Journal of Computer Science

Original Research Paper

Classification Algorithm in Predicting the Diabetes in Early Stages

[1]Subitha Sivakumar, [2]Sivakumar Venkataraman and [2]Asherl Bwatiramba

[1]New Era College, Gaborone, Botswana
[2]Department of Health Information Management, Botho University, Botswana

Article history
Received: 06-08-2020
Revised: 14-10-2020
Accepted: 23-10-2020

Corresponding Author:
Sivakumar Venkataraman
Department of Health Information Management,
Botho University, Botswana
Email: vsivakumarniit@gmail.com

Abstract: Diabetes is a standout amongst the deadliest and Chronical diseases which can increase the blood sugar in the human body. Diabetes gives several complications if it is not diagnosed and treated where it might lead to lifeless. Diabetes could be actively controlled when it is primary predicted. To solve this problem and to predict the diabetes in early stage, the machine learning process is used. In this research work, the classifiers like Naive Bayes, KSTAR, ZeroR, OneR, J48 and Random Forest are implemented to predict the diabetes at primary point. Diabetes dataset is sourced from UCI repository and used for this study. The results are evaluated against the performance, accuracy and time. This research work shows the Naïve Bayes classification algorithms is the best in predicting the diabetes diseases in primary stage where it helps the health professional to start in diagnosing the patient for diabetes and to save the patient life.

Keywords: Diabetes, Classification, Naïve Bayes, KSTAR, Filtered Classifier, OneR, J48 and Random Forest

Introduction

Research in health sector has shown that Type 2 Diabetes Mellitus (DM) has emerged as the new pandemic of the 21st century and it is estimated that 80% of people with diabetes live in low- and middle-income countries. Diabetes Mellitus (DM) is a disease of global public health importance associated with high morbidity and mortality (Wee *et al*., 2005). According to International Diabetes Federation report of 2015, about 415 million people have DM globally with the figure projected to have increased to 642 million by 2040 or maybe even doubled by the year 2040 (Ogurtsova *et al*., 2017).

Technology advancements in different sectors of the economy, have enabled a considerable amount of data to be collected for different purposes. There is opulence of diabetes datasets available within the health systems and online, however there is scarcity of useful analysis tools to find hidden relationship data. It is difficult for health professionals to predict the diabetes as this requires experience and knowledge.

Data mining is a knowledge discovery technique to analyze data and encapsulate it into useful information (Patel and Shukla, 2013). This technique is an inter-disciplinary subfield in Information and Communications Technology (ICT) where used to extract the relevant features for pattern matching from huge dataset. Data mining uses the artificial intelligence and the various machine learning algorithms to acquire the knowledge.

Classification is one of the data analyses techniques used in medical data mining. Classification algorithm used as an supervised learning methods which is used to classify the features to various related groups. With this technique, huge and voluminous medical data sets of patients in hospital or health institution can be used to predict future outcomes. Applications such as Waikato Environment for Knowledge Analysis (WEKA) are data mining software's with algorithms that provide mechanism to explore further medical data sets to reflect hidden patterns, which then be used to predict the patient status. Classification technique provides the insight on patient's details, thereby providing clinical support through analysis.

Medical data mining plays an important role in exploring the hidden patterns which can be used for clinical diagnosis of any disease dataset (Nikhar and Karandikar, 2016). Different classification techniques using WEKA have been applied for different medical datasets, however finding a better classification algorithm is a challenge, as it is difficult to compare different classification in different collection of data.

Many research works are doing their experimentations in improving the performance of predicting the diabetes in primary stage by using the classification algorithms (Kavakiotis *et al*., 2017). Data mining is an intelligent automated system and combined with number of algorithms as a machine learning tool which helps in decision making (Subhadra and Vikas, 2019). By using automated tool, the health professionals are helped in providing a better treatment to the patients in early stage (Vikas *et al*., 2018).

This research is focused on the data mining classification techniques that can predict a certain outcome based on a given input. For that, the six classifiers are used to analyze the diabetes dataset sourced from UCI repository (Dua and Graff, 2019).

Datamining

Datamining is a technique used to obtain the useful features from huge volume of dataset or unstructured data (Saouabi and Ezzati, 2020). In simple, this technique can be used to change the unstructured data to structured data. By using this datamining technique, the patterns can be grouped and the relation between the data can be classified (Neelamegam and Ramaraj, 2013). Based on the patterns the model or framework can be built were it helps for prediction. Researcher are using datamining in most of the fields like education, medicine, business, weather prediction, cyclone and to built the customer relationships (Sujni and Beulah, 2017).

