The Accuracy of Supervised Learning Algorithm on Machine Learning Implementation: a Literature Review

Bagas Tarangga [a,1,*], Evania Trafika [a,2]

[a] Information Systems,Faculty of Computer Science, UPN "Veteran" Jawa Timur
[1] 20082010080@student.upnjatim.ac.id*; [2] 20082010088@student.upnjatim.ac.id
* corresponding author

ARTICLE INFO

Keywords
Machine Leaning
Supervised Learning
SVM Model
LSH
k-NN

ABSTRACT

Machine Learning has become an integral element in technological development, having a significant impact on various sectors of life. This study explores the contribution of Machine Learning in big data processing, automated decision making, and predictive system development. The advantages of Machine Learning, especially in supervised learning, are emphasized by discussing algorithms such as regression, Support Vector Machines (SVM), and Neural Networks. Literature research includes five journals related to supervised learning applications, highlighting findings such as the effectiveness of the Random Forest algorithm in diagnosing pregnancy, the contribution of the SVM model in predicting student study periods, and the level of accuracy with the hybrid LSH and k-NN methods for weather prediction. The practical implementation of fruit detection using cameras shows real application in facilitating price checks and fruit recognition. In conclusion, the literature review confirms the potential and relevance of Machine Learning techniques, especially supervised learning, in providing solutions to various challenges in various sectors. It is recommended that further research explore different industrial sectors or specific case studies to gain a more comprehensive and relevant perspective on current trends in the development of Machine Learning techniques.

1. Introduction

Machine learning (ML) has become an integral part of modern technological development, having a significant impact on various aspects of our lives. This paper explores the importance of machine learning in the context of developments in information technology and artificial intelligence. Through literature analysis methods, we discuss the contribution of ML in big data processing, automated decision making, and predictive system development.

ML's superiority in handling data complexity and its ability to learn from experience make it a vital tool in a variety of industries, including healthcare, finance, and manufacturing. The application of ML not only improves operational efficiency, but also provides valuable insights through in-depth pattern analysis. However, along with positive advances, ML also faces challenges related to ethics, security, and interpretation of results. Therefore, a deep understanding of these advantages and challenges is key to optimizing the use of ML. In ML itself, there are several approaches that can be taken, one of which is supervised learning.

Supervised learningis an important paradigm in machine learning that has made major contributions in complex problem solving and automated decision making. This paper discusses in detail the concepts and applications of supervised learning in various domains. Through a literature review, we investigate the basic principles behind supervised learning and explain how learning models can be guided by labeled training data. We describe several popular algorithms in supervised learning, including regression, Support Vector Machines (SVM), and Neural Networks, and analyze the advantages and disadvantages of each.

The importance of labeled data in supervised learning is highlighted, and we explore strategies used to overcome the challenge of a lack of labeled data. We also highlight the latest developments in supervised learning, including the integration of artificial intelligence technology into everyday life through applications such as facial recognition, language translation, and autonomous cars. Although supervised learning has made significant progress, there are still challenges such as overfitting and model interpretability. Therefore, further research is needed to improve the performance and robustness of the model in the face of real-world complexity.

2. Method

Literature research (literature review) is a form of research that involves investigation and analysis of literature or written works that are relevant to a particular research topic so that the research method applied in this literature review is a systematic approach. The systematic approach itself is an approach that involves a structured literature search using predetermined criteria. The stages in this approach can be seen in Figure 1

