Performance evaluation of machine learning classification techniques for Diabetes disease

G Emmanuel, G G Hungilo and A W R Emanuel
Magister Informatika, Universitas Atma Jaya Yogyakarta, Indonesia

*andi.emanuel@uajy.ac.id

Abstract: Diabetes is a noncontagious disease where Diabetes of type two Mellitus is among the top five leading the cause of global death. Not knowing the status of patients leads to complications such as kidney neuropathy and retinopathy, eventually lead to death. Knowing the patient's stand using machine learning techniques can assist in early treatment will be useful in lowering the burdens mentioned above caused by Diabetes. In this work, researchers focused on evaluating the patient's status of Diabetes. In this study, the Cross-Industry Standard Process for Data mining (CRISP-DM) used as a research methodology of research. Where Support Vector Machine, Decision Tree, Naive Bayes used as a classification technique, the study aims to predict the patient status for optimizing the complication caused by Diabetes. The data set used for the model was retrieved from the Pima Indian diabetic database Diabetes Database (PIDD), which is obtained from the UCI machine learning database with 768 records in total. KNN algorithm can be made best with an accuracy of 76% for the condensed dataset with the nine attributes as identified from the comparison of the result of different models.

1. Introduction
Currently, Diabetes is terrifying and risked disease. Now, there is an increasing number of type two Diabetes. In Indonesia, the number of T2DM patients has increased rapidly, both metropolitan and countryside regions, this makes the nation among top countries with a high number of T2DM where 60% affected. The top area with the uppermost frequency of disease in the entire country is Sulawesi, located in the Eastern part of the country [1]. This paper emphasizes the performance evaluation of machine learning classification to know that the patient has Diabetes or doesn't have Diabetes. For Predicting (which performed in prediabetes) and the classification work by grouping the two parties, the patients by comparing different techniques of classification of Diabetes which related to Data Mining [2]. Data Mining helps to analyze the assets of data sets, and it is not having the limit because it works on both continuous and categorical data formats. Machine Learning is a branch of data science that spotlights on structuring, and it can learn from data set and the ability to make predictions on the data [3].

Diabetes is classified into two, which are diabetes type 1 and Diabetes type 2, which is more dangerous [4]. The consequence of Diabetes of type two are many such as making the body weak, and the patient feel pain in internal organs. This type 2 Diabetes has symptoms like feeling thirst, polyuria, eye sickness, and decreasing weight. And if it doesn't treat well, it can cause a loss of life. If Diabetes is found and treated early, it cannot reach the later stage, which results in many complications. To perceive the infection soon, having appropriate information about illness helps to fight against the effect of it.
this study, the authors made a computerized machine innovation to aid in the detection of Diabetes using nine attributes like Body Mass Index, Glucose, blood pressure derived from the diabetes database. Usually, the accuracy results of the various techniques are compared, and the one with high prediction scores often helps to identify diabetic patients.

Diabetes is characterized by the failure of the pancreas in the production of insulin. Failure to treat Diabetes on time complications like heart diseases, poor vision, kidneys, and nerve complications malfunction of blood vessels may come up [5]. The number of diabetics is more than 400 million population [6]. The research shows that by 2035, the number of people with Diabetes will rise from 130 million to 592 million. The majority of people with Diabetes live in poor and middle-income countries; therefore, their way of living is not stable; thus, Diabetes will still increase over the next 22 years [7]. Indonesia has a population of more than 269 million people, and in 2019 it has significantly many numbers of diabetic patients according to IDF, which was 6.3% in 2017 individuals aged (20-79) [8]. Classification is the best choice for interpreting many diabetes data for diagnosis [9]. The statistics show that the diabetic of type 2 will be 439 million in 2030 [10]. Its growing changes are depending on the nature and the lifestyle risk factor of the region [11]. Knowing the existence of diseases in the early stages is most important. Diabetes disease prognosis and interpretation of the diabetes data is a vital type of trouble [9]. In the case of T2DM additionally, the number is expected that eighty-two million people had DM in 2017. The number of type two diabetes patients is expected to be 439 million in 2030. The dominance of T2DM differs from one region to the other [3].

A clinical diagnosis is a class procedure. A physician has to analyze a lot of factors earlier than recognize the presence of Diabetes, which makes the doctor's activity difficult. Currently, devices gaining knowledge and statistics mining strategies were taken into consideration to layout the computerized analysis system for diabetes [12]. Now, there are a lot of methods and classifier used to predict the sicknesses ordinarily commonplace are SVM, Decision Tree, SVM, and Naive Bayes, K – Nearest Neighbor (KNN), Ensemble methods, Random Forest, Genetic Algorithms (GA) [13-15]. All of the above methods are grouped in supervised and Unsupervised machine learning [16].

2. Methodology
In this study, a machine learning classifier is used for classifying Diabetes, while the CRISP-DM used as methodology. The CRISP-DM method offers a designed technique for planning of facts mining projects [16].

