JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Diabetes Prediction using Data Mining Techniques - A Comparative Study

Prof. Ruchika Patel[a]

[a]*Computer Engineering Department, Gandhinagar Institute of Technology, Gandhinagar University, Moti Bhoyan, Khatraj-Kalol Road, Gandhinagar - 382721, India*

Abstract

Diabetes is one of the fastest increasing chronic diseases, affecting millions of people worldwide. Its diagnosis, prognosis, appropriate treatment, and administration are essential. 382 million people worldwide have diabetes, according to the International Diabetes Federation. This will double to 592 million by 2035. Diabetes is a condition brought on by elevated blood glucose levels. The symptoms of this elevated blood sugar level include frequent urination, increased thirst, and increased hunger. Diabetes is one of the leading causes of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Data mining-based forecasting strategies for diabetes data analysis can aid in the early detection and prediction of the condition. For the diagnosis, prediction, and classification of diabetes, numerous approaches have been developed. The goal of this project is to create a system that, by merging the findings of several data mining approaches, can more accurately perform early diabetes prediction for a patient. A decision tree with the techniques Naive Bayes, K Nearest Neighbour, Logistic Regression, Artificial Neural Network, and Support Vector Machine are employed. Each algorithm's accuracy is calculated along with the model's accuracy. The model for predicting diabetes is then chosen from those with good accuracy. We also emphasise the difficulties and directions for future research.

Keywords: Diabetes, Data mining, Classification, Cluster, Dataset, Diagnosis, Weka, SMO, KNN, ANN, SVM

1. Introduction

Data mining is an essential step of the knowledge discovery process by analysing the massive volumes of data from various perspectives and summarising it into useful information. Data mining is widely used in various application domains such as market analysis, credit assessment, stock market, fraud detection, fault diagnosis in production systems, hazard forecasting, medical discovery, buying trends analysis, knowledge acquisition and science exploration. In general, a data mining system accomplishes one or more of the following data mining tasks. Those are class description, association rule mining, classification, prediction, clustering, time series analysis and outlier analysis. Classification is one of the most necessary and essential tasks in data mining. The use of data mining techniques in the medical field is expanding quickly as a result of their increased accuracy in categorization and prediction. It is crucial for developing strategies to increase patient outcomes, as well as for lowering medical expenses and promoting the early detection of diseases. Using data mining tools like WEKA, the main goal of this study is to evaluate and compare the effectiveness of various categorization techniques.

2. Diabetes : An overview

Diabetes is a long-lasting, non-communicable illness that disrupts the body's normal ability to regulate blood glucose levels. Two hormones—insulin and glucagon—that are released by the pancreas beta (β) and alpha (α) cells, respectively, are typically responsible for controlling blood glucose levels. [4], [5]. The normal secretion of both hormones sustains normal blood glucose concentrations in the body, which are in the range of 70 – 180 mg/dl (4.0 – 7.8mmol/L).Insulin lowers glucose levels, whereas glucagon raises it. However, diabetes results from the improper secretion of these hormones. Type 1 diabetes, type 2 diabetes, and gestational diabetes mellitus (GDM) are the most prevalent kinds of diabetes, while there are many more with varying incidence rates. While GDM appears in women and is diagnosed during pregnancy, type 1 diabetes frequently develops in youngsters, type 2 diabetes is more common in middle-aged and elderly people. In type 1 diabetes, the loss of pancreatic beta cells results in the failure of insulin secretion, but in type 2, both insulin secretion and action fail. GDM is a condition of glucose intolerance of any degree that is first recognized during pregnancy; mainly, it occurs in the second half of pregnancy. It can be mild, but it can also be associated with considerable hyperglycemia and high insulin requirements during pregnancy. All of these contribute to an

