Supervised Machine Learning Techniques: A comparison

NIK SI QIN (161205)
Bachelor of Management, School of Management (SOM)
Universiti Sains Malaysia (USM)
Penang, Malaysia
littlenik1010@student.usm.my

***Abstract*—Machine Learning is a subset of Artificial Intelligence which plays a crucial role in data-driven decision-making worldwide. This research paper focuses on supervised machine learning, comparing the strengths, weaknesses, processes, and tools of four techniques: Linear Regression, K-nearest neighbors, Decision Tree, and Support Vector Machine.**

Keywords—Machine Learning (ML), Supervised Machine learning, Linear Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM).

I. INTRODUCTION

The field of machine learning developed from the more general field of Artificial Intelligence, which seeks to replicate human intelligence in machines. [23]. Machine Learning can be broadly defined as a development of computer systems capable of improving their performance on specific tasks through experience, rather than relying solely on explicit programming instructions [30]. However, the concepts underlying machine learning are not new. Alan Turing, often regarded as the father of machine learning, proposed as early as 1950 that computers could improve their performance on assigned tasks without requiring additional programming [30].

Machine Learning is widely used in many fields, such as bioinformatics [23], medical [30], logistics, oil and gas, and other fields. Machine learning is highly valuable because it can uncover hidden patterns within historical data and generate predictions by applying various ML techniques [24]. The field of machine learning is generally divided into three sub-categories, which are supervised learning (SL), unsupervised learning (UL), and reinforcement learning (RL) [16].

This study will focus on supervised learning techniques, including Linear Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), and Support Vector Machines (SVM). The goal is to compare these techniques based on their strengths, weaknesses, processes, tools, and algorithms.

II. Literature Review

A large number of supervised learning methods have been introduced in the field of machine learning. Supervised learning has emerged as a key research area in machine learning, employing various methods in process and evaluate different types of data [21]. Supervised learning excels in classification and regression analysis to predict outcomes [13]. One of the main characteristics of supervised learning is it can annotate training and testing data [21]. During the training process, the learning algorithm builds the learning model to extract features [26]. In the testing process, the learning model uses the execution engine to predict the test or production data [21]. Finally, the learning model's output, known as tagged data, provides the final prediction or classification results.

Regression analysis is a statistical method proposed in early social research to explore the relationship between variables, such how income levels impact consumer spending behavior. Linear regression is commonly used for numeric prediction and has two main types which are simple linear regression and multiple linear regression [1]. One independent variable is examined by simple linear regression, whereas two or more independent variables are examined by multiple linear regression. When the outcome variable is categorical, logistic regression is better used for classification tasks [9].

K-nearest neighbor method has widely been used in machine learning due to its simple implementation and distinguished performance [34]. KNN is easy to use and effective at solving regression and classification problems. KNN can handle a variety of data kinds, such as text, photos, and videos, and classifies data using similarity metrics [10]. KNN operates by storing all available data and classifying new data points using a distance function to identify the nearest neighbors [4]. Despite its simplicity, KNN remains a powerful tool for classification and regression analysis in machine learning [9].

A decision tree is a visual model that shows every option for a decision depending on a series of conditions, which makes it easy to understand. [9]. It is widely used to uncover patterns in large datasets, supporting algorithms like Classification and Regression Trees (CART), C4.5, ID3, and CHAID to improve predictive accuracy [17]. When the target variable has a finite set of values, the tree is called classification tree [31]; when it has continuous values, it is a regression tree [31]. The DT process starts at the root node and branches out according to conditions until it reaches the leaf node. The leaf node represents the final classification or

prediction [29]. This branching continues until the outcome (leaf) is reached [14].

Support Vector Machine widely used for classification and pattern recognition [8]. It is considered one of the top techniques in data mining due to its robustness and accuracy [32]. SVM works by constructing an N-dimensional hyperplane to separate data into two categories. SVM divides datasets by applying a margin, which stands for the distance between the two classes [32]. SVM can handle both binary and multi-class classification problems, using various kernel functions, including linear, polynomial (degree 2 & 3), and radial, to address different data complexities [8].

