

IMPLEMENTATION OF K-NEAREST NEIGHBOR (KNN) AND SUPORT VECTOR MACHINE (SVM) FOR CLASIFICATION CARDIOVASCULAR DISEASE

Wiji Lestari¹⁾, Sri Sumarlinda²⁾

^{1), 2)}Faculty of Computer Science, University of Duta Bangsa Surakarta
Email : wiji_lestari@udb.ac.id

ABSTRACT

Data mining and machine learning are two tools that play an important role in the study of data analysis and decision systems. Classification is a function of data mining. In the classification function, sorting or mapping occurs based on the proximity or similarity of data attributes to the specified label. K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). The KNN algorithm is a non-parametric method used for classification and regression. Cardiovascular disease prediction models with KNN and SVM algorithms are used to identify and predict cardiovascular disease. The KNN algorithm uses Euclidian distance for the prediction process of training data. The SVM algorithm uses a hyperplane for the data prediction training process. The dataset used is 400 with 7 attributes, namely age, gender, systolic blood pressure, cholesterol, talach, oldpeak and slope. The results of the implementation of the KNN and SVM algorithms produce performance with an accuracy of 75.75% on KNN and 76.00% on SVM. The precision value is 76.78% for KNN and 83.93% for SVM. Meanwhile, the recall resulted in 77.14% for KNN and 67.14% for SVM.

Key words: K-Nearest Neighbor (KNN), Support Vector Machine (SVM), data mining, machine learning, cardiovascular disease

INTRODUCTION

The digital era is marked by the increase in data quickly and abundantly. Such massive data then gave rise to the term big data. The development of big data is indicated by volume, velocity, veracity, variability, variety and so on. The development in quantity has also given rise to new studies such as data science, data engineering, data analysis, and so on. The study of data collection is an interesting study and has been widely carried out. Data mining and machine learning are two tools that play an important role in the study of data analysis and decision systems. The data set after being processed with data mining and machine learning will produce information and then a decision (Hastuti et al., 2018). There are three learning methods for analyzing data with machine learning, namely supervised learning, unsupervised learning and reinforcement learning. The three learning methods are closely related to various functions in data mining. Supervised learning is related to prediction, forecasting and classification. Meanwhile, unsupervised learning is related to the clustering function (Zhang et al., 2013).

In the classification function, sorting or mapping occurs based on the proximity or similarity of data attributes with the specified label. In classification, training is usually required as a performance test of the algorithm before it is implemented with data testing. One algorithm that is quite popular for classification and pattern recognition is K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). The KNN algorithm is a non-parametric method used for classification and regression (Prasad et al., 2019). The principle of KNN is to have a set of sample data as a training set that is connected by a label so that the proximity of the data can be known (Zhang et al., 2013), (Alkhatib, 2013). If new data is given without a label, then the data can be determined which label/class it belongs to. From the similarity, the nearest neighbor, the new data label can be determined (Purwanto and Sahid, 2021) (Novitasari et al., 2019), (Sagar et al., 2013). Support vector machine (SVM) is the best stock classifier. SVM has good decision making for data points

outside the training set (Tripathy, 2019), (Rantini et al., 2019). The characteristics of SVM are low generation error, computationally inexpensive, easy to implement and sensitive to tuning parameters and kernel choices. In SVM the dividing line in the dataset is called the separating hyperplane (Hristova et al., 2021), (Guenther and Schonlau, 2016). The closest points that separate the hyperplane are known as support vectors (Akramar et al., 2018), (Sobhani et al., 2013).

