|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | | |  |
| **Iraqi Journal of Statistical Sciences** | | |
| https://ojs.uomosul.edu.iq/index.php/stats/index | | |
|  | | | | |
|  | | | | |
| **Paper Title\* (use style: paper title)** | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
|  | | | | |
| **Article information** | |  | **Abstract** | |
| ***Article history:***  Received xxxxxx  Revised xxxxxx  Accepted xxxxxx  Available online xxxxxx | |  | Diabetes has become one of the most prevalent diseases in Iraq and is listed as one of the leading causes of death. Machine learning provides effective information extraction results by creating predictive models from diagnostic medical datasets collected from diabetes patients in Iraq. In this study, we applied machine learning classification to compare and contrast the performances of classification and regression trees (CART), support vector machines (SVM), random forests (RF), linear discrimination analysis (LDA), and K-nearest neighbors (KNN). We sought to design a model that can predict with maximum accuracy the probability that a person has, is healthy, or is expected to develop diabetes in the future using the two scales of accuracy and kappa. Based on the results obtained from the algorithms, it showed that the accuracy and sequence of the algorithms concerning the training data were Random Forest (RF), Classification and Regression Trees (CART), Support Vector Machine (SVM), Linear Discrimination Analysis (LDA), and K-Nearest Neighbors (KNN). While the test data results showed some differences, the sequence of the algorithms was as follows: SVM, RF, CART, LDA, and KNN were the highest, respectively. The training data set refers to the samples that were used to construct the model, whereas the testing data set is used to evaluate the model's performance. Based on the assessment criteria discussed above, we chose the best machine learning approach to predict diabetes mellitus in Iraq to achieve high performance. All of the strategies listed above are approximated using a supervised diabetes testing dataset. The approach that achieves the maximum performance in terms of accuracy and kappa is regarded as the best option. Based on the results, it can be seen that the SVM and RF algorithms predicted diabetes with more accuracy. | |
| ***Keywords***:  *First keyword; Second keyword; Third keyword; Fourth keyword; Fifth keyword.* | |
| ***Correspondence:*** | |
|  | |  |  | |
| |  | | --- | | DOI [0000000000000](https://stats.mosuljournals.com/article_183258.html) , ©Authors, 2024, College of Computer Science and Mathematics University of Mosul.  This is an open access article under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>). | | | | | |

1. **Introduction (12 Time New Roman)**

International Diabetes Federation (IDF) (2017) data shows that hundreds of millions live with diabetes worldwide. Diabetes now routinely tops lists of the leading causes of death worldwide. Over the past 30 years, based on World Health Organization (WHO) (2018) data, diabetes prevalence has increased rapidly, especially in low- and middle-income countries.1 The International Diabetes Federation (IDF) (2017) reported an 8.8% (425 million people) prevalence among adults in 2017. The Middle East and North Africa (MENA) region has the second highest rate at 9.2%. The MENA region is projected to grow 110 percent between 2017 and 2045, from 329 million to 629 million (IDF) (2017). Diabetes is a significant illness, with a 10.7% death rate in adults aged 20–79. The MENA Region, including Iraq, has the highest rate of fatalities due to diabetes in individuals under 60, ranking second among IDF regions (IDF) (2017). However, Only 2.9% of the world's diabetes investment is directed towards researching the development and consequences of the disease, leaving a significant knowledge gap (WHO) (2018). Iraq faces 1.4 million diabetes cases, but insufficient epidemiological studies and RCTs make it difficult to understand the prevalence and effective therapies for the population as described by Mansour *et al.* (2014).