III. Findings and Discussion

A. Strengths

The strength of Linear Regression is its mathematical equation is simple, interpretable, and easy to understand. Its mathematical equation is straightforward, making it easy to understand and apply, even for those new to machine learning. Moreover, LR has a lower time complexity compared to more complex algorithms, making it computationally efficient and suitable for large datasets [18]. The simplified form of the linear regression equation is shown below:

$$Y = MX+C \quad (1)$$

Y is dependent variable, X is independent variable, C is the y-intercept, and M is the slope of the line.

The strength of K-Nearest Neighbors is its effectiveness in handling huge and noisy training data [3]. KNN can handle various data types, including numeric, categorical, images, and even text making it more flexible [12]. KNN does not require model training because it utilizes the neighbors' existing classification to classify an unknown sample. [20]. Therefore, KNN is easy to implement compared to other ML algorithms, making it a suitable choice for those starting with machine learning [3].

The strength of Decision Trees is that they can handle both continuous and categorical variables and are resistant to outliers [22]. DT can provide a clear graphical representation of decisions and is easy to understand because it can generate understandable "if-then" rules [22]. In the tree, each node denotes a choice made in response to a feature, and the branches show potential results of that choice. It performs classification without requiring much computation, making it suitable for real-time or resource-limited environments [22].

The strength of Support Vector is in its capacity to manage a variety of classification issues, including those involving high-dimensional and non-linearly separable data [2]. SVM uses kernel functions to map data into higher dimensions in order to solve complicated classification problems. This effectively separates non-linear data and handles high-dimensional datasets with robust performance. Additionally, SVM solely impacts the hyperplane, making outliers less impact [9].

B. Weaknesses

The weakness of Linear Regression is it is sensitive to the effect of outliers which can distort the results and reduce its accuracy and performance [18]. Outliers are extreme values that deviate from the data distribution [18]. Additionally, Linear Regression is prone to overfitting when dealing with many features and assumes a linear relationship between variables, which may not always hold in real-world data [33].

The weakness of K-Nearest Neighbors is space requirement and classification time [3]. KNN requires high memory to store the entire dataset and has slow prediction times due to distance calculations with all data points. The classification rule of KNN can be weak because it is based on one known sample only [18].

The weakness of Decision Trees is its tendency to create overly complex models, especially when trained on noisy or detailed datasets, leading to overfitting [5]. It causes a model to fail to predict accurately on test or real-world data, reducing its usefulness. Additionally, DT takes a lot of time to calculate many numerical variables, sort the variables, and assign them to the subtree to find the optimal solution [14].

The weakness of Support Vector Machine is that it requires a lot of key parameters to be set appropriately to obtain excellent classification results [2]. Incorrect settings can lead to suboptimal performance, making it sensitive and complex to configure. Next, SVM is its high computational cost for large datasets, as the training kernel matrix grows quadratically with the dataset size, making the training process extremely slow [15].

C. Processes

Simple regression model describes the relationship between one independent variable and one or more dependent variable. The mathematical equation of simple and multiple regression model is shown as bellow [6]:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (2)$$

Y_i is dependent variable predicted by independent variable X_i. β0 is the intercept, β1 is the slope indicating the effect of X on Y, and ϵ_i is the random error term in Y_i.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \epsilon_i \quad (3)$$

The population coefficients $\beta_1, \beta_2, \beta_3, \ldots . \beta_k$,represent the unknown parameters that quantify the effect of each independent variable $X_{1i}, X_{2i}, X_{3i}, \ldots X_{ki}$ on the dependent variables (Y_i).

K-Nearest Neighbors focuses on estimating the distance between two data points to identify the nearest neighbors. One of the most commonly used distance metrics in KNN is the Euclidean distance method, particularly for continuous numerical data. Its formula is [32]:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

Where $x_i$and y_i are the i^{th} attributes (components) of x and y respectively, and d is the distance.