This study compared the implementation of the K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithms for cardiovascular disease susceptibility prediction. Cardiovascular disease is a global problem and a major cause of death and disability. Prediction and classification based on risk factors are important for prevention and early treatment. KNN is an algorithm that is widely used for prediction and classification of cardiovascular disease. Prediction of cardiovascular disease using KNN with parameters compared between 8 and 13 parameters (Enriko et al., 2016). KNN is combined with genetics algorithm for prediction of cardiovascular disease (Jabar et al., 2013). Extended KNN is used to predict heart disease with 11 parameters (Kumar & Fatima, 2020). Heart Disease Prediction System (HDPS) is developed using Logistic Regression, K Nearest Neighbor, Decision Tree, Random Forest Classifier, and Support Vector Machine algorithms to predict the heart disease risk level (Bhunias et al., 2021). The KNN algorithm is compared with random forest, Decision tree and Bayesian algorithms for cardiovascular disease prediction (Hasan, 2021), (Yuvah et al., 2022), (Mohit et al., 2021). Web-based KNN is used to predict heart disease (Kamalapurkar and Gunjal, 2020). KNN is used to identify the main risk factors for cardiovascular disease (Guarneros-Nolasco et al., 2021). There are also quite a number of studies using the SVM algorithm to predict cardiovascular disease. the polynomial-SVM with the use of the grid search optimization method had the best performance for the detection of coronary artery disease (Chen et al., 2020). SVM is used as a cardiovascular disease prediction tool (Gupta et al., 2022), (Kanikar & Shah, 2016), (Sandhya, 2020), (Sivaraman & Khanna, 2021). Intelligent cardiovascular disease prediction empowered with gradient descent optimization using SVM (Nawaz et al., 2021). The rule based model to compare the accuracies of applying rules to the individual results of support vector machines, decision trees, and logistic regression on the Cleveland Heart Disease (Mythili et al., 2013).

METHOD

This research was carried out using a comparative study approach on the implementation of the K-Nearest Neighbor and Support Vector Machine (SVM) algorithms in the prediction and identification of cardiovascular disease. The research dataset as training data comes from the UCI Machine Learning Repository <https://archive.ics.uci.edu/ml/datasets/heart+disease>. The KNN algorithm was used with a value of K = 5 with a measure of Euclidian distance. SVM algorithm used with kernel cache 200 and convergence epsilon 0.001. The testing of machine learning performance used accuracy, precision and recall. The stages of the research are as shown in Figure 1 below:

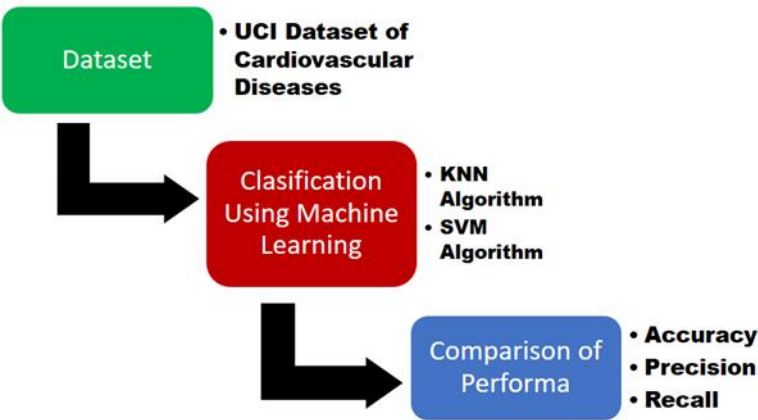


Figure 1. Research Stages

Dataset Cardiovascular Diseas

The heart disease dataset selected for this project came from the UCI Machine Learning Repository <https://archive.ics.uci.edu/ml/datasets/heart+disease>. The dataset consists of 400 individuals’ data, which describe the individual’s health factors and their diagnosis of heart disease. The 7 factors for parameters predictive and cardio for label. The 8 health factors in the dataset used in this project are outlined below.

- 1. **Age** – age of the patient in years
- 2. **Sex** – sex of the patient
 - 0 indicating Female
 - 1 indicating Male
- 3. **Systolic Blood Pressure** – resting blood pressure in mmHg
- 4. **Cholesterol** – serum cholesterol in mg/dl
- 5. **Thalach** – maximum heart rate achieved
- 6. **Oldpeak** – ST depression induced by exercise relative to rest
- 7. **Slope** – the slope of the peak exercise ST segment
 - 1 indicating upsloping
 - 2 indicating flat
 - 3 indicating downsloping
- 8. **Cardio** – diagnosis of heart disease
 - 0 indicating absence
 - 1 indicating presence

K-nearest Neighbor (KNN) algorithm

The K-Nearest Neighbor (KNN) algorithm is an algorithm with supervised learning and is widely used for prediction and classification. The advantages of the KNN algorithm are high accuracy, intensive on outliers and no assumptions about the data. Determining the value of K becomes important. Similarity data with labels used Euclidian distance with the formula:

$$d(x,y) = \sqrt{\sum_i^n (x_i - y_i)^2}$$

Where.
d(x,y) distance of data x and y
x_i is training data ith
y_i is testing data ith

The stages of the KNN algorithm are:

- Determine the value of k
- Calculate distance with Euclidian distance
- Sorting training data
- Define class

Support Vector Machine (SVM) Algorithm

Support vector machine (SVM) is a set of supervised learning methods for classification, regression and outlier detection. The advantage of SVM is that it is effective for high-dimensional spaces and uses a subset of training points on the decision function (support vector). SVM forms a hyperplane in an infinite dimensional space that can be used for regression and classification.

