treatment to prevent diabetes and its associated health problems. According to their lifestyle and family background, this study evaluates the risk of diabetes among individuals. Type 2 Diabetes (T2D) is often detected too late in its clinical course with many patients presenting with complications of unrecognised T2D at the time of diagnosis. This paper provides an overview of several supervised machine learning omics T2D gene categorization methods important theoretical question, pointing the researcher in fascinating new paths for study, and recommending possible combinations of biases that haven't been investigated yet.

1. Introduction

Diabetes mellitus is most caused by type 2 diabetes. Diabetes-related complications, such as cardiovascular disease, kidney disease, retinopathy, and neuropathy, can be increased by long-term hyperglycemia[1]. Basically, diabetes is caused by the pancreas not being able to produce sufficient insulin. The body fails to utilize insulin produced by the cells and tissues.

In Diabetes Mellitus Type 1, the pancreas produces insufficient insulin to meet the body's needs, which is also called insulin-subordinate diabetes mellitus (IDDM). Type-1 diabetes patients require external insulin doses to compensate for their pancreas' reduced production of insulin.

Diabetes Mellitus Type-2 is characterized by the body's resistance to insulin, which is caused by a different reaction between cells and insulin. Eventually, the body may be unable to produce insulin. Adults with diabetes are sometimes referred to as individuals with "non-insulin-subordinate diabetes mellitus" (NIDDM). An inactive lifestyle or a high BMI are common causes of this type of diabetes[2].

A systematic review analyzed the relationship between medication nonadherence and health outcomes in the elderly [3]. They found that medication nonadherence could be significantly associated with all-cause hospitalization and mortality in the elderly. We studied T2D patients instead of seniors, and we predicted that HbA1c and diabetic complications would be the outcome of T2D. An author explored in a cross-sectional study the significant predictors of poor adherence in T2D patients and the factors affecting compliance[4]. For predicting nonadherent T2D complications and HbA1c risk factors, statistical and machine learning methods were used. It was a quite different research method and outcome. A new research idea about what influences T2D progression and what predictive models can be developed was presented.

A few ML models that can forecast the unfavourable effects of nonadherent T2D. The local healthcare systems will be used in this study to forecast any negative effects of non-adherent T2D. Determining the predictors of complications and HbA1c was the main goal of this investigation, which also aimed to create and assess prediction models of diabetes complications and poor glycemic control (defined as haemoglobin A1c (HbA1c)

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2835-2841 | ISSN:2251-6727

7%) among nonadherent T2D patients. Ultimately, it sought to offer clinical practise risk prediction tools.

The rest of this paper is organized into Sections detailed as follows: section 2 Related work, data description in section 3, Prediction in section 4, results in section, section 5 conclusion

2. Related Work

The global health challenge of diabetes is one of the most chronic diseases [5]. When diabetes is uncontrolled, complications can occur. The patient may suffer from a diabetic coma or suffer from a heart attack or stroke as a result. Diabetes affects 422 million people worldwide, which is about 17% of the population [6]. Diabetes complications account for more than 68% of all diabetes-related deaths [7]. There are about 25% of Americans without diabetes diagnosis [8]. Identifying and monitoring patients at risk of diabetic complications is essential to addressing complications [9] and [10]. In order to slow the diabetes epidemic, early detection of diabetes-related complications can prevent or delay complications and allow for effective interventions at both the individual and population levels.

Most people with diabetes mellitus (90–95%) have type 2 diabetes[11] and [12]. The effects of long-term hyperglycemia on diabetes complications may include cardiovascular disease, kidney disease, retinopathy, and neuropathy [13]. Individuals with type 2 diabetes and its complications suffer harsh effects on their health and finances, and the health-care system suffers heavy burdens [14]. Generally, these complications are related to the degree of glycaemic control and the duration of diabetes. It has been shown that intensive glucose control in the early stages of T2D can reduce chronic complications of diabetes [15]. Several guidelines and principles have been developed for managing glycemic control and preventing long-term complications of T2D [16]. However, high therapy adherence is essential for effective treatment of T2D. It refers to how closely a person follows the recommendations of their healthcare provider when taking medication, monitoring indicators, or following a diet.

