Comparison Data Mining Techniques To Prediction Diabetes Mellitus

Aswan Supriyadi Sunge
School of Technology Pelita Bangsa, Jawa Barat, Bekasi, Indonesia
Bina Nusantara University, Jakarta, Indonesia
aswan.sunge@pelitabangsa.ac.id
aswan.sunge@binus.ac.id

Abstract. Diabetes is one of the chronic diseases caused by excess sugar in the blood. Various methods of automated algorithms in various to anticipate and diagnose diabetes. One approach to data mining method can help diagnose the patient's disease. In the presence of predictions can save human life and begin prevention before the disease attacks the patient. Choosing a legitimate classification clearly expands the truth and accuracy of the system as levels continue to increase. Most diabetics know little about the risk factors they face before the diagnosis. This method uses developing five predictive models using 9 input variables and one output variable from the dataset information. The purpose of this study was to compare performance analysis of Naive Bayes, Decision Tree, SVM, K-NN and ANN models to predict diabetes millitus

1. Introduction

Diabetes Millitus (DM) is dangerous disease number five in the world, In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2016, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 240 million deaths[1,2]. DM can attack human organs such as the kidneys, eyes, heart, nerves, legs and result in death[3]

1.1 Data Mining

The origin of data mining from the word mining that means mine if developed to dig past data. Collecting data mining is not just data collection but includes the analysis and prediction of the information you want to display and collected data stored in the database is then processed so that it can be used for decision making to view the information to be used[4]. Many data in large quantities in electronic form, and can be used into useful information and knowledge for predictions especially in market analysis, business management, and decision support, by the Stakeholder in future viewing.[5]

1.2 Diabetes Millitus

DM is one of the chronic diseases and causes of death, where people have high blood sugar levels. This affects the body to use energy derived from food. After the body supplies the feeding, then changing the simple sugar (sucrose) usually converts it into glucose and will act as the main source fuel for the body. Glucose is carried in the bloodstream and taken by cells[6]

2. Literature Review

Research in DM has been long researched, mostly in the study using classification data mining methods. Related research related to diabetes prediction mellitus.

- S. Yuvarani and R. Selvarani has used Decision Tree, C4.5 with classification method to predict diabetes millitus[7]

- Alkaragole and Kurnax comparison data mining techiques with Naïve Bayes, SVM, Decision Tree[8]
- Kadhm, Ghindawi and Mhawi used K-Means Clustering and Classification[9]
- Esmaily, Tayefi, etc, comparison with decision tree and random forest[10]
- Benbelkacem and Atmani, with method random forests[11]
- Vijayan and Ravikumar, study data mining algoritms with K-NN, K-Means[12]
- Vaishali, prediction with Decision Tree and SVM[13]
- Sisodia, prediction with SVM, Naïve Bayes [14]
- Steffi and Balasubramanian, used recommended to use ANN and SVM [5]
- Devi and Shyla, diabetes prediction and implemented using classification based data mining algorithm[15]
- Bano and Khan, implemented using K-NN and KNN[16]
- Saravanananathan and Velmurugan, Classification algoritms used C4.5 and K-NN[17]
- Ahmed and Jesmin, comparative analysis with data mining use Classification algoritms[18]
- Iyer, S and Sumbaly, diagnosis diabetes using decision tree and naïve bayes[19]
- Kandhasamy and Balamurali, using classifier models[20]
- Vigneswari, Kumar and Vikash, used machine learning with C4.5 [21]

3. Methodology

3.1 Naïve Bayes

One of the methods in the classification, sometimes referred to as Idiot's Bayes, simple Bayes, Bayes independence due to also a simple method in classification based on the probability theory is Bayesin theorem10. Excess NB is relying on that there are no hidden attributes that could affect in the predictive process[22].

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \tag{1}$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

3.2 Decision Trees

C 4.5 algorithm is an improvement of ID3 using Gain Ratio to update information gain then with formula[23]:

$$GainRatio(S.A) = \frac{Gain\ (S.A)}{SplitInfo(S.A)} \tag{2}$$

- S = Space/Sample Data for data training
- A = Atribute
- Gain(S,A) = Gain Informatioan in attribute A
- SplitInfo(S,A) = Split Information in attribute A

When building a decision tree, there may be noise or blank data in training data. Tree pruning can be done to recognize and remove the branches so that the trees are smaller and easier to understand for better classification.