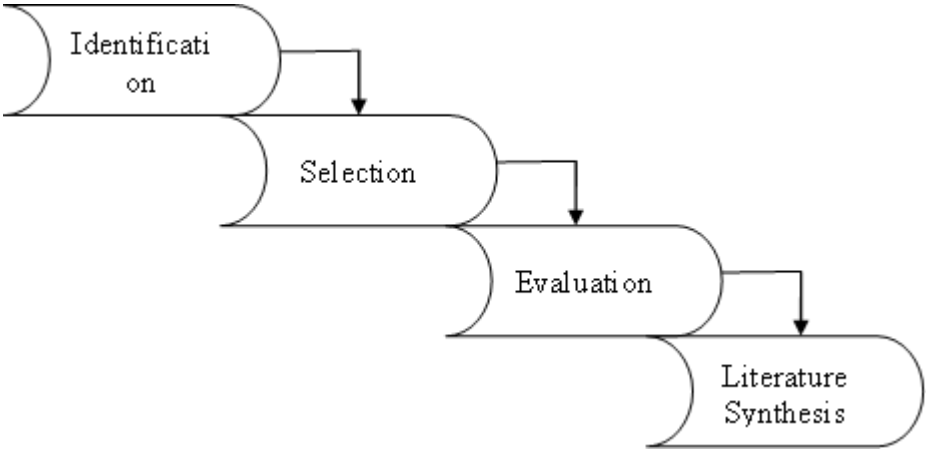

Figure 1. Research Methodology

Identification

The first stage involves identifying literature sources relevant to the research topic. This includes developing a detailed search strategy, including keywords, search phrases, and information sources to be used. This identification can involve academic databases, libraries, journals, conferences, and other literature sources.

Selection

After identification, researchers screen the literature according to predetermined criteria. These criteria may include year of publication, topic relevance, research methods, or specific types of literature. The purpose of selection is to narrow the research focus and ensure that the literature selected is appropriate to the research question or research objectives.

Evaluation

At this stage, the selected literature is critically evaluated. Researchers assess the quality of research methodology, validity of findings, and relevance to the research topic. This evaluation helps ensure that the literature included in the literature study is of sufficient quality to support the argument or research objectives.

Literature Synthesis.

After selection and evaluation, the synthesis stage begins. This involves combining information from various literature sources to form a more comprehensive picture of the research topic. Synthesis can be descriptive, analytical, or theoretical, depending on the purpose of the research and the type of literature at hand.

3. Results and Discussion

After carrying out the identification and selection stages of journals that will be reviewed in the literature, the next stage is evaluating the literature and carrying out a literature synthesis. The literature results can be seen in table 1.

Journal 1	
Title	Accuracy Analysis of Supervised and Unsupervised Learning Modeling Using Data Mining
Writer	Warnia Nengsih
Year	2019
Research purposes	This research aims to model each learning by measuring the accuracy of supervised and unsupervised learning using several testing methods and measuring their accuracy. Accuracy measurement uses the confusion matrix method for supervised learning and lift ratio accuracy testing for unsupervised learning.
Research subject	There are no specific subjects/case studies taken in this research, it only focuses on the accuracy results of supervised modeling and unsupervised learning from several datasets tested.
Research methods	The methods used in supervised learning are Decision trees, Support Vector Machines and Linear Regression. Meanwhile, the methods used for unsupervised learning are k-means, single linkage and a priori
Findings	This research shows that Decision Tree has the highest accuracy (87%) in supervised learning, while K-means has the lowest accuracy (76%) in unsupervised learning.
Implications	Supervised learning methods, such as Decision Trees and Linear Regression, tend to provide higher accuracy compared to unsupervised learning methods such as K-means and hierarchical clustering.
Journal 2	
Title	Supervised Learning Approach for Pregnancy Diagnosis
Writer	Fahira, Zian Asti Dwiyanti, and Roni Habibi
Year	2023

Research purposes	This research aims to find the most effective algorithm between decision trees and random forests in diagnosing pregnancy.
Research subject	Supervised learning model to diagnose pregnancy.
Research methods	This research uses quantitative data analysis methods. Data was collected by making direct observations of the clinic and analyzing the business processes that occurred there. The stages for conducting this research, the author followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) model, consisting of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.
Findings	This research shows that in diagnosing pregnancy, the Random Forest algorithm with the Gini impurity method provides the highest accuracy of 81%, outperforming Decision Tree accuracy which reaches 80%. This indicates that the use of Random Forest with a balanced dataset can increase efficiency in the pregnancy detection process.
Implications	Using the Random Forest algorithm with the Gini impurity method can be a more effective choice in supporting pregnancy diagnosis. This understanding can help develop decision support systems in the health sector to increase accuracy in diagnosing pregnancy conditions.
	Journal 3
Title	Prediction of Study Length and Student Graduation Predicate Using Supervised Learning Algorithm
Writer	Nur Baiti N, Dwi Hartanto, Dicky JS, Lasimin, Dewi M.
Year	2023
Research purposes	This research aims to develop a classification model using a supervised learning algorithm to predict students' study period and GPA when they graduate from college.
Research subject	Supervised learning algorithm to predict student study period and GPA predicate. The data used is Pekalongan University alumni data for 2018, namely 1208 alumni.
Research methods	This research uses a standard methodological method commonly used in data mining research, namely the Cross-Industry Standard Process for Data Mining (CRISP-DM) method, consisting of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.
Findings	This research shows that the Support Vector Machine (SVM) model provides the highest accuracy value of 70% in predicting students' study period, while the K-Nearest Neighbor (KNN) model provides the highest accuracy of 51% in predicting GPA predicate.
Implications	The use of the SVM model can make a significant contribution in

