EDRAAK

Vol. **(2023)**, 2023, **pp**. 27–32
**ISSN:** 3078-8412

Research Article

# Steps of Design and Implementation of a Diabetes Classification System Using Decision Tree and Bayesian Theory

Hadeel M Saleh[1,*]

[1] *Center For Continuing Education*

**ARTICLE INFO**

Article History
Received 12 Jan 2023
Revised: 1 Mar 2023
Accepted 1 Apr 2023
Published 17 Apr 2023

Keywords
Diabetes mellitus,
Classification System,
Decision Trees,
Bayesian Theory.

**ABSTRACT**

An effective tool for the early identification and diagnosis of diabetes mellitus is intended to be developed through the design and implementation of a diabetes categorization system utilizing decision trees and Bayesian theory. To increase classification accuracy and dependability, this system integrates two potent machine learning techniques: Bayesian classifiers and decision trees. While Bayesian theory offers a probabilistic framework that improves decision-making under uncertainty, decision trees provide an understandable way to represent decision-making based on input attributes. The system starts with data preprocessing, which includes feature selection, normalization, and handling missing values. Next, decision tree methods like ID3 or C4.5 are applied to create classification models based on patient attributes. The categorization is then further refined by incorporating prior knowledge and uncertainty using Bayesian techniques, especially Naive Bayes.. Performance parameters like accuracy, precision, recall, and F1-score are assessed while testing the system using publically accessible diabetic datasets. The results show that the hybrid model, which combines Bayesian theory with decision trees, provides a reliable solution for diabetes categorization and offers a promising way for doctors to more accurately diagnose diabetes. Furthermore, real-time data can be added to the system to enable ongoing observation and forecasting.

## 1. INTRODUCTION

Diabetes mellitus is a chronically, lifelong condition which impacts the body's capability of using the energy that exists in different foods. There are 3 main kinds of diabetes: type 1, type 2, and gestational diabetes. Diabetes Mellitus happens in the case where the level of the blood 's glucose is increased due to the fact that the body is not capable of using it in the proper way. [1, 2]

Diabetes was the 7th main reason of death in the US in the year of 2010 according to the 69,071 death certificates where this disease was listed as the underlying reason of death. In the same year, diabetes was named as the reason of death in a total of 234,051 death certificates. Diabetes may be underreported as a cause of death. Researchers have shown that only about 35% to 40% of people that have this disease who died had diabetes listed somewhere on the death certificate and about 10% to 15% had it listed as the contributing factor that caused death. [3]

Effects of diabetes have been reported to have a more serious and worsening effect on women than on men due to their survival rate which is lower and inferior quality of life. WHO reports indicate that almost 33% of the women suffering from diabetes have no idea that they do.

The impact of diabetes is unique in the case of mothers due to the fact that the disease is transferred to their yet to be born babies. Strokes, miscarriages, blindness, kidney failure and amputations are just a sample of the complications that spring from the diabetes. For the reasons of this thesis, the analyses of cases of diabetes have been restricted to pregnant women. In general, an individual is considered to be a patient of diabetes, when the levels of blood sugar exceed the normal level (4.4 to 6.1 mmol/L) [4].

### 1.1 Types of diabetes

The most three widely known types of diabetes are as following :

- Type 1: Previously known as insulin-dependent diabetes mellitus (IDDM) or juvenile-onset diabetes, may account for 5% to 10 % of every one of the diagnosed diabetes cases.

*Corresponding author email: Haddeel.mohammed@uoanbar.edu.iq
DOI: https://doi.org/10.70470/EDRAAK/2023/006

- Type2: Previously known as not insulin-dependent diabetes mellitus (NIDDM) or adult diabetes, it develops in a gradual manner. Over time, the patient‘s body becomes less able to use the insulin, or it begins to produce less amounts of insulin. This type is caused by a set of factors, such as genetic factors and lifestyle habits.
- Type3 (gestational diabetes): it is a kind of diabetes which occurs, or is initially diagnosed, during the pregnancy. The condition, such other forms of diabetes, involves high levels of blood sugar.

Often, this type of diabetes is a temporary issue that happens during the 2nd trimester of pregnancy, and disappears post giving birth.

