IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

IJARSCT

Impact Factor: 7.301

Volume 3, Issue 8, May 2023

To Study Different Types of Supervised Learning Algorithm

Falguni Ghatkar[1], Sakshi kharche[2], Priyanka Doifode[3], Jagruti Khairnar[4], Prof. Neelam Kumar[5]

Students, Department of Computer Engineering[1, 2, 3, 4]

Professor, Department of Computer Engineering[5]

Shree Ramchandra College of Engineering Pune, Maharashtra, India

Abstract: *This review paper aims to Study of the different type of algorithm is used for the detection. Machine learning has ability to detect or predict the result n basis of the data points data points are the any domain like image, text, video and speech also. The scientific discipline of machine learning enables computers to learn without explicit programming [2]One of the most intriguing technologies that has ever been developed is machine learning. The ability to learn is what, as the name suggests, gives the computer a more human-like quality. Today, machine learning is actively being employed, possibly in a lot more places than one might think [1] Machine Learning is has as multiple application one of Machine learning uses data to detect various patterns in a given dataset it can learn from past data and improve automatically[6].It is a data-driven technology in this review, we will delve into the fundamental concepts and principles of machine learning algorithms, exploring their strengths, weaknesses, and specific use cases. By understanding the intricacies of different algorithms, we can make informed decisions about which method to choose for a given problem or dataset [3]Whether you are a beginner in the field of machine learning or an experienced practitioner, this review aims to provide valuable insights into the diverse landscape of machine learning algorithms.*

Keywords: Support vector machine, Random Forest, Machine Learning, Supervised Learning

I. INTRODUCTION

In World to defend various cyberattacks and computer viruses, lots of computer security techniques have been studied in last decade. For prevents the cyber-attack we use the Machine Learning.Machine learning has revolutionized various industries by enabling computers to learn from data and make accurate predictions or decisions[6] With the increasing complexity of datasets and the demand for advanced predictive modelling. A wide range of machine learning algorithms have been developed. These algorithms serve as powerful tools for extracting meaningful insights and patterns from data, thereby aiding in decision-making, pattern recognition, and automation.

Support Vector Machines (SVM) is a versatile and powerful algorithm that has gained popularity in the field of machine learning. Having explored its capabilities extensively, I find SVM to be an excellent tool for a wide range of classification and regression tasks[11]. While it may not be the go-to choice for every scenario, SVM's strengths make it a valuable addition to any data scientist's toolkit.One of the key strengths of SVM is its ability to handle high-dimensional feature spaces efficiently. By finding the optimal hyperplane that separates different classes with the widest margin, SVM demonstrates robust generalization performance even in the presence of complex and overlapping data distributions[11]. This capability makes SVM particularly effective when dealing with datasets with large numbers of features, such as text classification or image recognition[2]

The Random Forest algorithm is an absolute gem in the realm of machine learning. With its exceptional ability to handle complex tasks and produce reliable results, it has earned its reputation as a game-changer in the field. Having worked extensively with Random Forest, I am delighted to share my experience and highlight the remarkable strengths that make this algorithm stand out.

1.1 Abbreviation and Acronyms

- ML: Machine Learning

DOI: 10.48175/IJARSCT-10256

- IDS: Intrusion Detection System
- A: Algorithms
- SVM: Support Vector Machine
- RF: Random Forest

1.2 Machine Learning:

The field of study known as machine learning enables computers to learn without being explicitly programmed. One of the most intriguing technologies that has ever been developed is machine learning.The ability to learn is what, as the name suggests, gives the computer a more human-like quality Machine learning is being actively used right now, possibly in a lot more ways.

There are two types of Machine learning:

- Un-supervised learning
- Supervised Learning

A. Un-Supervised Learning

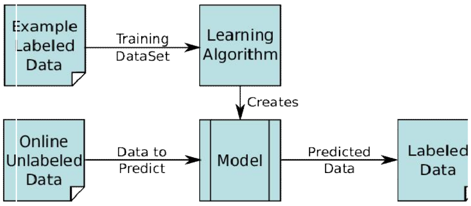
Unsupervised learning is a subfield of machine learning where the model learns patterns or structures in unlabeled data without explicit guidance or supervision. Unlike supervised learning, where the model is provided with labeled examples and aims to learn a mapping between input and output, unsupervised learning focuses on finding inherent patterns or relationships within the data itself [1]

The goal of unsupervised learning is to extract meaningful insights, discover hidden structures, or group similar data points together without any prior knowledge. It is commonly used in tasks such as clustering, anomaly detection, dimensionality reduction, and generative modelling [1]

B. Supervised Learning

In Supervised Learning, algorithms learn from labeled data. The algorithm decides the label to use after comprehending the input.[8]

The algorithm determines which label should be given to new data based on patterns and associating the patterns to the unlabeled new data.Supervised learning (SL) is a machine learning paradigm for problems where the available data consists of labeled examples, meaning that each data pointan associated label Supervised Learning can be divided into 2 categories i.e., Classification and Regression. Classification predicts the category that the data belongs to whereas Regression predicts a numerical valuebased on previous observed data[8]



Supervised learning is a popular approach in machine learning for various tasks, including text detection. In the context of text detection, supervised learning involves training a model using labeled examples, where each example consists of an input image and the corresponding ground truth annotations indicating the location of the text.