In certain studies, doctors failed to monitor glycemia regularly nor intensify treatments in a timely manner [17]. The identification of potential adverse outcomes associated with nonadherence should be an urgent priority for individualized treatment of T2D [18]. Therefore, it was necessary to establish a prediction model that could predict the prognosis of nonadherent T2D.

Tigga et al. applied logistic regression to the PIDD for the purpose of predicting the presence of diabetes and discovered that, out of all the parameters in the PIDD, the number of pregnancies, BMI, and glucose level are the most important factors for the diagnosis [19].

Model migration is used in a rapid model for glucose identification and prediction [20]. In many studies, even though data were collected from different sources and places, some attributes such as age, gender, body mass index (BMI), glucose, blood pressure, and time of diagnosis were used.

Glycemic control factors were assessed using a logistic regression model. Based on the model, patients older than 65 are more likely to experience complications than younger patients [21]. The logistic regression method was used to predict complications based on demographic and treatment data [22]. A study comparing 30-day complication rates was conducted using random forest and logistic regression, with age being the most significant factor in predicting complications [23]. Random forest and simple logistic regression methods showed the most effective performance compared to the evaluated algorithms [24].

Han Wu et al. in [25] used data mining techniques (such as enhanced kNN and logistic regression) to properly predict a person's probability of acquiring type 2 diabetes up to 95.42% of the risk.

A similarity measure was used to predict diabetes complications. First, they assessed the similarity between textual medical records after data cleaning, then topic mining is conducted, and last building the model [26].

3. Data Description

The National Health and Nutrition Examination Study (NHANES) is an annual cross-sectional study conducted by the American Centre for Disease Control and Prevention (CDC) [27]. This study examines the health and nutrition of a nationally representative sample of 5000 non-institutionalised, civilian people living across the US

4. Prediction of T2D Using Machine Learning Techniques

Diabetes-related research using gene data is increasingly using machine learning and data mining algorithms

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2835-2841 | ISSN:2251-6727

shown in fig (e.g., relationship between periodontitis and T2D and diabetes-related adverse medication impact). These studies have mostly concentrated on using T2D -related gene data mining for clinical reasons; for example, one such study sought to predict clinical risk of diabetes using gene dataset. Our work, which tries to find more situations and controls, differs from the preceding work in terms of its motivation and intended use. Selecting the particular learning algorithm to be used is a crucial first step. As soon as initial testing is judged sufficient, the classifier—which assigns labels to unlabeled cases—can be employed again. The classifier is usually assessed using prediction accuracy, It measures how many accurate forecasts there are out of all the predictions. To evaluate a classifier's accuracy, at least three methods are employed. Two thirds of the training set may be used for training, and the remaining third might be used for performance estimation. An alternative method called cross-validation is used to split the training set into equal-sized, mutually exclusive subgroups. The classifier is then trained using the union of all other subsets for all subsets. Thus, an approximation of the classifier's error rate can be obtained by averaging the error rates of each subset. Another type of cross-validation is leave-one-out validation. There is only one occurrence in each test subset. This kind of validation is helpful when the most precise estimate of a classifier's error rate is needed, but it inevitably requires more computing resources.

Since there is often just one size N dataset accessible, all calculations must start with this one dataset. Subsampling yields distinct training sets; cases not sampled for training are used for testing. Unfortunately, this is not in accordance with the independence concept that is required for appropriate significance testing.

An introduction of decision tree research and an example of its applicability to novices and experts in machine learning[28]. Decision trees use feature values to sort examples in order to classify them based on their feature values, instances are sorted and classified, beginning at the root node. Theorists have looked for effective heuristics for creating nearly-optimal decision trees because the task of creating ideal.

Decision tree induction techniques can prevent over fitting training data by using two typical approaches. Trim the decision tree that was induced. Preference frequently goes to the tree with fewer leaves if both use the same tests and have the same forecast accuracy. Post-pruning approaches are commonly used by decision tree classifiers to assess the decision trees' performance after they have been pruned using a validation set. Any node can be removed if it is sorted into the training instance class that is more common. A comparative analysis of common pruning approaches did not identify a single optimum technique. Post-pruning approaches are commonly used by decision tree classifiers to assess the decision trees' performance after they have been pruned using a validation set. Any node can be removed if it is sorted into the training instance class that is more common[29].