"Even if a woman needed quite a bit of therapy and treatment for the sake of keeping her blood sugars levels under control during the pregnancy usually the day that follows giving birth, [her] sugars go back down to their normal levels" Nevertheless women who've had this type of diabetes must be monitored regularly post giving birth, due to the fact that they‘re more susceptible to developing diabetes later in their life, based on researches by the National Institutes of Health (NIH).

According to the NIH Digestive and Kidney Diseases (NIDDK), risk factors for gestational diabetes including:

- Being overweight or obese.
- Being pre diabetic.
- Previously giving birth to a child that weighs over 9 pounds.

Women that have gestational diabetes usually don‘t show symptoms or mild, non-life-threatening symptoms, based on reports from the NIH. Tests for this type of diabetes are usually held about 24 to 28 weeks of pregnancy. Some women can be examined even earlier during the pregnancy in case that they‘re at higher risk to get gestational diabetes. At first, women undergo a screening test for glucose, where they drink a sugar solution, and the levels of their blood sugar are tested an hour later. In the case where the level of a woman's blood sugar exceeds the normal, they will have to undergo a 2nd test, known as glucose tolerance test. [4].

## 1.2 Feature selection

Selection of the subset is done by a process which is known as feature selection which is used for the construction of the model. In the learning of the machine including its statistics this process is known as the variable selection or also as attribute selection or variable subset selection. This process can also be used for the identification of the attributes which are not in need of which are irreverent in the data which do not contribute for having the accuracy in the predictive model. Feature selection do not mean dimensionality reduction, both methods try to summarize the number of attributes in the dataset, but the reduction of dimensionality process do so by creating new combinations of attributes, while the feature selection process includes and excludes attributes present in the data without changing them [27,28].

There are three main reasons behindthe use of feature selection techniques, these are:

a. Reduces over fitting: Less redundant data means less opportunity to make decisions based on noise.
b. Improves Accuracy: Less misleading data means high opportunity to approach accurate modeling.
c. Reduces Training Time: Less data means that algorithms train faster. Algorithm of the selection of the features has two different classes namely filter method and second is wrapper method .

1- Filter Method: In these types of methods, there is a crisp criterion on all the features including all the subsets of the features which is used in the evaluation of the suitability of the classification. There is dependency of this method on the algorithm which has been shown in (1) [29].

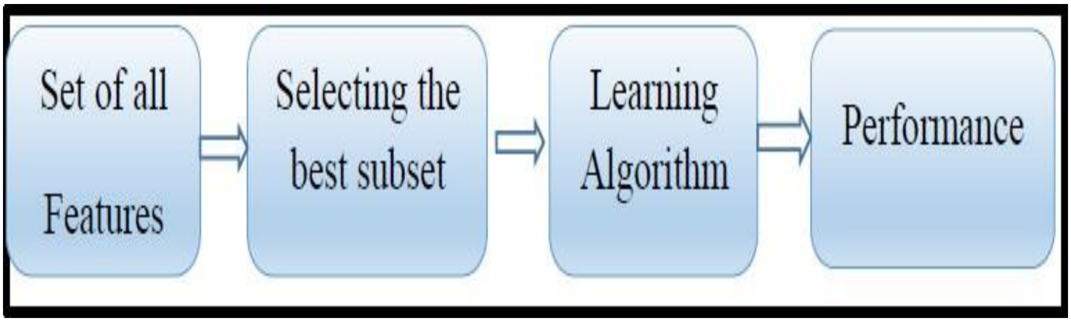

Fig (1) Filter Feature Selection Method

2- Wrapper Method: In this type of methods, the selection of the features has been done by the classification algorithm for making the selection of the feature and all the process sensitive to the algorithm of the classification. This method makes it clear that different features can be useful for better working of the different algorithm. This method is based on the performance of the classifier[29]. Figure (5) illustrates the Wrapper feature selection method.