Advantages of Supervised Learning:

- **Labeled Data**: Supervised learning relies on labeled training data, which provides explicit information about the input-output relationship. This labeled data helps the model learn patterns and make accurate predictions.
- **Predictive Accuracy**: Supervised learning algorithms tend to achieve high predictive accuracy, especially when trained on a large and representative dataset. They can generalize well to unseen data, making them useful for various applications.

Disadvantages of Supervised Learning:

- **Dependency on Labeled** Data: Supervised learning algorithms require labeled data for training, which can be expensive and time-consuming to obtain. The process of labeling data often relies on human expertise and may introduce biases or errors.
- **Limited Generalization**: The performance of supervised learning models heavily depends on the quality, representativeness, and diversity of the training data. If the training data does not capture the full range of possible scenarios, the model may struggle to generalize to unseen instances accurately.

1.3 Algorithms of Supervised Learning:

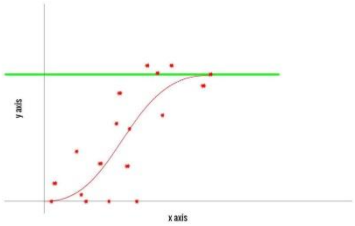
A. Logistic Regression Algorithm

In essence, the predictive model analysis technique of logistic regression uses discrete values for the output (target) variables for a given collection of characteristics or input (X).

This supervised classification approach in machine learning is extremely effective yet straightforward.[11] The logistic regression approach can be used to resolve about 60% of the classification issues encountered world-wide. One of the most popular algorithms for binary classification is logistic regression. A logit function is used to forecast the likelihood that a binary outcome will occur. Given that it uses the log function to estimate outcome probabilities, it is a specific example of linear regression.[10]

Simply said, linear regression uses the results of a second variable to predict the scores on a first variable.

The projected variable is referred to as the Criterion Variable. The element. They succeeded to record 98.3% as accuracy rate based on this methodology.



Advantages:

- **Simplicity**: Linear logistic regression is relatively simple to understand and implement. It is a straightforward extension of linear regression to predict binary outcomes.
- **Interpretable coefficients**: Linear logistic regression provides coefficient estimates for each predictor variable, allowing you to interpret their impact on the probability of the outcome. These coefficients can provide insights into the relationship between the predictors and the outcome.

Disadvantages:

- **Linearity assumption:** Linear logistic regression assumes a linear relationship between the predictor variables and the log-odds of the outcome. If the relationship is highly nonlinear, the model may not fit the data well and result in poor predictions.
- **Independence of observations**: Linear logistic regression assumes that the observations are independent of each other. If there is dependence or correlation among the observations, the model assumptions may be violated, leading to biased or inefficient coefficient estimates.

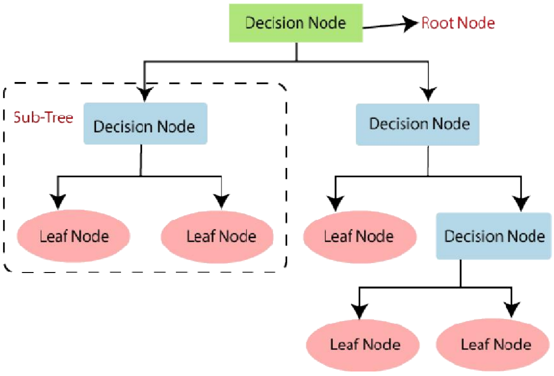
B. Decision Tree Algorithm

A supervised learning approach known as a decision tree is frequently employed in classification tasks.It works for both categorical and continuous input and output variables [7]

According to the most important differentiator or splitter in the input variables, we divide the sample into two or more homogeneous groups (or sub-populations) with this method.

DOI: 10.48175/IJARSCT-10256

In decision tree internal node represents a test on the attribute, branch depicts the outcome and leaf represents decision made after computing attribute.[2] The general motive of using Decision Tree is to create a training model which can be used to predict class or a value of target variables by learning decision rules inferred from prior data (training data).
Compared to other classification methods, the Decision Tree approach is quite simple to understand.
By representing the problem as a tree, the Decision Tree method attempts to find a solution. Each leaf node of the tree corresponds to a class label, whereas each internal node of the tree relates to an attribute [7]
Algorithm that produces a number of useful rules to find incursions. They succeeded to record 97.65% as accuracy rate based on this methodology



Advantages :

- **Interpretable and Easy to Understand**: Decision trees provide a clear and intuitive representation of the decision-making process. The tree structure with nodes and branches can be easily interpreted and visualized. This makes decision trees particularly useful for explaining the reasoning behind the decisions made by the algorithm.
- **Handling Both Numerical and Categorical Data:** Decision trees can handle both numerical and categorical data without requiring extensive pre-processing. They can automatically handle missing values and outliers, making them robust to data imperfections.