When solving issues requiring diagonal partitioning, most decision tree techniques are not very effective. The model computes new features by combining the previous ones in a linear fashion. The replication problem means that decision trees can represent some notions in a much more complicated way. One way to prevent duplication is to use an algorithm to incorporate complex characteristics at nodes.

Even more straightforward, ZeroR forecasts the average value (if numeric) or majority class (if nominal) of the test data. The elementary covering algorithm for rules is implemented by prism. A portion uses partial decision trees to obtain rules. Using the same user-defined parameters as J48, it constructs the tree using the heuristics from C4.5.Learning set of rules. Nevertheless, a range of rule-based techniques can also be used to directly induce rules from training data. Furnkranz gave a great summary of the state of rule-based method research.

Dividend-and-conquer algorithms, also called covering algorithms, aim to find a rule that at least partially explains their training cases. They adopt new rules to recursively conquer the remaining examples when these instances are divided, and they keep doing this until there are no more instances. Regarding rule-followers, it provides a general pseudo-code. Rule learning heuristics are not the same as decision tree heuristics in that the former only evaluate the quality of the set of instances that the candidate rule covers, while the latter evaluate the average quality of many disconnected. These checks can be either in the form of stopping the specialization process by using a quality metric or by generalizing rules that are overly specialized in a separate pruning phase.

Lazy classifiers

Idle learners hoard training instances and wait until classification time to put in any meaningful work. The IB1 instance-based learner is a fundamental model that identifies the training instance that is nearest in Euclidean distance to the test case and predicts the class of that training instance. When there are numerous examples that are the closest, the first one found is used. The k-nearest-neighbor classifier, or IBk, uses the same

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2835-2841 | ISSN:2251-6727

distance metric. In the object editor, the number of nearest neighbors can be directly defined or automatically determined using leave-one-out cross-validation (k = 1 by default). Subject to the maximum amount indicated by the supplied value. Two distinct methods are implemented to transform the distance into a weight, so that predictions from several neighbors can be weighted based on how far away they are from the test instance.

Combining Classifiers

Combining classifiers that use stacking for regression and classification issues is known as stacking. The number of cross-validation folds, the base classifiers, as well as the meta learner, are all specified. A more effective variation is implemented by stacking C, where the meta-learner needs to be a numerical prediction system.

5. Result and Discussion

The suggested approach was created with the intention of classifying gene individuals with Type 2 Diabetes. On both available and chosen characteristics, the effectiveness of several machine learning prediction models for the Type 2 Diabetes was evaluated.

Several fields can benefit from the application of supervised machine learning techniques. Regarding the application of each technique, we give our conclusions. This work is not intended to discuss the merits and drawbacks of individual algorithms or empirically compare different bias possibilities; the selection of an algorithm is always dependent on the particular task at hand. But still, The following observations are intended to assist practitioners in avoiding choosing an algorithm that is completely unsuitable for their task. The table1 is KStar and Decision Stump algorithms have positive average correlation coefficients, while the table 2 and table 3 REPTree and table 4 ZeroR algorithms have negative average correlation coefficients. With a mean absolute error of 0.04106 shown in figure 1and 2, the Decision Stump algorithm has a low mean. In the ZeroR algorithm, the highest mean absolute error is concurrently present shown in figure3 and 4 . With an average build time of 0.02 seconds, the ZeroR algorithm is the fastest when it comes to processing time in table 5. Decision Stump is the slowest.

Dataset	Time Taken to Built Model	Cor-Relation Co-effecient	Mean Absolute Error
lc4b035ex2	#########	0.8933	0.08
lc4b036ex2	0.59	0.8552	0.0089
lc4b041ex2	0.48	0.9032	0.0883
lc4b050ex2	0.59	0.8991	0.0877
lc4b053ex2	0.56	0.8866	0.0464
lc4b056ex2	0.51	0.797	0.0204
lc4b057ex2	0.47	0.83	0.0126
lc4b059ex2	0.47	0.8631	0.0167
lc5b040ex2	0.36	0.9094	0.0282
lc5b093ex2	0.52	0.8963	0.0214