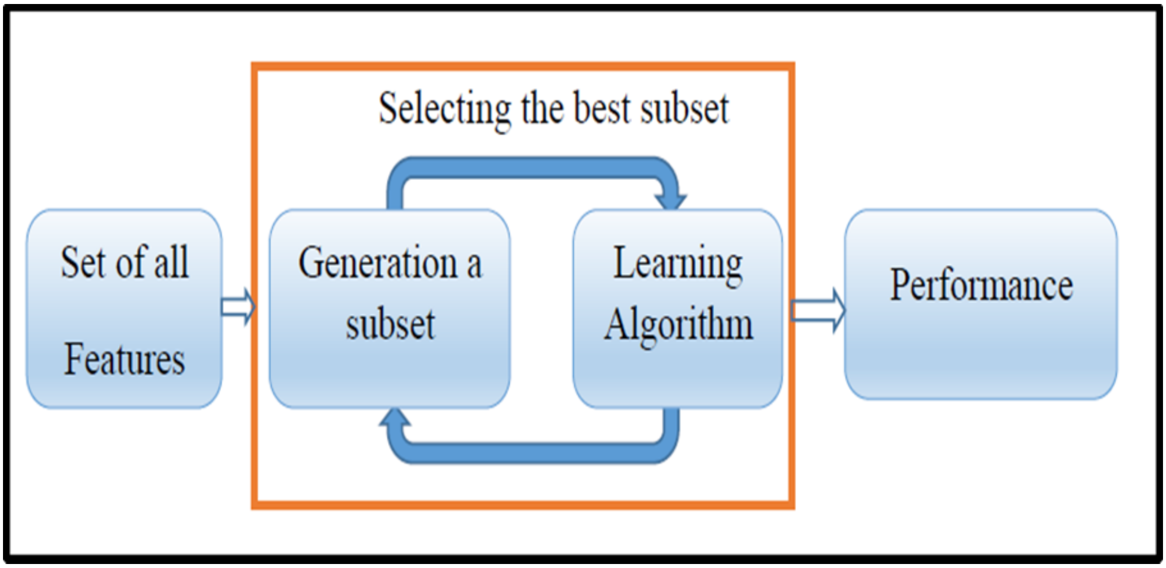

Fig (2) Wrapper Feature Selection Method

a. Evaluation procedures

The goal of structure procedures of evaluation is determining the way of applying specific performance measurements for the sake of obtaining the dependable assessment of the model's expected performance on new data, in other words for determining its generalization characteristics.

This is impossible as long as totally or partially identical data is utilized for both model generation and its evaluating. For several modeling algorithms that would be leading to over-optimistic performance estimates that is sometimes known as evaluation over fitting.

The basic attempt to design evaluation process is therefore for ensuring separating the validation or the test-set Q from the training set T without minimizing the goodness of the model because of non-sufficient training-data [30].

Then after the training is accomplished, the removed data can be utilized for testing the performance of the learned model on the new‖ data. This is the main idea for an entire class of model evaluation approaches known as cross validation.

- Cross-validation method:

It is a widely known approach utilized for estimating the generalization error for constructed models. According to various designs, 2 cross-validation strategies are usually depicted in literature, in other words, exhaustive cross- validation and non-exhaustive cross-validation, the last used in this thesis [30].

3- Exhaustive Cross-Validation

In order to separate the original data set into a training set and a testing set, exhaustive cross-validation techniques attempt to train and test every possible configuration. There are two approaches in this strategy: the Leave-p-out cross-validation method (LpOCV) and the Leave-one-out cross-validation (LOOCV), which is a special case of the Leave-p-out cross-validation method when p = 1. For a data set with N instances, for instance, to use (LpOCV). The leave-p-out cross-validation method chooses the reaming N −p instances as the training set and p instances as the testing set. Until every possible segmentation is covered, the process is repeated. Consequently, the total number of iterations needed for leave-p-out cross-validation is (np). The leave-one-out cross-validation method chooses one instance as the testing set and the remaining N −1 instances as the training set, which is also a computation problem [30]. In contrast, the leave-p-out cross-validation method has a computation problem for hug data sets.

- Non-exhaustive cross-validation

Non-exhaustive cross-validation rules do not have to test every possible data-set segmentation, in contrast to exhaustive cross-validation techniques. Holdout and K-fold cross-validation are the two fundamental procedures used in this approach.

a. K-fold cross-validation, to start.

10-fold cross-validation is the most popular K-fold procedure [30, 31].

For example, there are three phases:

1. Dividing the dataset into K nearly equal parts (folds).
2. Using K − 1 parts as training partitions and the ith portion of all K portions as the test set.
3. Using k time, repeat step 2.

b. Hold-out

In this kind, every fold, data-set randomly separated into 2 groups d1 and d2, therefore, the hold-out evaluation process is thought to be the most accurate approach of the separation of the training and test-data: a sub-set of the existing labeled data-set is chosen in a random way as the training-set and the rest of the instances are held out for the sake of model evaluating. It is popular partitioning the data in a (2:1) division for training and (3:1) division for testing, from rather dependable performance estimates of a rather insufficient structure. Its popular partitioning a single iteration of k-fold cross-validation is equivalent to the hold-out process [32].