	predicting students' study period, while the KNN model can be used to estimate their GPA predicate. This can help educational institutions identify students at risk and design more effective intervention strategies.
	Journal 4
Title	Supervised Based Weather Prediction in Palembang City Learning Using the K-Nearest Neighbor Algorithm
Writer	Alvi Syahrini Utami, Dian Palupi Rini, Endang Lestari
Year	2021
Research purposes	Predicting the weather in the city of Palembang on a supervised basis learning uses the K-Nearest Neighbor algorithm
Research subject	Data obtained from the Meteorology, Climatology and Geophysics Agency (BMKG) of Palembang city.
Research methods	LSH method to generate hash values for each record in the training data. The hash value obtained will then be used to classify test data which will be calculated using k-NN.
Findings	Weather prediction in the city of Palembang using the hybrid LSH and k-NN methods has been carried out and can work well. The MSE value obtained is 0.301 so the accuracy level of this hybrid method is around 70%.
Implications	The k-NN and LSH methods have a fairly good level of accuracy. In this research, the k-NN and LSH methods are used to predict the weather in the city of Palembang. After the implementation is carried out, the prediction results are analyzed against the actual data which will determine the level of accuracy of this method.
	Journal 5
Title	Fruit Detection Using Supervised Learning and Feature Extraction for Price Checkers
Writer	Kristiawan, Deon Diamanta, Try Atmaja, Andreas Widjaja
Year	2020
Research purposes	Create a system that is able to recognize fruit and help calculate prices easily for consumers.
Research subject	A dataset of 9,695 images as input consisting of: • Apples (3,110 images); • Bananas (2,838 images); • Lemons (3,747 images).
Research methods	In this research, the methods used are OpenCV, HoG, and k-NN. The task is carried out by following the workflow guidelines /

	process for creating a fruit detection learning model.
Findings	The result of this research is a prototype camera system that receives input in the form of photos of fruit. This camera is connected to a computer/laptop, so every time there is input in the form of a photo of fruit, the input will be entered into the machine learning algorithm. Then the process results issued by machine learning will be mapped to the database in the system. Then the price of the fruit will appear.
Implications	The system created is capable of predicting information on fruit names and prices per kilogram based on new data taken from the camera.

Based on a literature review conducted by researchers, in general, five journals that study supervised learning applications highlight the importance of algorithm selection in achieving optimal model accuracy.

In the first journal, researchers utilized data from various case studies which were adapted to the methods applied. The results of accuracy testing, which were evaluated through the confusion matrix and lift ratio, showed that the average accuracy for supervised learning was greater than unsupervised learning. Obtaining this accuracy value is influenced by the number and diversity of data dimensions. Therefore, different cases, quantities and dimensions may result in variations in accuracy values. In the second journal, the test results showed that the Random Forest algorithm using the Gini impurity method had the best accuracy. This indicates that utilizing a balanced dataset can increase efficiency in detecting pregnancy. The Random Forest algorithm is a popular choice in classification thanks to its optimal performance when combined with balanced data and the Gini method. Therefore, to diagnose pregnancy, you can use the Random Forest algorithm with a balanced dataset and use the Gini method to make predictions related to pregnancy. In the third journal, from the modeling results for the two dependent variables, a model was obtained that had the highest accuracy value for predicting future variables. study is the SVM (Support Vector Machine) model, while the model that has the highest accuracy value in predicting the predicate variable is the KNN model. In the fourth journal, the results of weather predictions in the city of Palembang using the hybrid LSH and k-NN methods have been carried out and can run well. The MSE value obtained is 0.301 so the accuracy level of this hybrid method is around 70%. And in the fifth journal, the results of testing with data where shape features and color features have been extracted. The testing results show that the accuracy level for shape features is 99.8% and for color features is 99.7%.