Table 1. KSTAR

Dataset	Time Taken (Sec)	Cor-relation Co-effecient	Mean Absolute Error
lc4b035ex2	0.02	0.7562	0.0855
lc4b036ex2	0.05	0.5575	0.009
lc4b041ex2	0.01	0.7689	0.0968
lc4b050ex2	0.02	0.7576	0.0986
lc4b053ex2	0.02	0.5896	0.0582
lc4b056ex2	0.02	0.6304	0.0184
lc4b057ex2	0 .00	0.5567	0.0136
lc4b059ex2	0.02	0.5182	0.0184
lc5b040ex2	0.02	0.59	0.0396
lc5b093ex2	0 .00	0.5607	0.0253

TABLE 2. DECISION STUMP

Dataset	Time Taken to Built Model	Cor-relation Co-effecient	Mean Absolute Error
lc4b035ex2	0.27 seconds	0.0114	0.1936
lc4b036ex2	0.23 seconds	0.0303	0.0187
lc4b041ex2	0.22 seconds	0.0019	0.2257
lc4b050ex2	0.2 seconds	0.026	0.2154
lc4b053ex2	0.28 seconds	0.0315	0.1061
lc4b056ex2	0.24 seconds	0.0331	0.0419
lc4b057ex2	0.23 seconds	0.0345	0.0264
lc4b059ex2	0.25 seconds	0.0335	0.0352
lc5b040ex2	0.15 seconds	0.0295	0.0789
lc5b093ex2	0.14 seconds	0.0156	0.0524

TABLE 3. REPTREE

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2835-2841 | ISSN:2251-6727

TABLE 4. ZEROR

Dataset	Time Taken to Built Model(Sec)	Cor-relation Co-effecient	Mean Absolute Error
lc4b035ex2	0.01	0.0223	0.1937
lc4b036ex2	0.02	-0.0315	0.0188
lc4b041ex2	0.03	-0.0147	0.2257
lc4b050ex2	0.03	-0.029	0.2154
lc4b053ex2	0.02	-0.0335	0.1062
lc4b056ex2	0.02	-0.0367	0.0419
lc4b057ex2	0.02	-0.0391	0.0264
lc4b059ex2	0.03	-0.034	0.0353
lc5b040ex2	0.02	-0.0358	0.0789
lc5b093ex2	0 .00	-0.0216	0.0523

TABLE 5. Comparative table

Algorithm	Time Taken to Built Model	Correlation Co-efficient	Mean Absolute Error
KSTAR	3	2	3
DECISION STUMP	1	3	3
REPTREE	2	1	1
ZEROR	3	1	1