imbalanced blood glucose content in the body, which causes serious health issues. Consequently, hyperglycemia is the term used to describe the state when the blood glucose concentration rises and exceeds the usual concentration range. Hypoglycemia, on the other hand, is a condition that occurs when blood sugar levels fall and are below normal. [6]–[8]. Both of these conditions can lead to adverse consequences on an individual's health, for instance, hyperglycemia has long-term complications and can cause nephropathy, retinopathy, cardiovascular and heart diseases, and other tissue injuries, whereas hypoglycemia has short-term effects that may result in life-threatening diabetic coma [4], [6], [7]. Diabetes has become one of the major public health problems in today's world due to its prevalence in children as well as in the adult population. According to [9], [10], Around 415 million adults globally, or 8.8% of the adult population, had diabetes in 2015; by 2040, that number is projected to rise to 642 million. Additionally, during this time the illness has killed about 5 million people and harmed more than 500,000 children. On the other side, it is predicted that the economic cost of diabetes worldwide in 2015 was close to USD 673 billion, and that cost is expected to increase to USD 802 billion by 2040 [10]. Self-monitoring of blood glucose (SMBG) using finger- stick blood samples is a common approach of diabetes therapy that has been introduced three decades ago [11], [12].In this method, diabetics use invasive finger-stick glucose metre to prick the skin of their finger three to four times a day to check their blood glucose levels. The idea is to measure blood glucose concentrations at various intervals and, in response, modify food, activity, and insulin intake to maintain normal glucose levels. However, if the estimation of inulin intake is based on just a few SMBG samples, this method can not only be difficult and uncomfortable but also deceptive. As a result, plasma glycemic levels could rise over the range that is considered normal. Continuous glucose monitoring (CGM), which offers the most information about daily variations in blood glucose concentration and enables diabetes patients to make the best therapeutic decisions, has been developed to address this issue. In this method, tiny wearable gadgets or systems that continually track the blood's glucose concentration levels act as continuous blood glucose monitors. Invasive, minimally invasive, or non-invasive systems may be used. Additionally, there are two categories into which CGM systems can be divided: retrospective systems and real-time systems. [13]. New prospects for diabetic patients to easily manage glucose control have emerged with the introduction and accessibility of a number of cutting-edge CGM devices/systems. Through continuous interstitial fluid (ISF) measurement using a minimally intrusive method, the majority of modern CGM devices typically compute and record the patient's current glucose condition every minute. Since they damage the skin's barrier without damaging any blood vessels, these systems and devices are minimally invasive. Additionally, there are non-invasive techniques, such as illuminating the body's blood capillaries with electromagnetic radiation to measure blood glucose levels. [14]. Additionally, the development of e-Health in the form of telemedicine allows doctors to regularly monitor patients from a distance. Additionally, the CGM data is sent to a remote database in the hospital, where it can be used to forecast critical events in glycemic control like hypo/hyperglycemia. The prevention of hypo/hyperglycemia episodes is one of the difficulties in managing diabetes, but this problem might be solved by properly predicting the blood glucose concentration using the CGM/SMBG and related data (such as exercise, food intake, insulin intake, etc.). Therefore, it is critical to develop tools for the processing and interpretation of data linked to diabetes, CGM/SMBG, and future glucose readings. To this purpose, data mining is crucial in the creation of tools for the detection and forecasting of diabetes. [15], [16]. Data mining is a process of extracting valuable information from a large volume of data in order to discover previously unknown trends, patterns, and relationships that could be used to build models for prediction [17]. Various data mining-based glucose forecasting tools and techniques have been created based on various models in the literature. In order to make clinical judgements, these tools extract, analyse, and interpret the available diabetic data. Figure 1 depicts a general foundation for such strategies. In this article, we give a cutting-edge review on the use of data mining for the diagnosis and prediction of diabetes in the field of glycemic management. Based on the underlying model used, we categorise the frequently employed data mining-based solutions for diabetes diagnosis and prediction. Additionally, we contrast them using important criteria and data. Finally, we highlight the issues that must be resolved as well as potential future research topics.