Classification Algorithms

Classification algorithms are the best in Health divisions to predict in diagnosing the patients through ordering the records, where helps to compare the record with the new patient (Isra'a *et al*., 2016). Classification is determined as a supervised learning tool which helps to classify the features to various groups. According to (Sujni and Beulah, 2017) with the sample dataset the classification algorithms can be trained to build a classification model. The six classification algorithms namely Naive Bayes, KSTAR, ZeroR, OneR, J48 and Random Forest are used in this experimentation.

Naïve Bayes Algorithm

Naïve Bayes Algorithm is derived from Bayës theory and works based on the conditional probability. Naïve Bayes algorithm brings the high classification accuracy when the size of the dataset is vast (Subhadra and Vikas, 2019) and take on the features as independent (Benjamin and Antony, 2018). Naïve Bayes is modest and the performance is good in classifiers because getting better accuracy for medical datasets (Subhadra and Sumithra, 2016).

KStar Algorithm

Kstar Algorithm which can be referred as *K** is one of the classifications algorithms which is made by summing the probabilities from the new instance to all the members of a category (Tejera Hernández, 2015). This must be done with the rest of the categories, to finally select that with the highest probability. According (Cleary and Trigg, 1995) to treat the missing values in datasets assumed that the probability of transforming to that kind of values is the mean of the probability of transforming to each of the specified attributes in the dataset.

ZeroR Algorithm

ZeroR is one of the classification algorithms which focuses the goal where it discards the predictors. By using the ZeroR algorithm the main classes can be predicted easily (Arka *et al*., 2018). ZeroR follows the rule based in classification and helps as the baseline performance for other classification algorithms (Nookala *et al*., 2013).

OneR

OneR is a short form of "One Rule". According to (Mahajan and Ganpati, 2014) OneR algorithms builds one rule for all individual features and then picks the best rule which is having the less error percentage as its rule. This OneR algorithm ranks the features based on the error percentage. OneR rules are created by determining the highest frequent classes for each feature.

Random Forest

Breiman (2001) built the Random Forest algorithm for both the classification and regression purpose. By using this method various decision tree can be built to find the rate of classification. Random Forest is one of the furthermost common and adaptable algorithms used in classification. Random Forest is best for big dimensional dataset and is easy to use. According to (Li *et al*., 2015) Random Forest algorithm can be used to improve the computational efficiency and helps to improve the accuracy by not increasing the implementation cost.

Literature Review

Diabetes as one of the deadliest diseases which upsurges the blood sugar in the body (Sisodia and Sisodia, 2018). Diabetes affects the body when the insulin creation is insufficient or when the body is not in a state to use the insulin properly. According to (Nai-Arun and Moungmai, 2015) state that Diabetes Mellitus affects most of the publics in the world of age above 20. With references to the World Health Organization (WHO) report dated on 2014 says that 8.5% of people early from age 18 are in prevalence of diabetes. This

leads to other diseases as loss of sight, kidney failure, high blood pressure and Heart attacks.

According to (Sneha and Gangil, 2019) state that the datamining is a procedure that helps to break the huge data to useful dataset by finding the comparable relations between the data, finding the co-relationships between the data, removing the noisy data and inappropriate data where is used to solve the problems and helps to create new rules by using this data.

Komi *et al*. (2017) elucidate the various classifiers by using the various attributes in Diabetes Dataset. The authors used only a small dataset to predict the diabetes by using the five algorithms like Gaussian Mixture Model, Artificial Neural Network, Support Vector Machine, EM and Logistic regression. From the obtain result, it has been proved that Artificial Neural Network algorithm has the high accuracy in predicting the Diabetes diseases.

Iyer *et al*. (2015) detailed that the classification algorithms like the Decision Tree and Naive Bayes algorithms used to diagnose the diseases using the pattern's in the dataset. According to (Nai-Arun and Moungmai, 2015) state that the Random Forest Classification algorithm performance well in identifying the disease risk and implemented though the Web application to predict a group of diabetes in risk. Lai *et al*. (2019) suggested two machine learning predicting systems like Gradient Boosting Machine and Logistic Regression in identifying the highly risky patients of Diabetes Mellitus.

