Predicting Diabetes Using Diabetes Datasets and Machine Learning Algorithms: Comparison and Analysis

Bekim Fetaji[1](✉), Majlinda Fetaji[2], Mirlinda Ebibi[1], and Maaruf Ali[3]

[1] Informatics/Computer Sciences, Mother Teresa University, Skopje, Republic of Macedonia
{bekim.fetaji,mirlinda.ebibi}@unt.edu.mk
[2] Computer Sciences, South East European University, Skopje, Republic of Macedonia
m.fetaji@seeu.edu.mk
[3] Computer Engineering, Epoka University, Vorë, Tiranë, Albania
maaruf@ieee.org

Abstract. The performance of three popular machine learning algorithms to predict diabetes based upon using three diabetes datasets is presented. Two of the datasets are from the public domain and the third is composed from a research study group. The J48, Random Forest and Naïve Bayes machine learning algorithms were evaluated. Machine Learning (ML) is used to both analyse and make predictions from data that is simply too voluminous for humans to process. This is especially true with medical data where the use of machine learning and data analytics is still in its infancy. More specifically this research investigates the application of ML algorithms on the growing data from the healthcare industry on the global diabetes epidemic. The performance of ML algorithms to predict diabetes is lacking. This paper provides an analysis of the challenges of machine learning in this field and covers this gap in the research.

Keywords: Machine learning · Diabetes · Diabetes prediction · J48 · Random forest · Naïve Bayes · ML · Healthcare · Datasets · Data analytics

1 Introduction

Machine Learning (ML) [1] has risen rapidly [2] to become one of the most important branches of Artificial Intelligence (AI) and IT with myriad of applications. ML can be defined generally as algorithms that computers are programmed with so that they can learn from the available inputs or in response to external data. ML is usually used for tasks beyond human capabilities such as: analysing large complex datasets, Big Data or making predictions based on the available data analysed.

ML is used for "automated detection of meaningful patterns in data" [2]. Applications [3] of ML include: image and speech recognition, medical diagnosis, learning associations, predictions, classifications etc.

M. H. Miraz et al. (Eds.): iCETiC 2021, LNICST 395, pp. 185–193, 2021.
https://doi.org/10.1007/978-3-030-90016-8_13

Recently interest has grown of the application of ML for the study of diabetes [4]. This approach deals with the creation of techniques and algorithms that facilitate computers to acquire knowledge and procure intelligence that relies on previous experience. ML represents a member of AI heavily associated with statistics [1]. Here, the system would be capable of recognising and understanding the data related to the input, such that it could make predictions and decisions by depending on that data. One of the most important issues in healthcare recently is dealing with Diabetes. Diabetes represents a well-known metabolic disease that could adversely affect the complete body system. Usually, type two diabetes onset occurs in middle age and rarely in old age. However, incidences of diabetes are also identified in children. Diabetes is driven by multiple etiologic factors such as sedentary lifestyle, food habits, body weight and genetic susceptibility. An undiagnosed diabetes could cause the levels of blood sugar to become very high. This condition is known as hyperglycaemia and could lead to complications such as: cardiac arrest, stroke, diabetic foot ulcer, neuropathy, nephropathy and retinopathy. Hence, diabetes detection at the earliest stage is central to enhance the patient related QOL (quality of life) and extend life expectancy [5].

2 Literature Review

Machine Learning as a concept was pioneered and introduced by the American computer scientist in the field of artificial intelligence and computer gaming, Arthur Samuel, in 1959 [6]. In his paper [6] he actually stated, "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort."

Since 1959 there has been continuous development in the field of ML especially in categorisation and classification of learning and learning algorithms. There are many approaches employed by ML including memorisation, extraction of information and learning by example. It differs from traditional software engineering by instead of providing instructions about the function f (as in traditional software engineering), the computer is provided input x and output y and is expected to determine or predict the function f using what has been provided [i.e., $Y = f(x)$, which can be understood as Output = function(Input)]. Machine learning programs learn through reasoning to solve a problem from examples, rules and information. It can also learn to generalise and help with issues of uncertainty with the use of statistics and probability-driven techniques. Models can also learn from previous computations or experiences to produce subsequent reliable and repeatable decisions and results [5].