In decision trees, Entropy and the Gini index are the most often utilized impurity indices. These metrics help determine the best attribute for splitting the data to reduce impurity and improve classification accuracy. The ID3 and C4.5 algorithms use entropy to calculate information gain, while the CART algorithm uses the Gini index for the same purpose [29]. Another key concept is Information Gain (IG), which evaluates the reduction in uncertainty achieved by a split. These three concepts—Entropy, Gini Index, and IG are explored in detail below.

Entropy in Decision Trees measures the randomness or unpredictability in a dataset. It assesses the homogeneity of data, with values ranging from 0 to 1. Sub-trees are built recursively until the desired outcome is achieved. The entropy formula is:

$$\text{Entropy(T)} = -\sum P(C_i) log_{29} P(C_i)) \quad (5)$$

$p(C_i)$ is the probability of class $C_i$at a node.. An entropy of 0 indicates all samples at a node belong to the same class (poor for training), while maximum entropy reflects a uniform class distribution, which is ideal for training (good for training dataset).

Gini index measures the probability of misclassification in a dataset [S. Dash]. Its value ranges from 0 to 1, where a lower Gini index indicates a lower likelihood of misclassification. The equation of Gini index is shown below:

$$G=1-\sum_{i=1}^{n} P_i^2 \quad (6)$$

G is the Gini index, P_i^2 is the probability of an item belonging to class, n is the total number of classes. The Gini index ranges between 0 (perfect purity) and 1 (maximal impurity), with lower values indicating better splits in decision tree models.

Information gain (IG) refers to how much information a feature can provide when split into two classes [9]. The higher its value, the better the IG knowledge [29]. The equation of IG is shown below:

$$IG\ (T, X) = Entropy(T) - \sum_{v \epsilon Values(X)} \frac{|T_v|}{|T|} Entropy(T_v) \quad (7)$$

T is the dataset, X is the feature, $T_v$is the subset of T for value v of X, $|T|$and $|T_{v|}$are the sizes of T and T_v, and Entropy(T) calculated as -$\sum P(C_i) log_{29} P(C_i))$, where $p(C_i)$ is the probability of each class.

Support Vector Machine (SVM) selecting a hyperplane that maximizes the distance between the hyperplane and the nearest data points from each class while also effectively dividing the data into two classes. SVM process involves the following steps:

1) Hyperplane: The hyperplane is defined as:

$$W^T x + b = 0 \quad (8)$$

where w is the weight vector, and b is the bias. to separate two classes while maximizing the margin.

2) Distance Calculation: The shortest distance to the hyperplane (r) is calculated using below formula [29]:

$$r = \frac{g(x)}{\|W\|} \quad (9)$$

3) Optimization with Lagrange Multipliers: The next step is calculating the Lagrange function. Where a_i are the Lagrange multiplier The Lagrange function is:

$$L(w, b, a) = \frac{1}{2} w^T w - \sum a_i [y_i (w^T x_i + b) - 1] \quad (10)$$

4) The optimal weight vector($w^{*)}$ and bias are computed as:

$$w^* = \sum a_i y_i x_i \quad (11)$$

5) Finally, the optimal bias b∗ can be calculated by using the following equation [29]:

$$b^* = 1 - w^* x_s (\text{for } y_s = +1) \quad (12)$$

For non-linearly separable data, kernel-based SVM such as RBF and polynomial is employed. It translates data into a higher-dimensional feature space using kernel functions, allowing for linear separation in the transformed space. The non-linear SVM score is then a linear combination of additional variables created by the kernel change.

D. Tools

Linear regression is a fundamental supervised learning method implemented in Scikit-learn. It supports model fitting and regularization with Ridge and Lasso to address overfitting. Python libraries like Pandas, NumPy, and Statsmodels, combined with visualization tools like Matplotlib, enable detailed statistical analysis. Software

such as SPSS and SmartPLS further enhance regression analysis capabilities for diverse applications.