3.3 Support Vector Machine

Some studies have researched with the SVM method generally capable of delivering performance in terms of classification accuracy of other data algorithms. SVM has been used in a variety of real-world problems such as one of them in the medical world. SVM has proven to be consistently superior with other algorithm methods.

3.4 K-Nearest Neighbors (K-NN)

K-NN is instance based or lazy learning method used for classification in data mining. K-NN is also known as lazy learner because of sample-based learning, when there are training examples present, k-NN learns from example and built a model. K-NN is simple and good classifier.

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

3.5 Artificial Neural Network (ANN)

is a method in neural tissue backpropagation in their approach to predicting diabetes. In the study showed 8-10-1 (1 Hidden layer with 10 nodes) network, which uses the algorithm Levenberg-Marquardt. But it is also through an iterative process, which finds a local minimum of models to best fit the data. Includes applying a backpropagation approach to calculating the gradient cost function in the neural networks of small word comtext to classify diabetes.

4. Performance Evalutioan

4.1 Dataset Description

Data set has applied Pima Indians Diabetes from UCI, there are 768 data set and 9 attributes. The table 1 show descriptions attribute

Attribute	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure(mmHg)
SkinThickness	Triceps skin fold thickness(mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body Mass Index(BMI)
DiabetesPedigreeFunction	Diabetes Pedigree function
Age	Age(in years)
Outcome	Class variable(0 or 1)

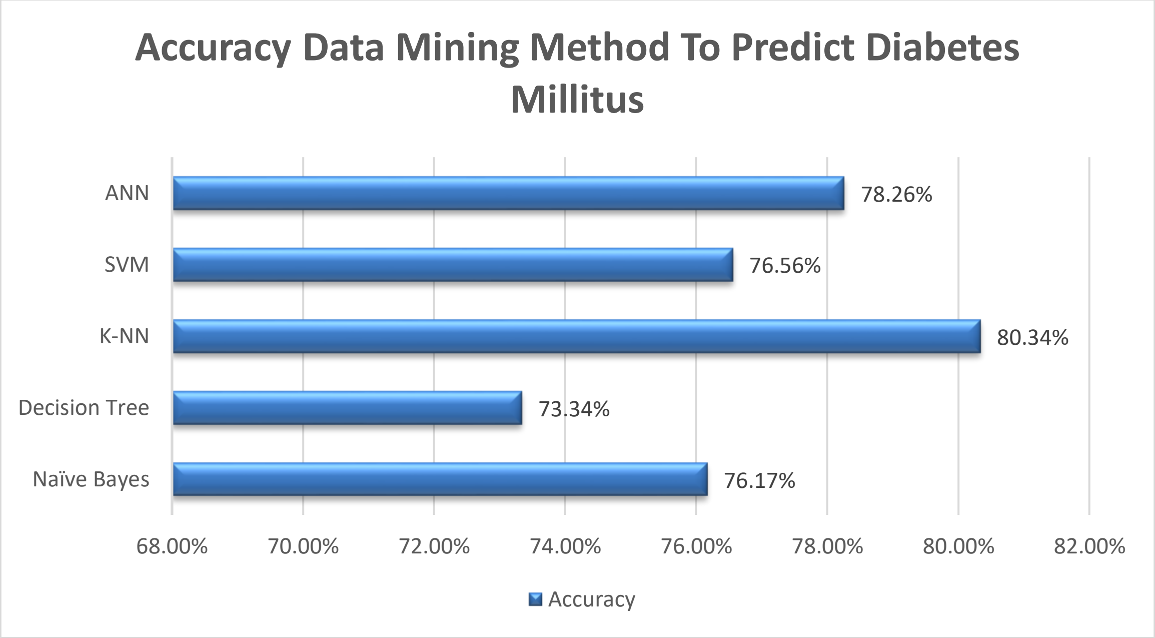
4.2 Evaluation Performance Vector

This paper used confusion matrix to appraise the performance of the five models for incidence of diabetes and five evaluated indices for accuracy, precision, recall and AUC. Where true positives, true negatives, false positives and false negatives. The model with highest the is the best predictive model.