According to (Meng *et al*., 2013) in the study used various datamining methods to predict the diabetes diseases by the dataset collected using the questionnaire and used SPSS, WEKA for analyzing the dataset. For this research work, the researchers used Artificial Neural Network, J48, Logistic regression algorithms and concluded that J48 provide the high accuracy.

Symptoms and Treatment

Diabetes can be predictable by the following symptoms:

- Polyuria
- Polyuria is a state when the urinary from the body is more than the normal or abnormal
- Polyphagia
- Polyphagia is a state when the body is in excessive hunger
- Polydipsia
- Polydipsia is a state when the body is having excessive thirst
- Gain or loss in body weight
- Body wounds not healing fast
- Blur in eye vision
- Itching in body skin

Diabetes can be diagnosis by the following common tests to examination the glucose level in the body. If the glucose level is more the normal count, then the person may have the diabetes:

- Urine test
- Blood test

Further, the diabetes can be diagnosis and confirmed by additional special test or examinations. Diabetes lead to hypertension, stroke, cancer, obesity, blindness, vascular complications, kidney damage and some nerve damage and more.

Diabetes is diagnosed in the early stage, and then the complications can be reduced. This paper identifies the best method to predict the diabetes in the primary stage.

Methodology

Datamining is a procedure to select the best features from the diabetes dataset and helps in predicting the diseases at early stages. By using the WEKA an open source tool, the datamining task is performed with the help of various algorithms. In this research, the WEKA application is used to do the experiment, from there the results are compared to find the best classification algorithm for predicting the diabetes diseases. Figure 1 shows the model for the classification algorithms in predicting the diabetes diseases.

Table 1: Diabetes dataset structure

Name	Description	Measurement in units	Type
Preg	Number of times pregnant	Number	Numeric
Plas	Plasma glucose concentration a 2 h in an oral glucose tolerance test	mmol/L	Numeric
Pres	Diastolic blood pressure	mm Hg	Numeric
Skin	Triceps skin fold thickness	mm	Numeric
Insu	2 H serum insulin	mu U/ml	Numeric
Mass	Body mass index	kg/m^2	Numeric
Pedi	Diabetes pedigree function	pedi	Numeric
Age	Age	years	Numeric
Class	Class variable	0 or 1	Nominal

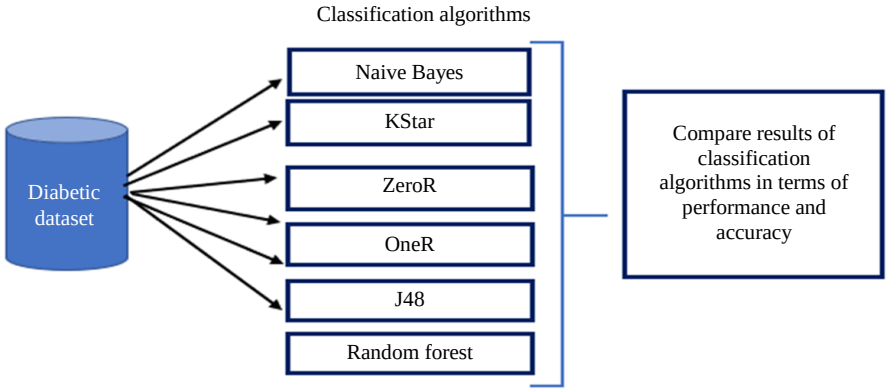

Fig. 1: Classification prediction model for diabetes dataset

For this study, the diabetes dataset is obtained from the UCI repository. The diabetes dataset has 9 features with 768 instances shown in Table 1.

The Diabetes dataset loaded in weka against the different classification algorithm. The result for each algorithm is noted and compared with the other results to find the best classification algorithm in predicting the diabetes diseases in early stages. The best classification algorithm is selected based on the optimal feature selection, performance and accuracy.

Experiment and Results

In this section the diabetes dataset with all features, the experiments and the evaluation schemes were discussed. The classification experiments were implemented by using the Weka Environment tool with a 10-fold cross validation was used. The study evaluated the effectiveness of all the classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy. The results are shown in Table 2.