Khanam and Simon (2021) state that diabetes identification is one of the most difficult challenges in healthcare. Baran (2020) the rapid increase of so-called data sources gives diversity and importance to studies in machine learning. The development of technology has led to the introduction of multi-label classification for increasing datasets.Alan (2020) choosing the optimal classifier is one of the most important difficulties when developing a model in machine learning. Parthiban and Srivatsa (2012) data classification is a typical job in machine learning. Data mining is critical for extracting knowledge from huge datasets. Keskin (2018), Nahzat and Yaanolu (2021), in recent years, several academics have discussed their experiences with various machine learning methods, including Decision Tree, Naive Bayes, Random Forest, and K-Nearest Neighbour Support Vector Machine. Research has shown that machine learning algorithms can predict outcomes for a variety of diseases with a high degree of accuracy. The power of machine learning algorithms comes from their capacity to handle vast amounts of data, mix data from many sources, and incorporate fundamental knowledge into their research.

The focus of this study is to develop prediction models using diagnostic and interventional datasets from diabetic patients in Iraq. We employed various machine learning techniques while considering their features and performance, and compared them to obtain the best disease prediction. We explored multiple supervised learning algorithms in the R programming language. Our study employs machine learning classification algorithms to predict the likelihood of diabetes. We evaluated the performance of all algorithms across multiple measures and found that the Support Vector Machine and Random Forest machine learning classification algorithms achieved perfect accuracy.

1. **Materials and methods**
   1. **Diabetes dataset**

The data for this study was initially collected from the laboratories of Medical City Hospital and the Specialized Centre for Endocrinology and Diabetes Al-Kindy Teaching Hospital in Baghdad, Iraq's capital, and contains 1000 records of diabetes patients of all ages. Attributes of the dataset (Gender, AGE, Urea, Cr (Creatinine ratio), HBA1c (Haemoglobin A1c Test), Chol. (Cholesterol), TG (Triglycerides), HDL (High-Density Lipoprotein, or good cholesterol), LDL (Low-Density Lipoprotein, sometimes called bad cholesterol), VLDL (Very Low-Density Lipoproteins), BMI (body mass index), and CLASS (Diabetic=Positive (P), Non-

Diabetic=Negative (N), or Predict-Diabetic=Y). The data set was acquired from a particular location. <https://data.mendeley.com/datasets/wj9rwkp9c2/1>

* 1. **Methods**

Supervised machine learning algorithms give critical and high-accuracy results for prediction. Classification and Regression Trees (CART), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Random Forest (RF) were utilized. Rebala *et al.*, (2019) showed the algorithms were evaluated based on metrics like accuracy and kappa to determine the best model for predicting. Resampling was used to ensure compatibility between the training scheme and models. Rating metrics for each algorithm were stored, and the model was tested using the test dataset. The prediction confusion matrix was created to confirm the results. The error rate for the best model is then determined to optimize the prediction. An extensive collection of data points with responses, also known as a labeled data set, was delivered to the learning algorithm. As a result, the algorithm should be able to predict the result and respond correctly the next time it is presented with a new data point based on essential qualities. Where the model learns about different types of inputs. The models were trained using a labeled dataset in guided learning.

1. **Conclusion**

Based on the results obtained from the previous five algorithms with a comparison function for machine learning algorithms, it shows that the sequence of the accuracy of the algorithms concerning the training data is utterly identical to the results of the overall comparison because the latter mainly depends on the results of the RF, CART, SVM, LDA, and KNN training data. The test data results showed some differences in the accuracy sequence of the algorithms shown, with SVM, RF, CART, LDA, and KNN being the highest, respectively. The training data set refers to the samples used to build the model, while the test or validation data set is used to check performance.

Building a model that understands underlying data patterns is vital to providing long-lasting predictions with little retraining. At their most basic, machine learning models are statistical equations that run at high rates on several data points. As a result, statistical tests on the algorithms are essential for fine-tuning them and verifying whether the model's equation best fits the dataset at hand. We often generate multiple viable models when working on a machine learning project. Each model will have its own set of performance attributes. Using resampling techniques like cross-validation, we can determine how accurate each model is on unseen data. We must be able to use the estimations to choose one or two of the best models from our array of models.