5. Performance Analysis

5.1 Accuracy

The Accuracy measures of Data mining from Naïve Bayes, Decision Tree, SVM, K-NN and ANN and their analysis report is given in the following figure 1

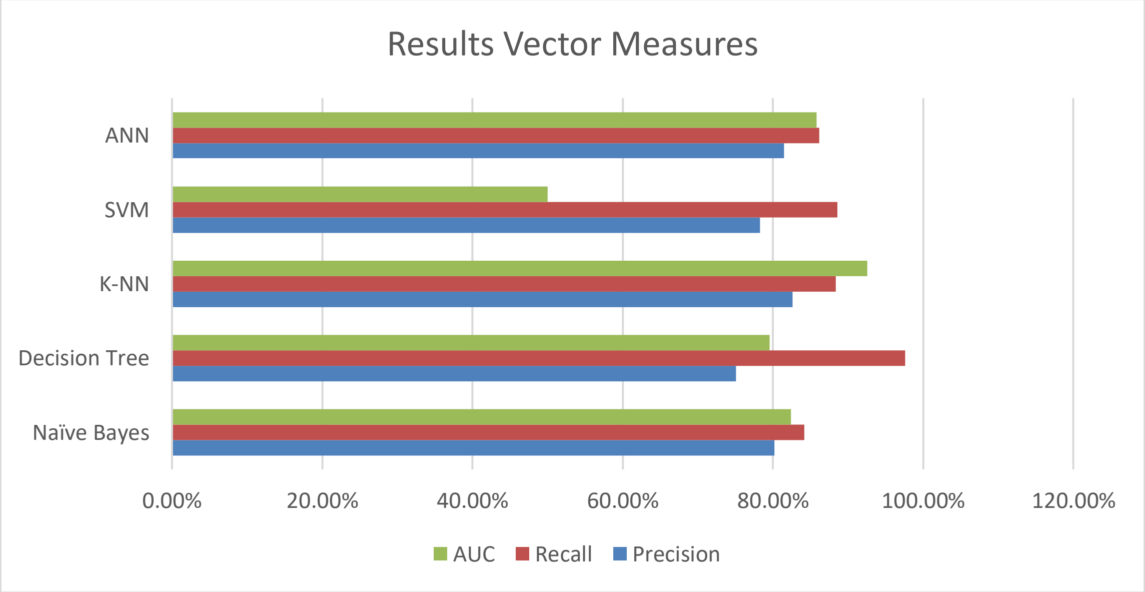


5.2 Other Vector Measures

The other Vector Measures are also calculated and it is shown in the following table 2

ALGORITHMS	PRECISION	RECALL	AUC
Naïve Bayes	80.19%	84.20%	0.824
Decision Tree	75.08%	97.60%	0.796
K-NN	82.62%	88.40%	0.926
SVM	78.27%	88.60%	0.500
ANN	81.47%	86.20%	0.858

Analyzing the Vector Measures figure 2



6. Conclusion

Diagnosis of diabetes is an important real-world problem of medical problems. Detection of diabetes one way out before treatment. This research demonstrates how data mining algorithms are used for the actual model of diabetes mellitus Prediction and comparative analysis is made between them by utilizing vector measurements. As a result of the research detection of diabetes mellitus that K-NN has the highest accuracy and ANN's algorithm is the second.

References

[1] W. H. Organization, "Top 10 causes of death," 2019.

[2] S. Lavery and J. Debattista, "Leveraging pharmacy medical records to predict diabetes using a random forest & artificial neural network," *CEUR Workshop Proc.*, vol. 2259, pp. 279–290, 2018.

[3] K. Suhre *et al.*, "Metabolic footprint of diabetes: A multiplatform metabolomics study in an epidemiological setting," *PLoS One*, vol. 5, no. 11, 2010.

[4] Sunge, Aswan S, "Optimasi Algoritma C4.5 Dalam Prediksi Web Phishing Menggunakan Seleksi Fitur Genetic Algoritma," *Paradigma*, vol. 10, no. 2, pp. 27–32, 2018.