2.1. CRISP-DM life cycle process
The CRISP-DM life cycle model consists of six phases which are explained below.

2.1.1. Business understandings. This is the first stage, where the needs are apprehended business objectives evidently to fund out what the commercial enterprises' desires are [16]. Within this research, the undertaking goal is predicting diabetes popularity (Diabetes or not Diabetes).

2.1.2. Data understanding. The statistics knowledge section begins with the first statistical series. The information deployed for making the model was taken from the UCI repository India. It contains information of about 768 instances with nine attributes to categorize the diabetes disease in its early stages, and the information has no missing data.

2.1.3. Data preparation. Information resources have to be well cleaned and arranged in the same format. In this step, take advantage of learning insight into the challenges deeply related to the primary operation of the business style [16]. The final record set is dispatched via some classification techniques to analyze and pick out the import data in the facts set, thus incorporates several missing values.

2.1.4. Modeling. During modeling, special approaches with unique techniques are formed with various changes in parameters. Identified problems are solved by creating models as their answers. The
prediction was conducted and focused on the diabetes conditions (Diabetes or not Diabetes) of the patients. The typical technique of classification algorithms, namely, KNN, Decision Tree, Support Vector Machine, and Naïve Bayesian were used.

2.1.5. Evaluation. Interpreting whether the model is easy to implement and fit with the organization's needs. A thorough evaluation of designing a model to ensure that the goal of business is well served. The main target is to understand whether there are crucial points left for the business to achieve the vision. Finally, this stage gives the main reason if using data mining results can be put into consideration.

2.1.6. Deployment. During this phase, it is important to determine the results that need to be utilized, the user of the results, and how often the results need to be utilized.

2.2. Classification task and algorithms
A classification is just a simple act of identifying to which category new algorithms belong. This is done based on the training sets.

2.2.1. Support Vector Machine (SVM). An SVM is grouped in a supervised learning algorithm. The SVM is best against the overfitting and easily gives the accuracy of sample data [17]. SVM is the best for providing accuracy for a wide range of applications for excessive simplification. Hyperplane ought not to lie in the direction of the information points which belong to the other class [18]. The points that lie near the margin of the classifier are called maintenance vectors [19].

2.2.2. Decision tree classifier. Decision Tree is a supervised learning set of rules used to solve specific problems. Root nodes may have two or other branches, even as the leaf nodes represent the category. In each degree, the choice tree chooses every node utilizing using comparing the fine statistics advantage among all the attributes [20].

2.2.3. Naïve bayes classifier. Naïve Bayes is a classification technique with a notion which assumes that all features are independent and unrelated to each other. It assumes that the status of a specific feature in a class does not affect the status of another feature. This classifier employs the Bayes Theorem for its function. Naïve Bayes introduces significant benefits, since it requires a few parameters for estimation in comparison with its competitors for classification, and does not require structural learning [21].

2.2.4. K-Nearest Neighbour classifier (KNN). K-Nearest Neighbour (KNN) is a classifier used for continuous data. It is also known as a case-based reasoning algorithm used in many applications such as sample popularity and statistical estimation. The KNN algorithm is a non-parametric method that can be used for either classification or regression tasks [22]. This algorithm is preferred over other different types of algorithms because of its excessive convergence speed.

3. Evaluation and results
All the four classifiers have been carried out to the same dataset using Anaconda navigator on editor referred to as Jupiter, and the effects have been acquired and examined in the period of predictive accuracy. And the percentage of n% accuracy in the predictive part indicates the technique is capable of classifying almost n% of instances efficaciously [23].

3.1. Data visualization
Unimodal Data Visualizations used for visualizing data, it is often useful to look at data by constructing histograms of each attribute to get a sense of the data distributions.
3.2. The output of evaluate algorithms

The data set that have been tested using four machine learning algorithms; SVM, Decision Tree, Naïve Bayes, and KNN, the result of each algorithm is shown in the Table2.

Table 2. The table of comparison between four classifications techniques used for experimental.

| Algorithms          | Overall accuracy |
|---------------------|------------------|
| SVM                 | 75%              |
| Decision Tree       | 69%              |
| Naïve Bayes         | 75%              |
| KNN                 | 71%              |

Below is the table which displays the comparison of the performance of each classification techniques.