Currently selecting effective algorithms for a specific application is a very difficult decision due to the sheer diversity of these algorithms. The decisions of the systems that use ML are made when the machine is able to learn from the data provided as an input and provide prediction as an outcome through pattern matching and data analysis. "A central challenge in building a machine-learning model is assembling a representative, diverse data set" [7]. Choosing the best algorithm depends on many factors such as: performance, speed, accuracy, data used for training, etc. Several databases in the field of biomedical sciences were searched extensively to identify articles that employ ML in healthcare, specifically in diabetes: PubMed, IEEE Xplore, ACM digital library, the DBLP Computer Science Bibliography, etc.

ML is a diverse field. A range of algorithms are given in [8]. However, there are four important types of Machine Learning Algorithms [9]:

1. Supervised Learning
2. Unsupervised Learning
3. Semi-supervised Learning
4. Reinforcement Learning.

Since semi-supervised learning is similar to supervised learning, it will not be covered in this systematic literature review (SLR) - the focus will be on the other three algorithms. This SLR will provide a short description of the three main types of ML algorithms, though the research will be focusing on Supervised Learning and its use cases. Supervised learning is one of the most popular machine learning type and it is widely used on cases with precise demand on mapping between the input and output data.

In the case of supervised learning, the dataset is labelled where the operator providing a dataset knows the correct answer and the algorithm makes predictions based on identified patterns. The categories of algorithms that fall under the supervised learning category include classification, regression and forecasting [10]. According to [5], "The term supervised means that the "machine" (the system) learns with the help of something—typically a labeled (sic) training data". Some of the application types where supervised learning is used are, viz.: fraud detection, email and spam detection, diagnostics, image classification, risk assessment and score prediction.

3 Research Methodology

The research methodology encompassed fundamental research that covered analyses of different ML algorithms for diabetes prediction using the three diabetes datasets. Applied research was then used to test the hypothesis. The main objective of the research study was to investigate the performance of the machine learning algorithms on predicting diabetes and the possibilities of this approach and its applications in improving healthcare.

3.1 Research Methodology and Hypothesis

The research hypothesis was: Using three different machine learning algorithms for three different diabetes datasets will give us more comparable results and increase insight on different machine learning algorithms and their performances.

This research focused on supervised learning algorithms especially on classification algorithms, which were:

- J48
- Naïve Bayes
- Random Forest.

The evaluation of the algorithms proceeded by comparing the results and performance of each algorithm on a specific dataset. All the above algorithms were implemented in WEKA (Waikato Environment for Knowledge Analysis) [4]. WEKA was developed in New Zealand by the University of Waikato. It consists of a collection of algorithms and tools for data mining and machine learning implemented in the JAVA language.

The data collection, processing and analysis of the result were performed using WEKA for this research. The diabetes datasets that have been used for testing were the:

1. UCI Diabetes [11] - this dataset is from AIM'94. It has been used in 29 research projects.
2. Pima Indians Diabetes Dataset [12] - this dataset predicts "the onset of diabetes based on diagnostic measures" [12]. It has been used in 184 projects.
3. Kosovo Dataset [13] - it has been used in a PhD thesis on Data Science and Machine Learning and two further research projects.

4 Results

This section presents the results for all three algorithms using the three diabetes datasets. The first dataset to experiment with was the UCI Diabetes Dataset [11]. After loading the dataset and running the experiment, the results shown below, in Tables 1–3 , were obtained. The speed performance of analysing the dataset is the same for all three algorithms, which is one second. All the tables show the accuracy of the results for each one of the three algorithms performed on the breast-cancer dataset.

For this type of application and data the best performing algorithm is the J48 (Table 1) followed by Naïve Bayes algorithm (Table 2) and the least performing one was the Random Forest algorithm (Table 3).

Table 1. Results for J48 algorithm.

J48 algorithm	Results
Correctly classified instances	216 (75.5%)
Incorrectly classified instances	70 (24.5%)
Kappa statistic	0.2826
Mean Absolute Error (MAE)	0.3676
Root Mean Squared (RMS) error	0.4324
Relative absolute error	87.86%
Root relative squared error	94.61%
Total number of instances	100,000

Table 1 correctly classified 216 instances or 75.5% for the J48 algorithm.