3. Classification Techniques

One of the critical problems of data mining and machine learning research is the development of accurate and effective classifiers for enormous databases. In the medical field, classification systems are crucial for detecting the disease and its early treatment.The classification process is then split into two stages: the training set, which is used to create the model, and the testing set, which is used to assess the model's accuracy. There are several classification methods available in data mining such as decision tree based algorithms, rule-based algorithms, Naïve Bayesian algorithms, nearest-neighbour algorithms, neural network, rough set, support vector machine, distance based methods, associative classification and genetic algorithms. This study focuses on the following six classification techniques.

1. Decision tree

Decision trees have become one of the most powerful and popular classification approaches in knowledge discovery and data mining, which classify the labeled trained data into a tree or rules. Test data are randomly selected from training data after the learning phase's tree or rules have been developed to evaluate a classifier's accuracy. Unlabeled data is categorised using the tree or rules discovered during the learning phase after accuracy has been confirmed. Theorists and practitioners alike are always looking for ways to improve the process's effectiveness, economy, and accuracy. Applied disciplines including finance, marketing, engineering, and medical regularly use decision trees. Decision tree classifiers are widely used for disease diagnosis, including diabetes prediction, breast cancer, ovarian cancer, heart sound diagnosis, and other conditions. ([19],[20]). There are several algorithms to classify the data using decision trees. The frequently used decision tree algorithms are J48, ID3, C4.5 and CART. In this study J48 algorithm has been used to analyze the performance.

2. Artificial Neural network

An Artificial Neural Network (ANN) is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases, an ANN is a robust system that changes its structure based on external or internal information that flows through the

network during the learning phase [21]. ANN has been confirmed as a powerful method for diabetes prediction [21].One of the key advantages of ANN over traditional approaches is their capacity to capture the complex and nonlinear interplay between prognostic markers and the result to be predicted. Depending on the nature of its input and output data as well as its intended use, an ANN can have a variety of forms. Multilayer perceptron (MLP) has been utilised more frequently than other existing structures for the purpose of predicting diabetes. [21]. Back propagation neural network is also used for detection of diabetes disease and it is considered more suitable to compare with other neural network models [22]. In this study, MLP algorithm has been selected to compare the performance.

3. Logistic Regression

Logistic Regression is one of the most common classification algorithms. LR is considered as the standard statistical approach to modelling binary data [36].It is a better option than linear regression, which assigns a linear model to each class and makes predictions about the future based on the models' consensus. Instead of anticipating the event's point estimate during prediction, it creates a model to forecast the likelihood that it will occur. For instance, in two class problems, the case is placed in the class labelled "1" for "YES" and "0" for "YES" and "NO" when the probabilities are greater than 50%.

4. Support Vector Machines

It is a method of classification that was proposed by Vapnik [24], it is based on the use of a hyperplane separator or a decision. The plane is responsible for defining the limits of decision between a set of data points classified with different labels according to what is mentioned in [25], To find an ideal separation plane or hyperplane that fairly separates the two classes or groups of data points and is equally distant from both of them, they offer the straightforward geometric explanation of this strategy. SVM was first developed to handle problems with linear data distribution, but it may also be applied to non-linear data issues by using the kernel function. Various agents have used SVM to accomplish tasks like recognising the digits in the Vapnik manuscript digits [24], used in object recognition problems [27], text classification [28]. According to [29], the availability of strong tools and algorithms to quickly and effectively find the solution is said to be one of the benefits of SVM. SVM vector support machines have a solid theoretical base and great empirical findings. [26]. In this study, SMO algorithm has been selected to compare the performance.

5. Naive Bayes

It is a supervised classification method developed using the Conditional Probability Theorem, they perform well in different situations, such as text classification and spam detection. Only a small amount of training data is necessary to estimate certain parameters [25]. Bayesian networks are considered an alternative to classic expert systems oriented to decision making and prediction under uncertainty in probabilistic terms [30]. The NB algorithm is a probabilistic classifier that calculates a set of probabilities based on the frequency and combination of the values given on the dataset [31], the algorithm uses the Bayes theorem and assumes that all data are independent of the values of class variable [32], rarely the assumption of conditional independence is met in real world applications, this is a naïve characterization, but the algorithm tends to work properly and learn fast in several supervised classification problems [33]. The most common algorithms that implement this method are: Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN), Bayesian Network (BN) [25].

6. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) classification [33] classifies instances based on their similarity. An object is classified by a majority of its neighbors. K is always a positive integer. The neighbors are selected from a set of objects for which the correct classification is known. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance. The unknown sample is assigned the most common class among its k nearest neighbors.

When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space. In WEKA this classifier is called IBK.

4. Methodology

Figure 1 describes the methodology. The benchmark diabetes dataset obtained from UCI repository site is used to carry out this research.

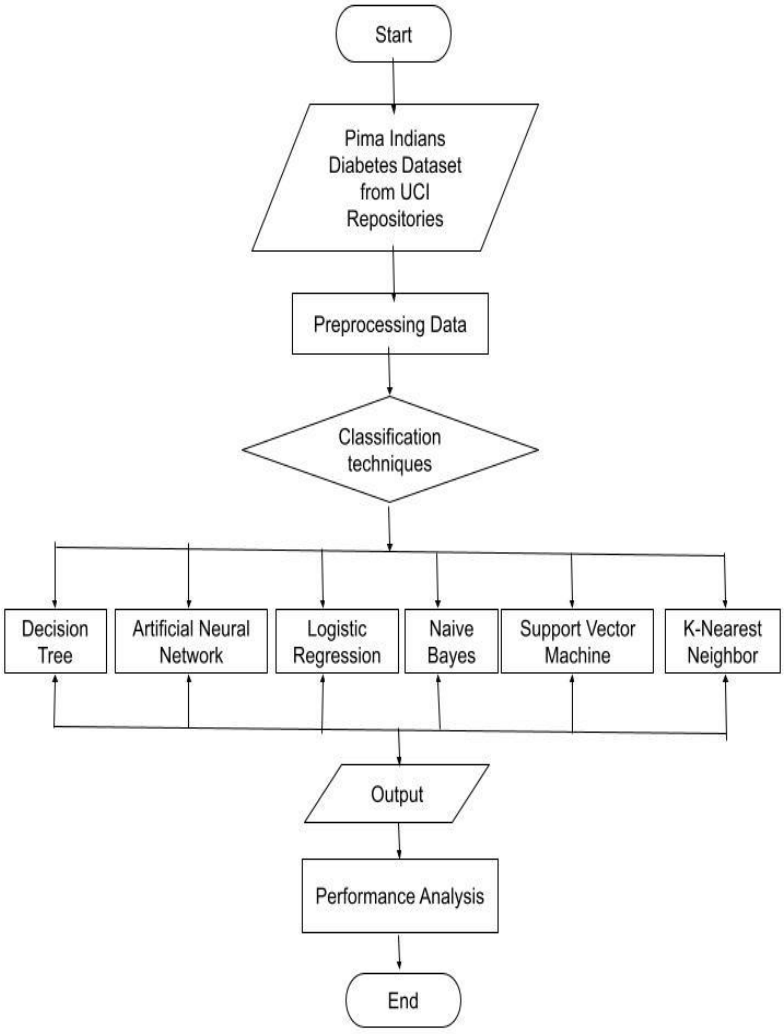

Figure 1. Methodology for diabetes prediction

6.1 Data preprocessing

In this study used a Pima Indians Diabetes dataset with 768 records supplied by the UCI machine learning repository. There are 09 attributes which includes preg, plas, pres, skin, insu, mass, pedi, age and 1 diagnosis result record (tested_negative and tested_positive). The classification experiment is run with various data allocation (training set and testing set) on the same diabetes dataset. There are eight missing values in node-caps and one missing value in skin. Missing values were replaced by using the WEKA pre-processing techniques.

6.2 Developing classification model

WeKA (Waikato Environment for Knowledge Analysis) was utilised to construct the categorization model for the six algorithms. This programme is an extensive collection of Java class libraries that carry out numerous data mining procedures.