4- Performance Evaluation

Accuracy as measured on unknown data (the test-set) is typically very distinctive and may be used to evaluate performances. When the correct classification is unknown, the practical value lies in the accuracy of the unseen data. Using the provided data and assuming that each class membership is known as follows is the most popular method for estimating this. using a sizable portion (the training-set) of the designated data to train the procedure. The remaining data (the test-set) is then used to test this rule, and the results are compared to the established classifications The proportion that is proper in the test-set is an unbiased estimation of the precision of the rule supplied that the training set is sampled in a random way from the specified data depending on the formula (9). [33]

$$\text{Accuracy} = \frac{\text{Number of Correctly Classified data}}{\text{Total Number of data}} \times 100\% \qquad (1)$$

Also, the performance of the system has estimated by using confusion matrix,

TABLE (I): CLASSIFICATION CONFUSION MATRIX

| Classification confusion matrix | | |
|---|---|---|
| | Actual class | |
| **Predicated class** | normal | Abnormal |
| **normal** | (TP) | (FP) |
| **abnormal** | (FN) | (TN) |

A confusion matrix can be broken down as follows [34]:

1. True Positives (TP): this measure shows how many instances were correctly classified as positive, meaning they were initially believed to be successful and ultimately turned out to be so.
2. The number of instances that were mistakenly categorized as positive that is, words that were anticipated to be successful but turned out to be a failure is known as False Positives (FP).
3. The number of cases that were appropriately categorized as negative that is, those that were anticipated to be failures and turned out to be such is known as True Negatives (TN).
4. False Negatives (FN): this is the quantity of cases that were incorrectly classified as negative, meaning that they should have been negative.
5. True Negatives (TN) = it represents the number of cases that were properly classified as negative, i.e., were expected to be a failure and resulted to be such.
6. False Negatives (FN) = it represents the number of cases that were classified incorrectly as negative, i.e., were expected to be negative but actually resulted to be the opposite.

## 2. CONCLUSION

The Diabetes Classification System (DCS) presented in this research effectively addresses the challenge of classifying diabetes cases by leveraging machine learning algorithms, specifically ID3 and Naive Bayesian classifiers. The system showed notable performance, achieving an accuracy of 80% with the ID3 classifier and 91% with the Naive Bayesian classifier. The study demonstrates the importance of preprocessing steps, including discretization, normalization, and feature selection, to improve the quality of data and enhance the classifiers' performance.

The comparative analysis with previous studies highlights the robustness of the proposed system, achieving competitive accuracy levels with reduced computational time. The application of cross-validation methods ensured a reliable evaluation of the model's generalization capabilities. Moreover, the inclusion of performance metrics such as precision, recall, F1 score, and sensitivity further validated the system's effectiveness.

This research underscores the potential of integrating machine learning algorithms into medical diagnostics, particularly for chronic diseases like diabetes. Future work may explore combining multiple classification techniques, expanding datasets, and employing advanced deep learning models to further improve accuracy and applicability in real-world clinical settings.

**Funding:**

This study was not funded by any governmental, private, or institutional grant. All work and expenses related to the study were borne by the authors.

**Conflicts of Interest:**

The authors declare that no conflicts of interest exist in connection with this work.

**Acknowledgment:**

The authors extend their gratitude to their institutions for the invaluable advice and logistical support provided during the research.

## References

[1] D. Care, "Classification and diagnosis of diabetes," *Diabetes Care*, vol. 40, no. Suppl 1, pp. S11–S24, 2017.

[2] World Health Organization, *Global patient safety action plan 2021-2030: towards eliminating avoidable harm in health care*. World Health Organization, 2021.

[3] Centers for Disease Control and Prevention, *National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014*. Atlanta, GA: US Department of Health and Human Services, 2014.

[4] US Department of Health and Human Services, *National Diabetes Information Clearinghouse (NDIC): National diabetes statistics*, 2011.

[5] J. F. Elder and D. W. Abbott, "A comparison of leading data mining tools," in *Fourth International Conference on Knowledge Discovery and Data Mining*, Aug. 1998, vol. 28.