K-Nearest Neighbors is a non-parametric technique implemented in Scikit-learn via the KNeighborsClassifier. It allows parameter adjustment for neighbor count and distance metrics. KD-Tree and Ball-Tree are efficient search methods for huge datasets, and R and Python provide implementation and visualization tools. Other platforms, such as RapidMiner, can also run KNN efficiently.

Decision Trees are versatile and interpretable machine learning algorithms supported by various tools. Scikit-learn's DecisionTreeClassifier allows tuning of depth, splitting criteria, and pruning to control overfitting. Visualization tools like Graphviz enhance understanding, while R and RapidMiner provide specialized tree construction and visualization packages.

Support Vector Machines are strong algorithms implemented in Scikit-learn using SVC and Linear SVC. It provides flexibility with kernel types for a variety of datasets. Other tools, such as Weka, support SVM using Sequential Minimal Optimization, while Python and R offer model implementation and evaluation choices.

IV. Conclusion

In conclusion, this study provided a comprehensive comparison of supervised machine learning techniques: Linear Regression, K-nearest neighbors, Decision Trees, and Support Vector Machines. Each method's strengths and weaknesses were analyzed, emphasizing their unique capabilities and limitations in various scenarios. Linear Regression's simplicity contrasts with KNN's adaptability and SVM's robustness for complex data. Decision Trees stand out for interpretability but require careful tuning to prevent overfitting. Overall, understanding these methods aids in selecting the most suitable technique for specific applications, showcasing the critical role of machine learning in decision-making and predictive modeling.

REFERENCES

[1] A. O. Sykes, “An introduction to regression analysis,” The Inaugural Coase Lecture, vol. 20, pp. 1–33, 1993.

[2] A. Soofi and A. Awan, “Classification Techniques in Machine Learning: Applications and Issues,” Journal of Basic & Applied Sciences, vol. 13, pp. 459–465, Jan. 2017, Retrieved from https://doi.org/10.6000/1927-5129.2017.13.76.

[3] Aized Amin Soofi, & Arshad Awan. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, *13*, 459–465. Retrieved from http://setpublisher.com/index.php/jbas/article/view/1715

[4] Analytics Vidhya. (2021). How KNN Uses Distance Measures. Retrieved December 18, 2024, retrieved from https://www.analyticsvidhya.com/blog/2021/08/how-knn-uses-distance-measures/

[5] Asogwa, D. C., Anigbogu, S. O., Onyesolu, M. O., & Chukwuneke, C. I. (2019). Review of some machine learning techniques for big data, their uses, strengths, and weaknesses. Retrieved from http://repository.unizik.edu.ng/handle/123456789/718

[6] Bangdiwala, S. I. (2018). Regression: simple linear. *International Journal of Injury Control and Safety Promotion*, *25*(1), 113–115. Retrieved from https://doi.org/10.1080/17457300.2018.1426702

[7] Caruana, R and Mizil, A, An empirical comparison of supervised learning algorithms, ICML '06: Proceedings of the 23rd international conference on Machine learning, Pages 161–168,2006.

[8] Chaudhuri, S., & Saha, S. (2011). Support Vector Machine Classification Algorithm and Its Applications. In Advances in Computing and Communications, Volume 1 (pp. 217-225). Springer. Retrieved from https://doi.org/10.1007/978-3-642-34041-3_27

[9] Chris Albon (2018). Machine learning with Python cookbook: Practical solutions from preprocessing to deep learning. O'Reilly Media.

[10] G. E. A. P. A. Batista and D. F. Silva, “How K-nearest neighbor parameters affect its performance,” *Computer Science*, 2009.

[11] H. Roopa and T. Asha, “A Linear Model Based on Principal Component Analysis for Disease Prediction,” *IEEE Access*, vol. 7, pp. 105314–105318, 2019, retrieved from 10.1109/ACCESS.2019.2931956. https://www.researchgate.net/publication/354921553_Application_of_Regression_Techniques_with_their_Advantages_and_Disadvantages

[12] Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. Retrieved https://doi.org/10.1186/s40537-024-00973-y

[13] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” SN Computer Science, vol. 2, no. 3, Mar. 2021,

retrieved from https://doi.org/10.1007/s42979-021-00592-x.