Table 2. Results for Naïve Bayes algorithm.

Naïve Bayes - algorithm	Results
Correctly classified instances	205 (71.7%)
Incorrectly classified instances	81 (28.3%)
Kappa statistic	0.2857
Mean absolute error	0.3272
Root mean squared error	0.4534
Relative absolute error	78.21%
Root relative squared error	99.18%
Total number of instances	100,000

Table 3. Results for random forest-algorithm.

Random forest - algorithm	Results
Correctly classified instances	199 (69.6%)
Incorrectly classified instances	87 (30.4%)
Kappa statistic	0.1736
Mean absolute error	0.3727
Root mean squared error	0.4613
Relative absolute error	89.09%
Root relative squared error	100.9%
Total number of instances	100,000

The Naïve Bayes algorithm performance on correctly classified instances is: 205 or 71.7%.

The Random Forest performance on correctly classified instances is: 199 or 69.6%.

The second dataset to experiment with was the Pima Indians Diabetes Database [12]. After loading the dataset and running the experiment, the results shown in Tables 4–6 were produced. For the Pima Indians dataset [12], the best performing algorithm is the Random Forest algorithm (Table 4) followed by the J48 algorithm (Table 5) and the least performing was the Naïve Bayes algorithm (Table 6).

The correctly classified instances are 19,301 or 96.5% for the Random Forest Algorithm.

The J48 algorithm performance on correctly classified instances is: 17,596 or 88.0%.

The Naïve Bayes performance on correctly classified instances is: 12,823 or 64.2%.

The third dataset to experiment with was the Kosovo Diabetes dataset [13]. After loading the dataset and running the experiment, the results produced are given in Tables 7, 8, 9. For the Kosovo dataset [13], the best performing algorithm is the Naïve Bayes

Table 4. Results for random forest-algorithm.

Random forest-algorithm	Result
Correctly classified instances	19301 (96.5%)
Incorrectly classified instances	699 (3.5%)
Kappa statistic	0.9637
Mean absolute error	0.0131
Root mean squared error	0.0622
Relative absolute error	17.65%
Root relative squared error	32.4%
Total number of instances	768

Table 5. Results for J48-algorithm.

J48 -algorithm	Result
Correctly classified instances	17596 (88.0%)
Incorrectly classified instances	2404 (12.0%)
Kappa statistic	0.875
Mean absolute error	0.0105
Root mean squared error	0.0903
Relative absolute error	14.24%
Root relative squared error	46.94%
Total number of instances	768

Table 6. Results for Naïve Bayes-algorithm

Naïve Bayes-algorithm	Result
Correctly classified instances	12823 (64.1%)
Incorrectly classified instances	7177 (35.9%)
Kappa statistic	0.6268
Mean absolute error	0.0323
Root mean squared error	0.1391
Relative absolute error	43.65%
Root relative squared error	72.33%
Total number of instances	768

algorithm (Table 7) followed by the Random Forest algorithm (Table 8) and the least performing one is the J48 algorithm (Table 9).

Table 7. Results for Naïve Bayes-algorithm.

Naïve Bayes-algorithm	Result
Correctly classified instances	96 (95.0%)
Incorrectly classified instances	5 (5.0%)
Kappa statistic	0.9352
Mean absolute error	0.0153
Root mean squared error	0.098
Relative absolute error	6.98%
Root relative squared error	29.7%
Total number of instances	243

Table 7 shows that the correctly classified instances are 96 or 95.0% for the Naïve Bayes Algorithm.

Table 8. Results for random forest-algorithm

Random forest -algorithm	Result
Correctly classified instances	94 (93.1%)
Incorrectly classified instances	7 (6.9%)
Kappa statistic	0.9074
Mean absolute error	0.1196
Root mean squared error	0.1924
Relative absolute error	54.6%
Root relative squared error	58.3%
Total number of instances	243

The Random Forest algorithm performance on correctly classified instances is 94 or 93.1%.

The J48 algorithm performance on correctly classified instances is 93 or 92.1%.