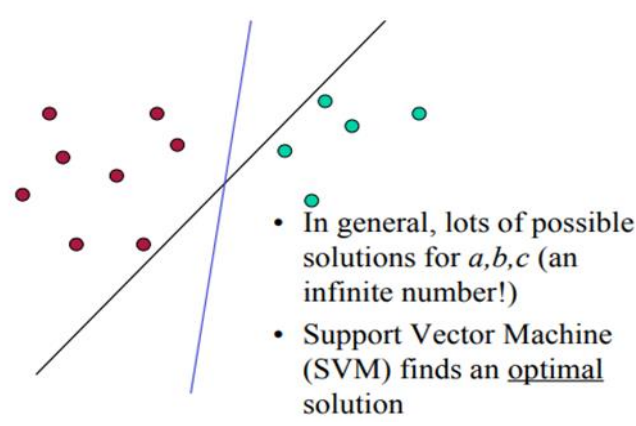


Figure 2. Hyperplane in SVM

DISCUSSION

Development of predictive models with machine learning with KNN and SVM algorithms used Rapidminer software. The dataset used the UCI dataset with 7 parameters used, namely age, sex, systolic BP, cholesterol, talach, oldpeak and slope with a cardio label. The dataset contain 400 patient data consisting of cardiovascular disease and unhealthy/unhealthy patient data. Implementation of the KNN and SVM algorithms for predictive models used Rapidminer as shown in Figure 3 below:

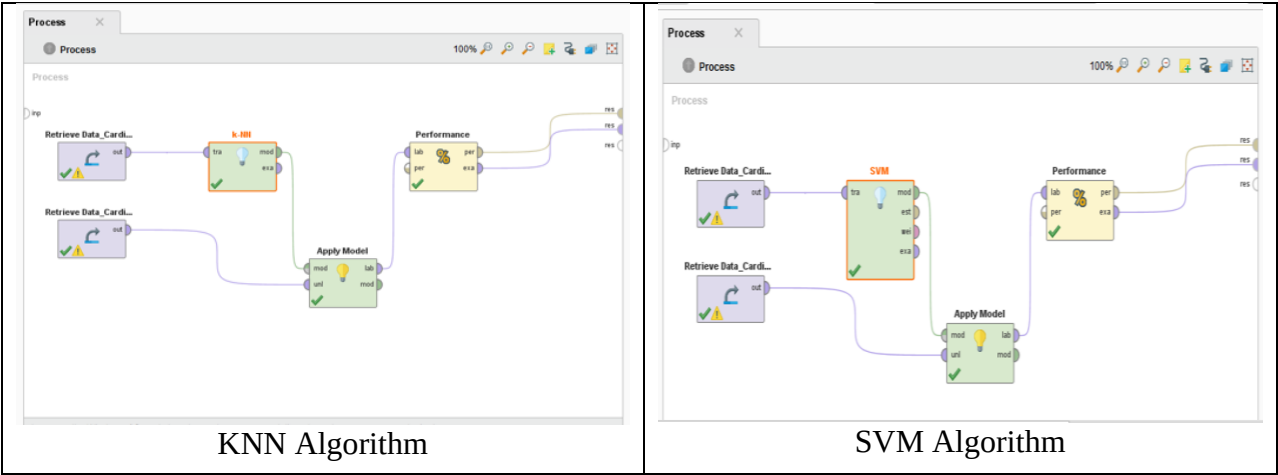


Figure 3. Implementation of KNN and SVM Algorithms for Predictive Model

The implementation of the KNN algorithm used $K = 5$, measure type with MixedMeasure and mixed measure with Mixed Euclidian. Implementation with SVM algorithm produced weight attributes as in table 1, below:

Table 1. Atributes Weight		
No	Atribute	Weight
1	Age	-0.020
2	Sex	1.245
3	Sistolic BP	0.166
4	Cholesterol	0.193
5	Thalach	-0.541
6	Oldpeak	0.613
7	Slope	0.395

From table 1, it can be seen that sex, slope and oldpeak have the greatest weight. Attributes age and talach have negative values.