[6] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 3rd ed. 2011.

[7] A. Syed, "Pattern Recognition Using Artificial Neural Network and Other Classifiers," M.Sc. Thesis, Dept. of Computer Science and Engineering, Jadavpur University, Kolkata, India, 2013.

[8] A. Franco and G. S., "ADT: A decision tree algorithm based on concepts," in *International Symposium on Robotics and Automation*, Mexico, Aug. 25–28, 2006.

[9] M. Vasantha, B. Subbiah, and R. Dhamodharan, "Medical image feature, extraction, selection and classification," *International Journal of Engineering Science and Technology (IJEST)*, vol. 2, no. 6, 2010.

[10] M. Vasantha, B. Subbiah, and R. Dhamodharan, "Medical image feature, extraction, selection and classification," *International Journal of Engineering Science and Technology (IJEST)*, vol. 2, no. 6, 2010.

[11] C. Satyanarayana and K. Yedukondalu, "Effect of subsetting of FCC image on the accuracy of the classified image," *International Journal of Electronics and Communication Technology (IJECT)*, 2012.

[12] M. Paulo and B. Ferreira, "Content-based image classification: A non-parametric classification," M.Sc. Thesis, Dept. of Electrical and Computer Engineering, Portugal, 2010.

[13] N. Matthew and G. Sajjan, "Comparative analysis of serial decision tree classification algorithms," *International Journal of Computer Science and Security (IJCSS)*, 2013.

[14] V. Lavanya and U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," *International Journal of Computer Applications (IJCA)*, 2011.

[15] M. Gupta and N. Aggarwal, "Classification techniques analysis," in *NCCI 2010 - National Conference on Computational Instrumentation*, CSIO Chandigarh, India, Mar. 19–20, 2010.

[16] M. Gupta and N. Aggarwal, "Classification techniques analysis," in *NCCI 2010 - National Conference on Computational Instrumentation*, CSIO Chandigarh, India, Mar. 19–20, 2010.

[17] V. Lavanya and U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," *International Journal of Computer Applications (IJCA)*, 2011.

[18] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[19] S. Marwaha and P. Singh, "Development of expert system through AGRIdaksh," Research Institute, Library Avenue, Pusa, New Delhi, 2012.

[20] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed., Chapter 7: Classification and Prediction.

[21] M. Gupta and N. Aggarwal, "Classification techniques analysis," in *NCCI 2010 - National Conference on Computational Instrumentation*, CSIO Chandigarh, India, Mar. 19–20, 2010.

[22] A. Fielding, *ANN Cluster and classification techniques for the biosciences*. Cambridge University Press, 2007.

[23] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, 2004.

[24] M. Green, U. Ekelund, L. Edenbrandt, J. Bjork, J. L. Forberg, and M. Ohlsson, "Exploring new possibilities for case-based explanation of artificial neural network ensembles," *Neural Networks*, vol. 22, pp. 75–81, 2009.

[25] H. S. Khamis, K. W. Cheruiyot, and S. Kimani, "Application of k-nearest neighbor classification in medical data mining," *International Journal of Information and Communication Technology Research*, vol. 4, no. 4, Apr. 2014.

[26] M. Wahde, *Biologically Inspired Optimization Methods: An Introduction*. Chalmers University of Technology, Sweden.

[27] J. Brownlee, *Machine Learning Resource Guide*, 2008.

[28] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[29] Chapman & Hall, *Data Classification Algorithms and Applications*. University of Minnesota, Dept. of Computer Science and Engineering, Minneapolis, Minnesota, USA, 2015.

[30] G. J. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data*. Wiley, 2004.

[31] G. J. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data*. Wiley, 2004.

[32] P. Cichosz, *Data Mining Algorithms: Explained Using R*. Wiley, 2015.

[33] D. Michie and T. Spiegel, "Machine learning, neural, and statistical classification," in *Ellis Horwood*, 1994, pp. 175–212.

[34] C. R. h. Chirag and H. Thiessen, "Key concepts and limitations of statistical methods for evaluating biomarkers of kidney disease," *Journal of the American Society of Nephrology*, vol. 25, no. 8, pp. 1621–1629, Aug. 2014.

[35] C. M. Velu and K. R. Kashwan, "Visual data mining techniques for classification of diabetic patients," in *3rd IEEE International Advance Computing Conference (IACC)*, 2013.