Firstly, the Diabetes dataset is executed against the Naïve Bayes in WEKA by using the 10 Folds. From the report, the values for the correctly classified instance 76.30 (in percentage), incorrectly classified instance 23.70 (in percentage) and the time taken to build the model 0 (in seconds) are noted. Then, the same dataset is executed against the KStar, ZeroR, OneR, J48, Random Forest the values for the correctly classified instance 76.30 (in percentage), incorrectly classified instance 23.70 (in percentage) and the time taken to build the model 0 (in seconds) are noted as shown in the Table 2.

Discussion on the Results

The experimentation is done and the results are captured as shown in the Table 2. Based on the result values, the graph for Fig. 2 is drawn based on the classification algorithms as in X-axis and values in percentage as in Y-axis. Another graph for Fig. 3 is drawn based on the classification algorithms as in X-axis and the time taken to build the model in seconds as in Y-axis.

From the Table 2, the value for the correctly classified instance for Naive Bayes is 76.30%, Kstar correctly classified instance value is 69.14%, ZeroR correctly classified instance value is 65.10%, OneR correctly classified instance value is 71.48%, J48 correctly classified instance value is 73.83% and Random Forest correctly classified instance value is 75.78%. By comparing the correctly classified instance values for this classification algorithms, the Naive Bayes algorithm illustrates the optimum result. Secondly, with a minimal variance the Random Forest algorithm stands as the next to Naive Bayes algorithm.

From the Table 2, the value for the time taken to build the model for Naive Bayes is 0 sec, Kstar value for the time taken to build the model is 0 sec, ZeroR value for the time taken to build the model is 0 sec, OneR value for the time taken to build the model is 0.01 sec, J48 value for the time taken to build the model is 0.09 sec and Random Forest value for the time taken to build the model is 0.26 sec. By comparing the values for the time taken to build the model for this classification algorithms, the Naive Bayes algorithm illustrates the optimum result when compared to other classification algorithm.

The results from these experiments are shown in Table 2 and Fig. 2. The Naïve Bayes classifier put accuracy at 76.30% and other classifiers provide accuracy of the following; Random Forest 75.78%, J48 73.83%, OneR 71.48%, KStar 69.14% and lastly ZeroR algorithm classifier gives an accuracy of 65.10% as all shown in Fig. 2 and Table 2. After considering these results, it is seen that the maximum accuracy is 76.30% and the minimum accuracy is 65.10%. It can then be concluded that Naïve Bayes classifier is better than the other classifiers considered.

Table 2: Comparison results for diabetes dataset

Classification algorithm	Correctly classified instance in %	incorrectly classified instance in %	Time taken to build the model in seconds
Naive bayes	76.30	23.70	0
KStar	69.14	30.86	0
ZeroR	65.10	34.90	0
OneR	71.48	28.52	0.01
J48	73.83	26.17	0.09
Random forest	75.78	24.22	0.26