The performance testing of the KNN and SVM algorithms for predictive models of cardiovascular disease used accuracy, precision and recall. The results of the performance testing of the implementation of the KNN and SVM algorithms are as shown in table 2 below:

Table 2. Performance of KNN and SVM Algorithms		
Performance	KNN	SVM
Accuracy	75.75%	76.00%
Precision	76.78%	83.93%
Recall	77.14%	67.14%

From table 2, it can be seen that the accuracy of the KNN and SVM algorithms is above 75%, so that both algorithms are good enough to be used as predictive models of cardiovascular disease. The SVM algorithm has higher accuracy than KNN. SVM precision is also higher than KNN. As for recall, KNN is superior to SVM.

CONCLUSION

Predictive model of cardiovascular disease with KNN and SVM algorithms used to identify and predict cardiovascular disease. The KNN algorithm uses Euclidian distance for the prediction process of training data. The SVM algorithm uses a hyperplane for the prediction process of training data. The dataset used is 400 with 7 attributes, namely age, sex, systolic blood pressure, cholesterol, talach, oldpeak and slope. The results of the implementation of the KNN and SVM algorithms produce performance with an accuracy of 75.75% on KNN and 76.00% on SVM. The precision value is 76.78% for KNN and 83.93% for SVM. Meanwhile, the recall resulted in 77.14% for KNN and 67.14% for SVM.

REFERENCES

AKramar, V., Alchakov, V.V., Dushko, V.R. and Kramar, T.V., 2018, Application of support vector machine for prediction and classification, *International Conference Information Technologies in Business and Industry 2018 IOP Publishing IOP Conf. Series: Journal of Physics: Conf. Series* 1015 (2018) 032070 doi :10.1088/1742-6596/1015/3/032070

Alkhatib, K., Najadat, H., Hmeidi, I. and Shatnawi, M.K.A., 2013, Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm, *International Journal of Business, Humanities and Technology*, Vol. 3 No. 3.

Badhiye, S.S., Sambhe, N.U. and Chatur, P.N., 2013, KNN Technique for Analysis and Prediction of Temperature and Humidity Data, *International Journal of Computer Applications* (0975 – 8887) Volume 61– No.14, January 2013.

Bhunia, P.K., Debnath, A., Mondal, P., Monalisa, D E, Ganguly, K. and Rakshit, P., 2021, Heart Disease Prediction using Machine Learning, *International Journal of Engineering Research & Technology (IJERT)*, Volume 9, Issue 11.

Chen, X., Fu, Y., Lin, J., Ji, Y., Fang, Y. and Wu, J., 2020, Coronary Artery Disease Detection by Machine Learning with Coronary Bifurcation Features, *Applied Science*, 2020, 10, 7656

Enriko, I.K.A., Suryanegara, M. and Gunawan, D., 2016, Heart Disease Prediction System using k-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters, *Journal of Telecommunication, Electronic and Computer Engineering*, Vol. 8 No. 12.

Guarneros-Nolasco, L.R.; Cruz-Ramos, N.A.; Alor-Hernández, G.; Rodríguez-Mazahua, L.; Sánchez-Cervantes, J.L., 2021, Identifying the Main Risk Factors for Cardiovascular Diseases Prediction Using Machine Learning Algorithms. *Mathematics* 2021, 9, 2537. <https://doi.org/10.3390/math9202537>.