5. Experimental Results

Pima Indians Diabetes Data Set was collected from the UCI Machine Learning Repository [28]. This data set contains 768 instances and has 09 patient attributes, which includes preg, plas, pres, skin, insu, mass, pedi, age and 1 diagnosis result record (tested_negative and tested_positive).

This research proposal's primary goal is to evaluate the performance of the diabetic data categorization algorithms in light of the many different input characteristics. Decision tree, ANN, Logistic Regression, Naive Bayes, Support vector machines, and KNN

classification methods are used to analyse them. The performance assessment is conducted using the WEKA programme. Each classifier is used for the 10-fold Cross Validation process. Figure 2 is a screen picture of the WEKA preprocessing stage.

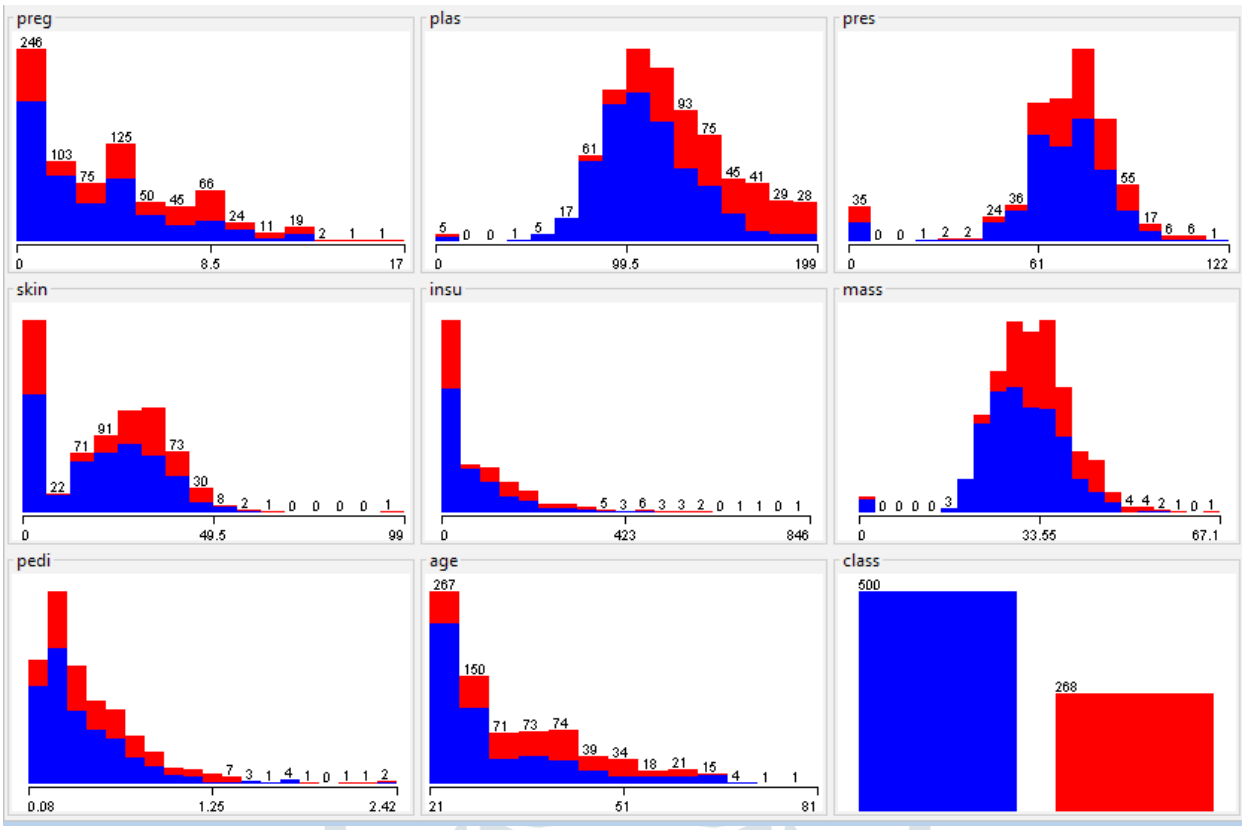

Figure 2: Plot diagram of diabetes attributes data

The following formula is used to calculate the proportion of the predicted positive cases, Precision P using TP = True Positive Rate and FP = False Positive Rate as,

$$\text{Precision P} = \frac{TP}{TP + FP}$$

It has been defined that Recall or Sensitivity or True Positive Rate (TPR) means the proportion of positive cases that were correctly identified. It will be computed as

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where FN =False Negative Rate

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The above formula will calculate the accuracy (the proportion of the total number of predictions that were correct) with TN = True Negative.