Table 9. Results for J48-algorithm

J48-algorithm	Result
Correctly classified instances	93 (92.1%)
Incorrectly classified instances	8 (7.9%)
Kappa statistic	0.8955
Mean absolute error	0.0225
Root mean squared error	0.14
Relative absolute error	10.2%
Root relative squared error	42.4%
Total number of instances	243

5 Conclusions

The main purpose of the research study was to investigate the performance of using three different machine learning algorithms on three different diabetes datasets. Taking this approach produced more comparable results and gave a better insight on the performance of the three different machine learning algorithms.

This research study focused on supervised learning algorithms specifically on these three classification algorithms:

- J48
- Naive Bayes
- Random Forest.

The evaluation was conducted by comparing the results and performances of each algorithm on a specific dataset described in the research methodology section. The main objective of this research study was to measure the performance in terms of the speed of execution and accuracy of the three chosen machine learning algorithms listed above on the three different datasets.

The datasets were chosen from two well-known public datasets and one dataset was collected from a research group. This private dataset represented the knowledge workflow that was built using the WEKA tool.

The results were compared for each dataset in terms of speed and accuracy. Test option cross-validation of ten folds was used for all the experiments. This is a technique that runs systematic repeated percentage splits. The dataset is divided into ten pieces where nine are used for testing and one for training. The results for all three chosen algorithms were presented for each dataset.

The differences in accuracy have been affected by the datasets that were selected for this study, thus proving the hypothesis. It can be concluded that the J48 and Random Forest classifiers are slightly slower in performance than the Naïve Bayes classifier for the datasets with large number of instances though more accurate and very powerful at finding very good results.

For future work, the recommendation is to combine the classifiers (algorithms) to compare how two or more classifiers would perform than using a single classifier.

References

1. Bansal, D., Chhikara, R., Khanna, K., Gupta, P.: Comparative analysis of various machine learning algorithms for detecting dementia. Procedia Comput. Sci. **132**, 1497–1502 (2018). https://doi.org/10.1016/j.procs.2018.05.102
2. Aher, S.B., Lobo, L.M.R.J.: Comparative study of classification algorithms. Int. J. Inf. Technol. Knowl. Manage. **5**(2), 239–243 (2012). http://csjournals.com/IJITKM/PDF%205-2/15_Sunita_B_Aher.pdf. Accessed 03 Aug 2021
3. Kevric, J., Jukic, S., Subasi, A.: An effective combining classifier approach using tree algorithms for network intrusion detection. Neural Comput. Appl. **28**, 1051–1058 (2017). https://doi.org/10.1007/s00521-016-2418-1. Accessed 03 Aug 2021
4. Kaur, G., Chhabra, A.: Improved J48 classification algorithm for the prediction of diabetes. Int. J. Comput. Appl. **98**(22), 13–17 (2014)
5. Bonaccorso, G.: Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning. Packt Publishing, 2nd Edition (2018)
6. Samuel, A.L.: Some studies in machine learning using the game of checkers. IBM J. Res. Dev. **3**(3), 210–229 (1959). https://doi.org/10.1147/rd.33.0210
7. Rajkomar, A., Dean, J., Kohane, I.: Machine learning in medicine. New Engl. J. Med. **380**, 1347–1358, (2019). https://www.nejm.org/doi/full/10.1056/NEJMra1814259. Accessed 28 July 2021
8. Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O., Akinjobi, J.: Supervised machine learning algorithms: classification and comparison. Int. J. Comput. Trends Technol. (IJCTT), **48**(3), 128–138 (2017). https://www.ijcttjournal.org/archives/ijctt-v48p126. Accessed 03 Aug 2021
9. Ayyadevara, V.K.: Pro machine learning algorithms. Andhra Pradesh: Apress Media, LLC (2018). https://doi.org/10.1007/978-1-4842-3564-5
10. Mohammed, M., Khan, M.B., Bashier, E.B.M.: Machine Learning Algorithms and Applications. 1st Edition. CRC Press, Boca Raton, 30 June 2020
11. https://data.world/uci/diabetes. Accessed 28 July 2021
12. https://data.world/data-society/pima-indians-diabetes-database. Accessed 28 July 2021
13. Tafa, Z., Pervetica, N., Karahoda, B.: An intelligent system for diabetes prediction. In: 4th Mediterranean Conference on Embedded Computing (MECO), pp. 378–382 (2015). https://doi.org/10.1109/MECO.2015.7181948