[14] J. R. Quinlan, “Learning decision tree classifiers,” *ACM Computing Surveys*, vol. 28, no. 1, pp. 71–72, Mar. 1996, retrieved from https://doi.org/10.1145/234313.234346.

[15] Jair Cervantes, 2020.A comprehensive survey on support vector machine classification: Applications, challenges and trends. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925231220307153?casa_token=N0jJ7TcG7SwAAAAA:ElGLMOwpnDz0Ii0k7dnhoiJbA24wxHmmmsZsMnD3CpdPott2uCDV3qy48ZfxWBlDXJLnZ8yQtU#s0045

[16] K. Yeturu, “Machine learning algorithms, applications, and practices in data science,” in Handbook of Statistics, vol. 43, Elsevier, 2020, pp. 81–206.

[17] L. Breiman, *Classification and regression trees*. New York: Routledge, 2017.

[18] M. A. Iqbal, “Application of Regression Techniques with their Advantages and Disadvantages,” *Electron Magazine*,

[19] Maulud and A. M. Abdulazeez, “A review on linear regression comprehensive in machine learning,” *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, 2020.

[20] Mucherino, A., Papajorgji, P.J., Pardalos, P.M. (2009). *k*-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, vol 34. Springer, New York, NY. https://doi.org/10.1007/978-0-387-88615-2_4

[21] Nasteski, V. (2017). An overview of the supervised machine learning methods. Retrieved from https://www.researchgate.net/publication/328146111

[22] PW Wagacha, 2003.Induction of Decision Trees. Retrieved from https://erepository.uonbi.ac.ke/bitstream/handle/11295/44263/decisionTrees.pdfs

[23] Rätsch G, 2004. A Brief Introduction into Machine Learning. Retrieved from https://citeseerx.ist.psu.edu/documentrepid=rep1&type=pdf&doi=fab926b5da15870777607679ebd56985735023d0

[24] S. Angra and S. Ahuja, “Machine learning and its applications: A review,” 2017 International Conference On Big Data Analytics and computational Intelligence (ICBDACI), pp. 167–170, 2017 Retrieved from https://doi.org/10.1109/ICBDACI.2017.8070809.

[25] S. Dash, “Decision Trees Explained — Entropy, Information Gain, Gini Index, CCP Pruning.,” *Medium*, Nov. 02, 2022. https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c

[26] Sandhya N. dhage, Charanjeet Kaur Raina. (2016) A review on Machine Learning Techniques. In International Journal on Recent and Innovation Trends In Computing and Communication, Volume 4 Issue 3.

[27] T. Goswani and G. R. Sinha, Satistical modling in machine leanring: COncepts and Applications. United States: Mara E. Conner, 2022, pp. 1–370.

[28] T. Jijo and A. M. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar.2021, retrieved from https://doi.org/10.38094/jastt20165 .

[29] Tigga, O., Pal, J., & Mustafi, D. (2023). A comparative study of rule-based classifier and decision tree in machine learning. *Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi*. Retrieved from https://sciencetransactions.com/ijascis/uploads/2023/02/j23-40-47.pdf

[30] Vanessa Nilsen,2018. An Introduction to Machine Learning. Retrieved from https://bafflerbach.github.io/test_software_carpentry/files/Module%201%20Lecture%20Notes.pdffrom

[31] Wei-Yin Loh, 2011.Classification and regression trees. Retrieved from https://doi.org/10.1002/widm.8

[32] X. Wu and V. Kumar, The top ten algorithms in data mining. Taylor and Francis Group, LLC, 2009

[33] Ying, X. (2019). *Title of the article*. Journal of Physics: Conference Series, 1168(2), 022022. Retrieved from https://doi.org/10.1088/1742-6596/1168/2/022022.

[34] Zhang, C., & Zhang, M. (2016). Learning k for kNN Classification: A literature review. ACM. Retrieved from https://dl.acm.org/doi/abs/10.1145/2990508