The F-Measure can be computed as some average of the information retrieval precision and recall metrics.

$$F = \frac{2 * Recall * precision}{precision + Recall}$$

ROC stands for Receiver Operating Characteristic.

Table 1 shows the weighted average accuracy of the classification algorithm for the diabetes prediction data.

Table 1. Accuracy by weighted average of classification algorithms

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Curve
Decision Tree(J48)	0.738	0.327	0.735	0.738	0.736	0.751

ANN(MLP)	0.754	0.314	0.750	0.754	0.751	0.793
Logistic Regression	0.772	0.321	0.767	0.772	0.765	0.832
Naive Bayes	0.763	0.307	0.759	0.763	0.760	0.819
SVM(SMO)	0.773	0.334	0.769	0.773	0.763	0.720
KNN(IBK)	0.702	0.378	0.696	0.702	0.698	0.650

Table 2. Performance accuracy of algorithms

Decision Tree	ANN	Logistic Regression	Naive Bayes	Support Vector Machine	KNN
73.8281 %	75.3906 %	77.2135 %	76.3021 %	**77.3438 %**	70.1823 %

From the above table we find that the highest accuracy of the Classification model is SVM (77.34%) and low error rate (22.66%) as shown in figure 3.

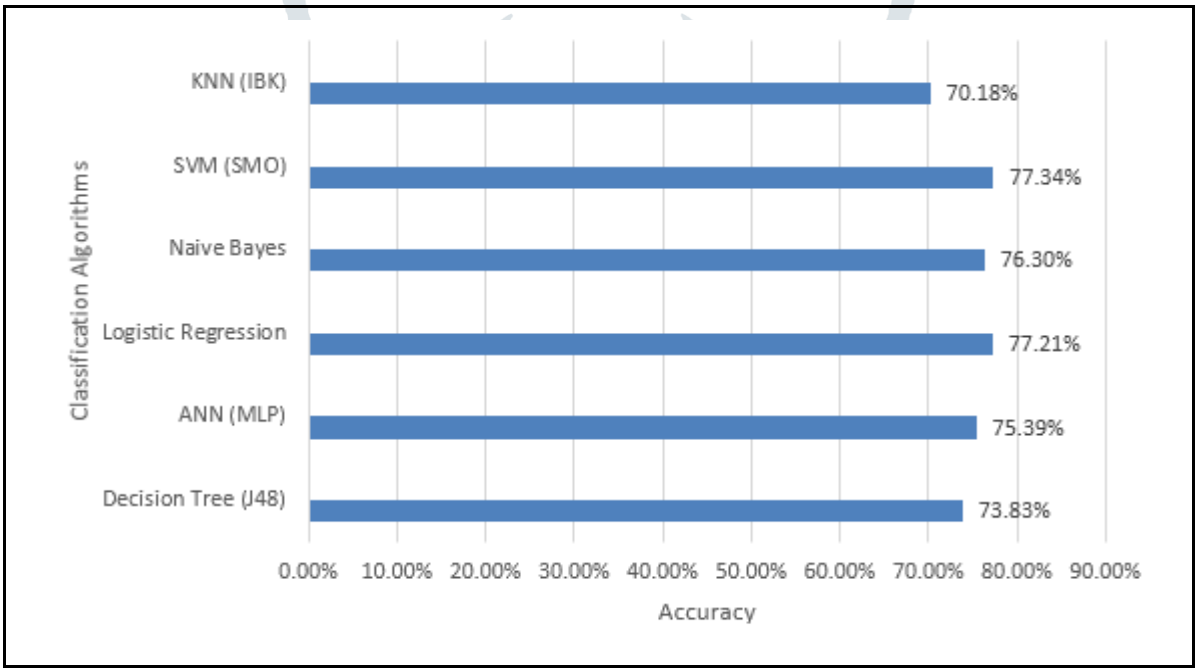

Figure 3: Accuracy of Classification methods

Conclusion

The classification accuracy of Decision Tree, ANN, Logistic Regression, Naive Bayes, SVM, and KNN algorithms is evaluated in this research work utilising a variety of accuracy measures, including FP rate, TP rate, Precision, Recall, F-measure, and ROC Area. According to the experimental findings, the SVM(SMO) classifier has the best accuracy (77.34%), followed by the J48 method (73.83%), the MLP algorithm (75.39%), the Logistic Regression algorithm (77.21%), the Naive Bayes algorithm (76.30%), and the IBK algorithm (70.18%). SVM(SMO) outperforms the other five methods for the selected data set, according to the classification results of the six algorithms. The development of accurate and computationally effective classifiers for medical applications is a significant challenge in the fields of data mining and machine learning. Early diabetes detection is essential for effective treatment. SVM performs at a high level when compared to other classifiers. As a result, SVM displays the numerical results of patient records with diabetic disease. Due to its accuracy, low error rate, and performance, SVM classifier is recommended for diabetes disease-based classification in order to obtain better results.

References

1. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and ol patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

2. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Dataset using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.
3. Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic
4. P. Dua, F. J. Doyle, and E. N. Pistikopoulos, ''Model-based blood glucose control for type 1 diabetes via parametric programming,'' IEEE Trans. Biomed. Eng., vol. 53, no. 8, pp. 1478–1491, Aug. 2006.
5. American Diabetes Association, ''2. Classification and diagnosis of dia-betes: Standards of medical care in diabetes—2020,'' Diabetes Care, vol. 43, no. 1, pp. S14–S31, Jan. 2020.
6. G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, ''Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series,'' IEEE Trans. Biomed. Eng., vol. 54, no. 5, pp. 931–937, May 2007.
7. S. Guerra, A. Facchinetti, G. Sparacino, G. D. Nicolao, and C. Cobelli, ''Enhancing the accuracy of subcutaneous glucose sensors: A real-time deconvolution-based approach,'' IEEE Trans. Biomed. Eng., vol. 59, no. 6, pp. 1658–1669, Jun. 2012.
8. J. M. Norris, R. K. Johnson, and L. C. Stene, ''Type 1 diabetes—Early life origins and changing epidemiology,'' Lancet Diabetes Endocrinol., vol. 8, no. 3, pp. 226–238, Mar. 2020.
9. National Diabetes Statistics Report, 2020. Accessed: Jan. 15, 2021. [Online]. Availabl https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf
10. ID Federation. IDF DIABETES ATLAS 9th Edition 2019. Accessed: Jan. 15, 2021. [Online]. Available: https://diabetesatlas.org/en/
11. L. Olansky and L. Kennedy, ''Finger-stick glucose monitoring: Issues of accuracy and specificity,'' Diabetes Care, vol. 33, no. 4, pp. 948–949, Apr. 2010.
12. J. B. Buse, D. J. Wexler, A. Tsapas, P. Rossing, G. Mingrone, C. Mathieu, D. A. D'Alessio, and M. J. Davies, ''2019 update to: Management of hyperglycaemia in type 2 diabetes, 2018. A consensus report by the American diabetes association (ADA) and the European association for the study of diabetes (EASD),'' Diabetologia, vol. 63, no. 2, pp. 221–228, Feb. 2020.
13. M. Langendam, Y. M. Luijf, L. Hooft, J. H. D. Vries, A. H. Mudde, and R. J. Scholten, ''Continuous glucose monitoring systems for type 1 diabetes mellitus,'' Cochrane Database Syst. Rev., vol. 2012, no. 1, pp. 1–144, 2012, Art. no. CD008101.
14. C. Choleau, J. C. Klein, G. Reach, B. Aussedat, V. Demaria-Pesce, G. S. Wilson, R. Gifford, and W. K. Ward, ''Calibration of a subcutaneous amperometric glucose sensor: Part 1. Effect of measurement uncertainties on the determination of sensor sensitivity and background current,'' Biosensors Bioelectronics, vol. 17, no. 8, pp. 641–646, Aug. 2002.
15. D. Sisodia and D. S. Sisodia, ''Prediction of diabetes using classification algorithms,'' Procedia Comput. Sci., vol. 132, pp. 1578–1585, Jun. 2018.
16. H. Kaur and V. Kumar, ''Predictive modelling and analytics for diabetes using a machine learning approach,'' Appl. Comput. Inform., vol. 16, pp. 1–11, Jul. 2020.