This research will contribute to a scientific addition to past studies in this field of knowledge, which must conducted in various sectors. Focusing on authentic data in all areas, notably health, because it is directly related to human life, which is at the heart of all life on Earth. It is vital to emphasize the importance of data from its sources and urge governments to open data centers, particularly in countries such as Iraq. In this pilot project, we used data from scientific research in Baghdad, Iraq's capital, and five machine learning algorithms. In the future, we hope that data will be available in a variety of disciplines so that we may give service to future generations a better living chance.

After carefully considering the assessment above criteria, we have selected the most optimal machine learning methodology to predict diabetes mellitus in Iraq, with the objective of attaining superior performance. The techniques above are estimated with a supervised dataset for diabetes testing. The optimal choice is the technique that attains the highest level of performance in terms of accuracy and kappa. The findings indicate that the Support Vector Machine (SVM) and Random Forest (RF) algorithms exhibited higher accuracy in predicting diabetes.

The study's findings might help healthcare providers in Iraq avoid diabetes earlier and make better clinical decisions to control it, perhaps saving lives. Our future study will include considering and evaluating new features for further investigation.

**Acknowledgments**

**Reference (Add DOI for all Reference)**

1. International Diabetes Federation. Chapter 3. The global picture. In: *Diabetes* *Atlas*. 8th ed. Brussels, Belgium: International Diabetes Federation; 2017. https://idf.org/e-library/epidemiology-research/diabetes-atlas/134-idf-diabetes-atlas-8th-edition.html. Accessed March (2019).
2. World Health Organization. ; (2018). Diabetes. Geneva, Switzerland: World Health Organization <https://www.who.int/news-room/fact-sheets/detail/diabetes>. Updated October 30, 2018. Accessed March (2019).
3. Mansour AA, Al-Maliky AA, Kasem B, Jabar A, Mosbeh KA. (2014). Prevalence of diagnosed and undiagnosed diabetes mellitus in adults aged 19 years and older in Basrah, Iraq. Diabetes Metab Syndr Obes. ;7:139-144.
4. Khanam, j.j., Simon, Y.F. (2021). Comparison of machine learning algorithms for diabetes prediction, Science Direct, ICT Express: 7, 432–439.
5. Baran, M. (2020). Classification of multi-label data with machine learning methods (Master thesis). Sivas Cumhuriyet University, Sivas.
6. Alan, A. (2020). Evaluation of performance metrics and test techniques on various data sets in machine learning classification methods (Master thesis). Firat University, Fen Bilimleri Enstitusu, Elazig.
7. Parthiban, G., Srivatsa, S.K. (2012). Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. International Journal of Applied Information Systems (IJAIS). Foundation of Computer Science FCS, Volume 3– No.7, 25-30.
8. Keskin, A.K. (2018). Investigation of Machine Learning Classification Algorithms (Master thesis). SINOP UNIVERSITY, Fen Bilimleri institute, Sinop.
9. Nahzat, S., Yaganoglu, M. (2021). Diabetes Prediction Using Machine Learning Classification Algorithms. European Journal of Science and Technology, (24), 53-59.
10. Rebala, G., Ravi, A., Churiwala, S. (2019). *An Introduction to Machine Learning*. Springer, Gewerbestrasse 11, 6330 Cham, Switzerland. 9-11.
11. Ramasubramanian, K., Singh, A. (2019). Machine Learning Using R with Time Series and Industry-Based Use Cases in R. Second Edition. Apress Media, LLC California LLC, USA. 3.

**اسم البحث باللغة العربية**

***الملخص (Italic, 12 PT)***

تكتب الخلاصة باللغة العربية *(Italic, 9 pt)*

الكلمات المفتاحية :***((12 PT***

*الكلمة الاولى، الكلمة الثانية، الكلمة الثالثة، الكلمة الرابعة، الكلمة الخامسة، الكلمة السادسة. ((10 PT*