Table 3. Performing measure.

| MODEL           | SE   | SP   | PRECISION | RECALL | F1-SCORE | CLASS |
|-----------------|------|------|-----------|--------|----------|-------|
| SVM             | 77%  | 62%  | 77%       | 79%    | 78%      | 0     |
|                 |      |      | 63%       | 60%    | 61%      |       |
| Decision tree   | 77%  | 62%  | 77%       | 79%    | 78%      | 0     |
|                 |      |      | 63%       | 60%    | 61%      |       |
| Naïve bayes     | 77%  | 62%  | 77%       | 79%    | 78%      | 0     |
|                 |      |      | 63%       | 60%    | 61%      |       |
| KNN             | 77%  | 62%  | 77%       | 79%    | 78%      | 0     |
|                 |      |      | 63%       | 60%    | 61%      |       |

In table 3, the class is 1 is a diagnosis with Diabetes, while Class 0 is non-diagnosis with Diabetes.
4. Conclusion

The study analyzes and predicts nine attributes related to diabetes patients for different machine learning classification using algorithms, namely; SVM, Decision Tree, Naïve Bayes, and KNN. From the study, the results show that the KNN algorithm outputs the highest accuracy of 75. The primary goal of this employed model is to create a novel patient's health status based on the situation, which will lead to developing a fast and better early diagnosis solution.

Furthermore, the model developed can help in designing an automatic system for diagnosing the patients of Diabetes. Using the model in a large hospital where there are many diabetic patients will assist in knowing the insight and measuring the performance of the accuracy of the current model with other models. This will eventually, without a doubt, help in further assessment of the model's performance.

Reference

[1] K RI 2013 Situasi dan analisis diabetes American Journal of Medical Genetics Part A 161(5) pp 1058–1063
[2] Layth M, Alkaragole Z, and Sefer Kurnaz A 2019 Comparison of Data Mining Techniques for Predicting Diabetes or Prediabetes by Risk Factors Int. J. Comput. Sci. Mob. Comput. 8(3) pp 61–71
[3] Alberti K G M M and Zimmet P Z 1998 Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation Diabet. Med. 15(7) pp 539–553
[4] Tripathi K et al. 2016 Chapter-06 Classification and Diagnosis of Diabetes Fundam. Diabetes 39 no. January pp 61–73
[5] Ramesh D and Katheria Y S 2019 Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach Health Technol. (Bert)
[6] Roglic G 2014 Global report on diabetes World Heal. Organ. 58(12) pp 1–88
[7] Guariguata L, Whiting D R, Hambleton I, Beagley J, Linnenkamp U, and Shaw J E 2014 Global estimates of diabetes prevalence for 2013 and projections for 2035 Diabetes Res. Clin. Pract. 103(2) pp 137–149
[8] Hannibal D, Zolet N E, de Souza J C, Speretta G F F, Leite R D, and Prestes J 2010 Exercício físico e obesidade: O impacto de diferentes modalidades Rev. Bras. Prescrição e Fisiol. do Exerc. v4 pp 219–229
[9] Barakat N, Bradley A P, and Barakat M N H 2010 Intelligible support vector machines for diagnosis of diabetes mellitus IEEE Trans. Inf. Technol. Biomed. 14(4) pp 1114–1120
[10] Chamnan P et al. 2011 Incidence of type 2 diabetes using proposed HbA1c diagnostic criteria in the european prospective investigation of cancer-norfolk cohort: Implications for preventive strategies Diabetes Care 34(4) pp 950–956
[11] Zimmet P, Alberti K G M M, and Shaw J 2001 <Nature 2001 Zimmet.pdf> Nature 414 December pp 782–787
[12] Hastie T, Tibshirani R, and Friedman J 2009 Machine Learning Book 27(2)
[13] Samant R, Campus S, and Rao S 2013 Performance of SVM Classifiers in Predicting Diabetes 2(12) pp 512–517
[14] Wei S, Zhao X, and Miao C 2018 A comprehensive exploration to the machine learning techniques for diabetes identification IEEE World Forum Internet Things, WF-IoT 2018 Proc. 2018-Janua. pp 291–295
[15] Kayvaktios I, Tsavo O, Saligogolou A, Maglaveras N, Vlahavas I, and Chouvarda I 2017 Machine Learning and Data Mining Methods in Diabetes Research Comput. Struct. Biotechnol. J. 15 pp 104–116
[16] Chapman P et al. 2000 Step-by-step data mining guide SPSS inc 78 pp 1–78
[17] Apostolakis J 2010 An introduction to data mining Struct. Bond. 134 pp 1–35
[18] Barbella D, Benzaid S, Christensen J, Jackson B, Qin X V, and Musicant D 2009 Understanding
Support Vector Machine Classifications via a Recommender System-Like Approach Proc. Int. Conf. Data Min. pp 305–11

[19] Sisodia D, Singh L, and Sisodia S 2014 Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28–30 236, no. SocProS 2012 pp 1027–1038

[20] Ramanathan T T and Sharma D 2015 An SVM-Fuzzy Expert System Design For Diabetes Risk Classification 6 no 3 pp 2221–2226

[21] Shiri Harzevili N and Alizadeh S H 2018 Mixture of latent multinomial naive Bayes classifier Appl. Soft Comput. J. 69 pp 516–527

[22] Maillo J, Ramirez S, Triguero I, and Herrera F 2017 kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data Knowledge-Based Syst. 117 pp 3–15

[23] Sharma S, Sharma V, and Sharma A 2016 Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis Int. J. Mod. Comput. Sci. 4(3) pp 11–16