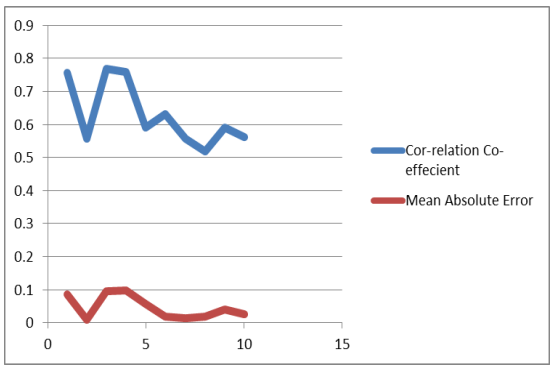

Figure 1. K-Star for Mean Error

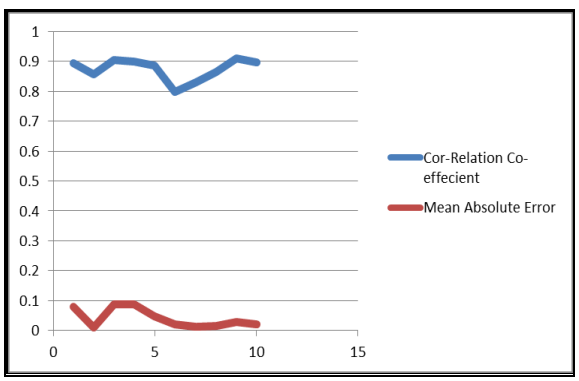

Figure 2. Decision Tree for Mean Error

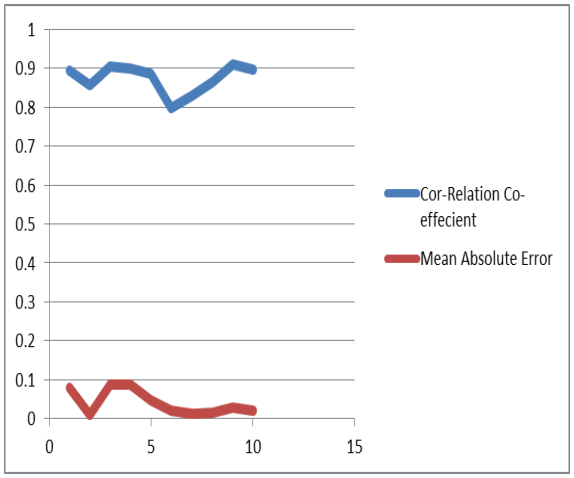

Figure 3. RepTree for Mean Error

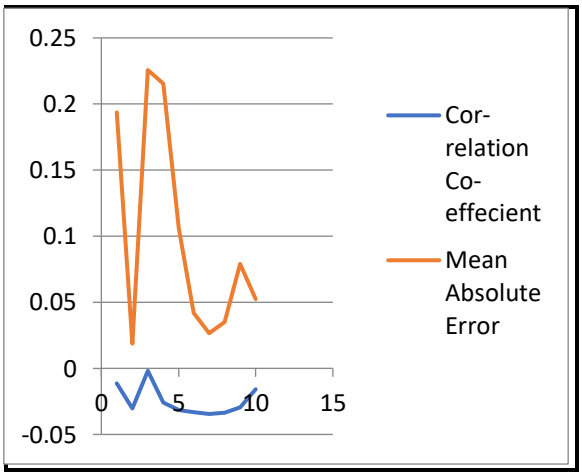

Figure 4. Zeror for Mean Error

6. Conclusion

This may hold promise as a means of creating an e-alert system within gene to aid with the timely recognition and appropriate management of T2D. However, further testing with sparse real-life clinical data and ultimately, clinical trials are required for a more robust assessment of such a system.

The most well-known supervised approaches are covered in considerable detail in this study direction is being pursued by meta-learning In order to achieve this, meta-learning looks for connections between the performance of learning algorithms and a set of variables, referred to as meta-attributes, which represent the qualities of learning tasks. ML working with desktop platform algorithms that runs in a few minutes or seconds, as well as flat files. According to these researchers, the range of "very large" datasets starts at

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2835-2841 | ISSN:2251-6727

100,000 cases with 20 attributes. But gigabyte databases are what the database community works with. Naturally, it is improbable that a data warehouse's whole contents would be mined at once. Since all data must live in main memory, the majority of the learning techniques used today are computationally expensive and unworkable for a large number of practical issues and databases.

References

[1.] Li, W., Lin, L., Yan, D., Jin, Y., Xu, Y., Li, Y., ... & Wu, Z. (2020). Application of a pseudotargeted MS method for the quantification of glycated hemoglobin for the improved diagnosis of diabetes mellitus. Analytical chemistry, 92(4), 3237-3245.

[2.] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706-716.

[3.] Walsh, C. A., Cahir, C., Tecklenborg, S., Byrne, C., Culbertson, M. A., & Bennett, K. E. (2019). The association between medication non-adherence and adverse health outcomes in ageing populations: a systematic review and meta-analysis. British journal of clinical pharmacology, 85(11), 2464-2478.

[4.] Demoz, G. T., Wahdey, S., Bahrey, D., Kahsay, H., Woldu, G., Niriayo, Y. L., & Collier, A. (2020). Predictors of poor adherence to antidiabetic therapy in patients with type 2 diabetes: a cross-sectional study insight from Ethiopia. Diabetology & Metabolic Syndrome, 12, 1-8.

[5.] J. P. Kandhasamy and S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus," Procedia - Procedia Comput. Sci., vol. 47, pp. 45–51, 2015

[6.] Malik, S., Khadgawat, R., Anand, S., & Gupta, S. (2016). Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. Springerplus, 5, 1-12.

[7.] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

[8.] Anderson, A. E., Kerr, W. T., Thames, A., Li, T., Xiao, J., & Cohen, M. S. (2016). Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study. Journal of biomedical informatics, 60, 162-168.

[9.] Anand, A., & Shakti, D. (2015, September). Prediction of diabetes based on personal lifestyle indicators. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 673-676). IEEE.