- Guenther, N. and Schonlau, M., 2016, Support vector machines, *The Stata Journal* (2016) 16, Number 4, pp. 917–937.
- Gupta, C., Saha, A., Reddy, N.V.S. and Acharya, U.D., Cardiac Disease Prediction using Supervised Machine Learning Techniques, *AICECS 2021 Journal of Physics: Conference Series* 2161 (2022) 012013 IOP Publishing doi:10.1088/1742-6596/2161/1/012013.
- Hasan, R., 2021, Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction, *ITM Web of Conferences* 40, 03007 (2021).
- Hastuti, I., Purnomo, S. and Lestari, W., 2018, The Guidance of Technopreneurship Using Expert System Computing Approach Based on Entrepreneurial Values and Multiple Intelligences, *International Journal of Economics, Business and Accounting Research (IJEBAR)*, Vol.2, issue 3.
- Hristova, T., Gabrovska-Evstatieva, K. and Evstatiev, B., 2021, Prediction of engineering students' virtual lab understanding and implementation rates using SVM classification, *Journal of E-learning and Knowledge Society*, Vol. 17, No. 1.
- Jabbar, M. A., Deekshatulua, B.L. and Chandra, P., 2013, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, *International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, Procedia Technology* 10 (2013) 85 – 94.
- Kamalapurkar, S. and Gunjal, S.G.H, 2020, Web Support System for Prediction of Heart Disease using k-Nearest Neighbor Algorithm, *International Journal of Computer Applications* (0975 - 8887) Volume 175 - No.14.
- Kanikar, P. and Shah, D.R.K., 2016, Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification, *International Journal of Computer Applications* (0975 – 8887) Volume 156 – No 2.
- Kumar, R.S. and Fatima, S.S., 2020, Heart Disease Prediction Using Extended KNN(E-KNN), *International Journal of Advanced Trends in Computer Science and Engineering*, Volume 9, No.5.
- Mohit, I., Kumar, K.S., Reddy, A.U.K. and Kumar, B.S., 2021, An Approach to detect multiple diseases using machine learning algorithm, *AMSE 2021 Journal of Physics: Conference Series* 2089 (2021) 012009 IOP Publishing doi:10.1088/1742-6596/2089/1/012009.
- Mythili T., Mukherji, D., Padalia, N., and Naidu, A., 2013, A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL), *International Journal of Computer Applications* (0975 – 8887) Volume 68– No.16.
- Nawaz, M.S., Shoaib, B. and Ashraf, M.A., 2021, Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization, *Heliyon* 7 (2021) e06948.
- Novitasari, H.B., Hadiananto, N., Sfenrianto, Rahmawati, A., Prasetyo, R., Miharja, J. and Gata, W., 2019, K-nearest neighbor analysis to predict the accuracy of product delivery using administration of raw material model in the cosmetic industry (PT Cedefindo), *International Conference On Engineering, Technology and Innovative Researches, Journal of Physics: Conference Series* 1367 (2019) 012008 IOP Publishing doi:10.1088/1742-6596/1367/1/012008.
- Prasad, D., Goyal, S.K., Bindal, A. and Kushwah, V.S., 2019, System Model for Prediction Analytics Using K-Nearest Neighbors Algorithm, *Journal of Computational and Theoretical Nanoscience*, Vol. 16, 4425–4430, 2019.
- Purwanto and Sahid, D.S.S., 2021, Using KNN Algorithms for Determining the Recipient of Smart Indonesia Scholarship Program, *Jurnal Komputer Terapan* Vol. 7, No. 2, November 2021, 163 – 173
- Rantini, D., Inas, R. dan Purnami, S.W., 2019, Predicting Popularity of Movie Using Support Vector Machines, *INFERENCE*, Vol. 2(1).
- Sandhya, Y., 2020, Prediction of Heart Diseases using Support Vector Machine, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*
- Sivaraman, K. and Khanna, V., 2021, Machine Learning Models for Prediction of Cardiovascular Diseases, *International Conference on Physics and Energy 2021 (ICPAE 2021) Journal of*

- Physics: Conference Series* 2040 (2021) 012051 IOP Publishing doi:10.1088/1742-6596/2040/1/012051
- Sobhani, J., Khanzadiand, M. and Movahedian, A.H., 2013, Support vector machine for prediction of the compressive strength of no-slump concrete, *Computers and Concrete*, Vol. 11, No. 4 (2013) 337-350.
- Tripathy, N., 2019, Stock Price Prediction Using Support Vector Machine Approach, *International Academic Conference on Management & Economics*.
- Yuvalı, M., Yaman, B. and Tosun, Ö., 2022, Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets. *Mathematics* 2022, 10, 311. <https://doi.org/10.3390/math10030311>
- Zhang, L., Liu, Q., Yang, W., Wei, N. and Dong, D., 2013, An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction, *Procedia - Social and Behavioral Sciences* 96 (2013) 653 – 662.