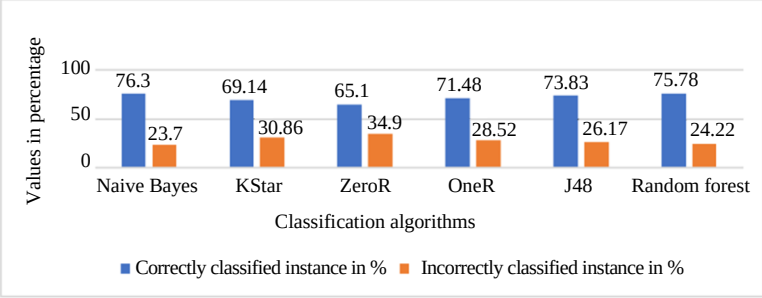

Fig. 2: Classification algorithm result comparison

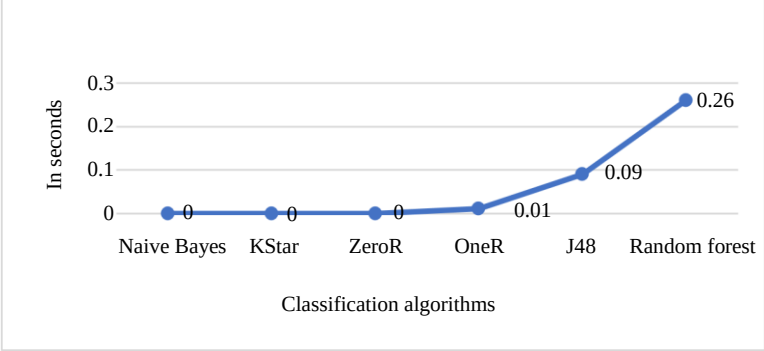

Fig. 3: Time taken to build the model

Conclusion

This study evaluated and investigated six preferred classified algorithms using the Weka tool. The diabetes dataset from UCI repository was used and the most effective algorithm is identified in terms of performance, accuracy and time was the Naïve Bayes classifier with accuracy of 76.30% and total time taken to build mode is at 0 sec. ZeroR algorithm Classifier has the least accuracy of 69.10% with the lowest accuracy in comparison with others. Accuracies are measured by using the values for the correctly classified instances and incorrectly classified instances. Time taken to build the model are measured by using the execution time values.

Acknowledgment

Authors are very thankful to New Era college of Arts, Science and Technology for providing great support and motivation in publishing the research article.

Author's Contributions

All authors equally contributed in this work.

Ethics

The research article is originally written by the authors and the datasets were obtained from the UCI machine learning repository.

References

Arka, H., Prudhvi, R. & Lakshmi, S. (2018). Comparison of different classification techniques using WEKA for diabetic diagnosis. International Journal of Innovative Research in Computer and Communication Engineering. 6(1).

Benjamin, F. D., & Antony, B. (2018). Heart disease prediction using data mining techniques. ICTACT Journal On Soft Computing. 9(1).

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cleary, J. G., & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. In Machine Learning Proceedings 1995 (pp. 108-114). Morgan Kaufmann.

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Isra'a, A. Z., Ahmad, M. A., & Mohammad, A. (2016). A comparative study for predicting heart diseases using data mining classification methods. International Journal of Computer Science and Information Security (IJCSIS), 14(12).

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.

Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 1006-1010). IEEE.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC endocrine disorders, 19(1), 1-9.

Li, L., Wu, Y., & Ye, M. (2015). Experimental comparisons of multi-class classifiers. Informatica, 39(1).

Mahajan, A., & Ganpati, A. (2014). Performance evaluation of rule based classification algorithms. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 3(10), 3546-3550.

Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.

Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132-142.

Neelamegam, S., & Ramaraj, E. (2013). Classification algorithm in data mining: An overview. International Journal of P2P Network Trends and Technology (IJPTT), 4(8), 369-374.

Nikhar, S., & Karandikar, A. M. (2016). Prediction of heart disease using machine learning algorithms. International Journal of Advanced Engineering, Management and Science, 2(6), 239484.

Nookala, G. K. M., Pottumuthu, B. K., Orsu, N., & Mudunuri, S. B. (2013). Performance analysis and evaluation of different data mining algorithms used for cancer classification. International Journal of Advanced Research in Artificial Intelligence (IJARAI), 2(5).

Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., ... & Makaroff, L. E. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes research and clinical practice, 128, 40-50.

Patel, A. K. S. S. B., & Shukla, D. P. (2013). A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level. International Journal of Engineering and Computer Science, 2(09).

Saouabi, M., & Ezzati, A. (2020). Data mining classification algorithms. Computer Science, 15(1), 389-394.

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

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data, 6(1), 13.

Subhadra, C., & Sumithra, P. (2016). A Comparative Study of heart disease prediction using data mining techniques. International Journal of Scientific & Engineering Research. 7(12).

Subhadra, K., & Vikas, B. (2019). Neural network based intelligent system for predicting heart disease. Int. J. Innov. Technol. Exploring Eng.(IJITEE), 8(5), 484-487.

Sujni, P., & Latha, B. C. (2017). Prediction of diabetes using a classification model. Al Dar Research Journal For Sustainability. 2.

Tejera Hernández, D. C. (2015). An Experimental Study of K* Algorithm. International Journal of Information Engineering & Electronic Business, 7(2).

Vikas, B., Anuhya, B. S., Bhargav, K. S., Sarangi, S., & Chilla, M. (2018). Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS). In Information Systems Design and Intelligent Applications (pp. 934-944). Springer, Singapore.

Wee, H. L., Cheung, Y. B., Li, S. C., Fong, K. Y., & Thumboo, J. (2005). The impact of diabetes mellitus and other chronic medical conditions on health-related Quality of Life: Is the whole greater than the sum of its parts?. Health and quality of life outcomes, 3(1), 2.