4. Conclusion

Overall, the literature review of five journals that examine supervised and unsupervised learning applications highlights the importance of algorithm selection in achieving optimal model accuracy. Findings, such as the effectiveness of the Random Forest algorithm in diagnosing pregnancy, the contribution of the Support Vector Machine (SVM) model in predicting student study periods, and the level of accuracy that can be achieved with the LSH and k-NN hybrid method for weather prediction, show that the application of this technique has positive implications in various fields, including health, education, and weather monitoring. Apart from that, the practical implementation of fruit detection using a camera with a supervised learning system provides a real example of the use of technology to facilitate price checking and fruit recognition. This conclusion emphasizes the potential and relevance of machine learning techniques, especially those that use supervised learning, in providing solutions to various challenges in various sectors.

For further research, it is recommended to increase the number of journals in the literature review by looking for research involving different industrial sectors or specific case studies. Including the most recent journals is also important to reflect current trends in the development of machine learning techniques. Selecting journals with diverse research scales can provide a more

comprehensive perspective on the application of supervised and unsupervised learning in various contexts.

References

[1] Nengsih, W. (2019) "Analysis of Supervised and Unsupervised Learning Modeling Accuracy Using Data Mining", Sebatik, Vol. 23, no. 2, pp. 285–291.

[2] Fahira, F., Dwiyanti, Z., & Habibi, R. (2023). Supervised Learning Approach for Pregnancy Diagnosis. Incentive Techno Journal, Vol. 17, no. 2, pp. 99-111.

[3] Nasution, NB, Hartanto, D., Silitonga, DJ, Lasimin, & Mardhiyana, D. (2023). Prediction of Study Length and Student Graduation Predicate Using Supervised Learning Algorithm. G-Tech: Journal of Applied Technology, Vol. 7, no. 2, pp. 386–395.

[4] Utami, AS, Rini, DP, & Lestari, E. (2021). Weather Prediction in Palembang City Based on Supervised Learning Using the K-Nearest Neighbor Algorithm. JUPITER: Journal of Computer Science and Technology Research, Vol. 13, no. 1, pp. 09–18.

[5] K. Kristiawan, DD Somali, TA Linggan Jaya, and A. Widjaja, "Fruit Detection Using Supervised Learning and Feature Extraction for Price Checkers", JuTISI, Vol. 6, no. 3.

[6] Chang, Z., Lei, L., Zhou, Z., Mao, S., & Ristaniemi, T. (2018). Learn to Cache : Machine Learning for Network Edge Caching in the Big Data Era. IEEE Wirell. Commun. Vol. 25. pp. 28–35.

[7] Khairudin, I. (2018). Machine Learning Will Become the Most Important Technology After the Internet in 2018. Cellular. ID. https://selular.id/2018/01/machine-learning-akan-jadi- Teknologi-tercepat-cepat-internet-di-2018/. accessed on December 25, 2023.

[8] AW Tawfik, H. Alhoori, CW Keene, C. Bailey and M. Hogan(2018). "Using a Recommendation System to Support Problems," Technology, Knowledge and Learning, vol. 23, no. 1, pp. 177-187.

[9] FH Pratama, A Triayudi, E Mardiani. (2022). Data Mining K-Medoids and K-Means for Grouping Potential Palm Oil Production in Indonesia. JIPI (Scientific Journal of Informatics Research and Learning) vol.7 no. 4, pp. 1294-1310.

[10] Mardiani, Eri, et al. (2023) Comparison of KNN Methods, Naive Bayes. Decision Tree, Ensemble, Linear Regression on High School Student Performance Analysis. INNOVATIVE Journal: Journal Of Social Science Research Vol. 3, no. 2, pp. 13880-13892.