17. K. Kincade, ''Data mining: Digging for healthcare gold,'' Insurance Technol., vol. 23, no. 2, no. 2, pp. 2–7, 1998.
18. Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
19. Aruna, S. P. R. a. L. V. N. (2011). An Empirical Comparison of Supervised learning algorithms in Disease Detection. International Journal of Information Technology Convergence and Services (IJITCS), 1(4), 81-92.
20. Stasis, A. C., Loukis, E., Pavlopoulos, S., & Koutsouris, D. (2003). Using decision tree algorithms as a basis for a heart sound diagnosis decision support system. Paper presented at the Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference on.
21. Burke, H. B., Rosen, D. B., & Goodman, P. H. (1994). Comparing artificial neural networks to other statistical methods for medical outcome prediction. Paper presented at the Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on.
22. S.Neelamegam, D. E. R. (2013). Classification algorithm in Data mining: An Overview. International Journal of P2P Network Trends and Technology (IJPTT), 3(8), 369 - 374.
23. Schwarzer, G., Vach, W., & Schumacher, M. (2000). On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Statistics in medicine, 19(4), 541-561.
24. Vapnik, V.: Statistical Learning Theory. Wiley, Hoboken (1998)
25. Das, K., Behera, R.N.: A survey on machine learning: concept, algorithms and applications. Int. J. Innov. Res. Comput. Commun. Eng. 5(2), 1301–1309 (2017)
26. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2(Nov), 45–66 (2001)
27. Papageorgiou, C., Oren, M., Poggio, T.: A general framework for object detection. In:Proceedings of the International Conference on Computer Vision (1998)
28. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0026683
29. Bekele, E., et al.: Multimodal adaptive social interaction in virtual environment (MASI-VR) for children with Autism spectrum disorders (ASD). In: 2016 IEEE virtual reality (VR), pp 121–130 (2016). https://doi.org/10.1109/vr.2016.7504695
30. Picard, R.W., et al.: Affective learning—a manifesto. BT Technol. J. 22(4), 253–269 (2004). https://doi.org/10.1023/B:BTTJ.0000047603.37042.33
31. Patil, T.R., Sherekar, S.S.: Performance analysis of Naive Bayes and J48 classification algorithm for data classification. Int. J. Comput. Sci. Appl. 6(2), 256–261 (2013)
32. O'Reilly,K.M.A., Mclaughlin,A.M., Beckett,W.S., Sime, P.J.:Asbestos-related lung disease. Am. Fam. Phys. 75(5), 683–688 (2007)
33. Peddabachigari, S., Abraham, A., Grosan, G., Thomas, J.: Modeling intrusion detection system using hybrid intelligent systems. J. Netw. Comput. Appl. 30(1), 114–132 (2007)
34. J. Han and M. Kamber, Data Mining—Concepts and Technique (The Morgan Kaufmann Series in Data Management Systems), 2nd ed. SanMateo, CA: Morgan Kaufmann, 2006.
35. https://archive.ics.uci.edu/ml/datasets/diabetes
36. Witten H.I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann Publishers, 2005.