[5] J. Steffi and D. R. B. 1M. Phil Student, "Predicting Diabetes Mellitus using Data Mining Techniques Comparative analysis of Data Mining Classification Algorithms," *Int. J. Eng. Dev. Res.*, vol. 6, no. 2, pp. 460–467, 2018.

[6] S. R. P. Shetty and S. Joshi, "A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique," *Int. J. Inf. Technol. Comput. Sci.*, vol. 8, no. 11, pp. 26–32, 2016.

[7] R. S. S.Yuvarani, "Analysis of Decision Tree Models for Diabetes," *Int. Res. J. Eng. Technol.*, vol. 3, no. 11, pp. 680–684, 2016.

[8] M. Layth, Z. Alkaragole, and A. Sefer Kurnaz, "Comparison of Data Mining Techniques for Predicting Diabetes or Prediabetes by Risk Factors," *Int. J. Comput. Sci. Mob. Comput.*, vol. 8, no. 3, pp. 61–71, 2019.

[9] M. Kadhm, I. Ghindawi, D. M.-I. J. Of, and U. 2018, "An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach," *Int. J. Appl. Eng. Res.*, vol. 13, no. 6, pp. 4038–4041, 2018.

[10] H. Esmaily, M. Tayefi, H. Doosti, M. Ghayour-Mobarhan, H. Nezami, and A. Amirabadizadeh, "A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes.," *J. Res. Health Sci.*, vol. 18, no. 2, p. e00412, 2018.

[11] S. Benbelkacem and B. Atmani, "Random Forests for Diabetes Diagnosis," *2019 Int. Conf. Comput. Inf. Sci.*, pp. 1–4, 2019.

[12] V. VijayanV and A. Ravikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus," *Int. J. Comput. Appl.*, vol. 95, no. 17, pp. 12–16, 2014.

[13] C. Engineering, "Diabetes Prediction using Linear Regression , Decision Tree & Least Square," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 6, no. 4, pp. 3756–3763, 2018.

[14] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018.

[15] M. Renuka Devi and J. Maria Shyla, "Analysis of various data mining techniques to predict diabetes mellitus," *Int. J. Appl. Eng. Res.*, vol. 11, no. 1, pp. 727–730, 2016.

[16] S. Bano, M. Naeem, and A. Khan, "A Framework to Improve Diabetes Prediction using k-NN and SVM," *Int. J. Comput. Sci. Inf. Secur. (IJCSIS)*, vol. 14, no. 11, pp. 450–460, 2016.

[17] K. Saravananathan and T. Velmurugan, "Impact of Classification Algorithms in Diabetes Data : A Survey," *3rd Int. Conf. Small Mediu. Bus. 2016*, pp. 271–275, 2016.

[18] T. Marnoto, "Drying of Rosella (Hibiscus sabdariffa) Flower Petals using Solar Dryer with Double Glass Cover Collector," *Int. J. Sci. Eng.*, vol. 7, no. 2, pp. 155–160, 2014.

[19] A. Iyer, S. Jeyalatha, and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 1, pp. 1–14, 2015.

[20] J. P. Kandhasamy and S. Balamurali, “Performance analysis of classifier models to predict diabetes mellitus,” *Procedia Comput. Sci.*, vol. 47, no. C, pp. 45–51, 2014.

[21] D. Vigneswari, N. K. Kumar, V. G. Raj, A. Gugan, and S. R. Vikash, “Machine Learning Tree Classifiers in Predicting Diabetes Mellitus,” *2019 5th Int. Conf. Adv. Comput. Commun. Syst.*, pp. 84–87, 2019.

[22] A. S. Sunge and W. D. Septiani, “Komparasi Algoritma Data Mining Dalam Prediksi Keamanan Website.”

[23] Sunge, Aswan S, “Prediksi Kompetensi Karyawan Menggunakan Algoritma C4.5 (Studi Kasus : PT Hankook Tire Indonesia)” *Seminar Nasional Teknologi Informasi dan Komunikasi 2018 (SENTIKA 2018)*. ISSN: 2089-9815