[10.] Peddinti, G., Cobb, J., Yengo, L., Froguel, P., Kravić, J., Balkau, B., Tuomi, T., Aittokallio, T. and Groop, L., (2017). Early metabolic markers identify potential targets for the prevention of type 2 diabetes. Diabetologia, 60(9), pp.1740-1750.

[11.] Deshpande, A. D., Harris-Hayes, M., & Schootman, M. (2008). Epidemiology of diabetes and diabetes-related complications. Physical therapy, 88(11), 1254-1264.

[12.] Inaishi, J., & Saisho, Y. (2020). Beta-cell mass in obesity and type 2 diabetes, and its relation to pancreas fat: a mini-review. Nutrients, 12(12), 3846.

[13.] Kidanie, B. B., Alem, G., Zeleke, H., Gedfew, M., Edemealem, A., & Andualem, A. (2020). Determinants of diabetic complication among adult diabetic patients in Debre Markos referral hospital, northwest Ethiopia, 2018: unmatched case control study. Diabetes, metabolic syndrome and obesity: targets and therapy, 13, 237.

[14.] Harding, J. L., Pavkov, M. E., Magliano, D. J., Shaw, J. E., & Gregg, E. W. (2019). Global trends in diabetes complications: a review of current evidence. Diabetologia, 62, 3-16.

[15.] Petrie, J. R. (2009). Follow-up of intensive glucose control in type 2 diabetes. New England Journal of Medicine, 360(4), 416-417.

[16.] Jia, W., Weng, J., Zhu, D., Ji, L., Lu, J., Zhou, Z., ... & Chinese Diabetes Society. (2019). Standards of medical care for type 2 diabetes in China 2019. Diabetes/metabolism research and reviews, 35(6), e3158.

[17.] Aujoulat, I., Jacquemin, P., Rietzschel, E., Scheen, A., Tréfois, P., Wens, J., ... & Hermans, M. P. (2014). Factors associated with clinical inertia: an integrative review. Advances in medical education and practice, 141-147.

[18.] Zarkogianni, K., Athanasiou, M., Thanopoulou, A. C., & Nikita, K. S. (2017). Comparison of machine learning approaches toward assessing

Journal of Chemical Health Risks

www.jchr.org

JCHR (2024) 14(2), 2835-2841 | ISSN:2251-6727

the risk of developing cardiovascular disease as a long-term diabetes complication. IEEE journal of biomedical and health informatics, 22(5), 1637-1647.

[19.] Tigga, N. P., & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, 706-716.

[20.] Zhao, C., & Yu, C. (2015). Rapid model identification for online subcutaneous glucose concentration prediction for new subjects with type I diabetes. IEEE Transactions on Biomedical Engineering, 62(5), 1333-1344.

[21.] Almetwazi, M., Alwhaibi, M., Balkhi, B., Almohaini, H., Alturki, H., Alhawassi, T., ... & Alshammari, T. (2019). Factors associated with glycemic control in type 2 diabetic patients in Saudi Arabia. Saudi pharmaceutical journal, 27(3), 384-388.

[22.] YimamAhmed, M., Ejigu, S. H., Zeleke, A. Z., & Hassen, M. Y. (2020). Glycemic control, diabetes complications and their determinants among ambulatory diabetes mellitus patients in southwest ethiopia: A prospective cross-sectional study. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 13, 1089.

[23.] Armstrong, B. N., Renson, A., Zhao, L. C., & Bjurlin, M. A. (2019). Development of novel prognostic models for predicting complications of urethroplasty. World Journal of Urology, 37, 553-559.

[24.] Rodriguez-Romero, V., Bergstrom, R. F., Decker, B. S., Lahu, G., Vakilynejad, M., & Bies, R. R. (2019). Prediction of nephropathy in type 2 diabetes: an analysis of the ACCORD trial applying machine learning techniques. Clinical and translational science, 12(5), 519-528.

[25.] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107.

[26.] Ding, S., Li, Z., Liu, X., Huang, H., & Yang, S. (2019). Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model. Information Sciences, 499, 12-24. https://cks.nice.org.uk/topics/diabetes-type-2/

[27.] Dietterich, T.G. (1998), “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms”, Neural Computation, 10(7), pp. 1895–1924.

[28.] Elomaa, T., Rousu, J. General and Efficient Multisplitting of Numerical Attributes. Machine Learning 36, 201–244 (1999).https://doi.org/10.1023/A:1007674919412.