Disadvantages:

- **Overfitting**: Decision trees have a tendency to over-fit the training data, especially when the tree becomes too complex or is allowed to grow too deep. Overfitting occurs when the tree captures noise or irrelevant patterns in the data, leading to poor generalization performance on unseen data.
- **Lack of Robustness**: Decision trees are sensitive to small changes in the training data. A slight variation in the data can lead to a completely different tree structure. This lack of robustness makes decision trees prone to high variance.

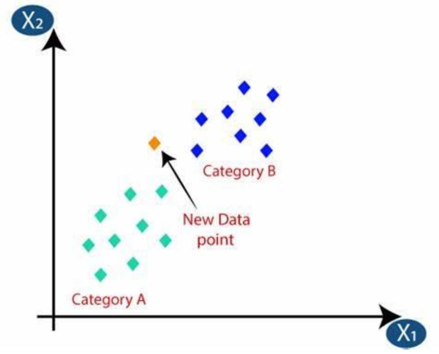
C. K Nearest Neighbor (KNN) Algorithm

The K Nearest Neighbour (KNN) technique calculates the separation between a group of scenarios in the data set and a query scenario. The full training dataset serves as the KNN model. The KNN algorithm will look through the training dataset for the k-most comparable cases when a prediction is needed for an unobserved data instance.
The most comparable examples' prediction attributes are compiled and sent back as the prediction for the unknown instance. The type of data has an impact on the similarity metric. The Euclidean distance can be applied to data with real values. Hamming distance can be applied to category or binary data, among other forms of data. Algorithms that model the problem using data instances (or rows) are known as instance-based algorithms., Instance-based algorithms are those that use data instances (or rows) to model the problem and then use that model to predict outcomes.[4] All training observations are kept in the model through the KNN algorithm, which is an extreme instance-based technique. Its underlying use of competition amongst model components (data examples) in order to arrive at a prediction conclusion makes it a competitive learning algorithm. Each data instance competes to "win" or be the most similar as a

DOI: 10.48175/IJARSCT-10256

result of the objective similarity metric between data instances. An algorithm was utilised to provide a variety of useful rules to find intrusions[11]

Algorithm used to generate a number of effective rules to detect intrusions. They succeeded to record 95.75% as accuracy rate based on this methodology.



Advantages:

- **Simplicity**: KNN is a simple and easy-to-understand algorithm. It's straightforward to implement and interpret the results.
- **No training phase**: KNN is a lazy learning algorithm, which means it doesn't require a training phase. The algorithm directly uses the training data during the prediction phase, making it suitable for dynamic or constantly changing datasets

Disadvantages:

- **Computationally expensive**: During the prediction phase, KNN needs to calculate distances between the query point and all the training points. This computation can be expensive, especially for large datasets.
- **Sensitivity to feature scaling**: KNN uses distance metrics (e.g., Euclidean distance) to determine the similarity between data points. If the features have different scales, it can lead to biased results. Feature scaling is often required to normalize the data

D. Random Forest

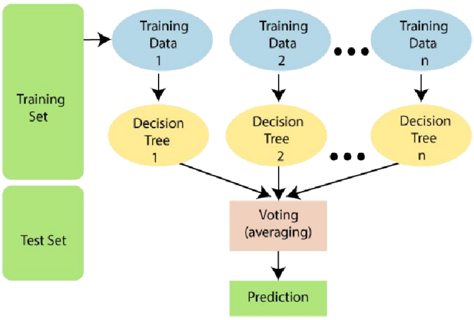
Random Forest algorithm is a supervised classification algorithm. From its name, it is clear that the goal is to haphazardly establish a forest. The more trees a forest has, the more accurate its results will be; conversely, the fewer trees a forest has, the less accurate its results will be. But it's important to keep in mind that building a decision using information gain or an index strategy is not the same as establishing a forest. The procedures of locating the root node and dividing the feature nodes in the Random Forest method operate at random, in contrast to the Decision Tree Approach[10]

There are two stages in Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first stage[1] For applications in classification problems, Random Forest algorithm will avoid the over fitting problem and for both classification and regression tasks, the same random forest algorithm can be used[8]The most crucial features in the training dataset can be found using the Random Forest technique. Random forest takes time to training and testing and also it does not give 96.1% accuracy. For overcome to accuracy part we use support vector machine for increases accuracy of intrusion detection system

Advantages:

- **Robustness**: Random Forest is highly robust to outliers and noisy data. It averages the predictions from multiple trees, reducing the impact of individual outliers or errors in the training data.
- **High accuracy**: Random Forest typically provides higher accuracy compared to individual decision trees. It can handle both classification and regression tasks effectively

- **Feature importance:** Random Forest provides a measure of feature importance, which helps in understanding the relative significance of different input variables. This information can be useful for feature selection and feature engineering.
- **Handles high-dimensional data:** Random Forest can handle datasets with a large number of input features and still maintain good predictive performance. It automatically selects a subset of features at each split, reducing the impact of irrelevant or redundant features

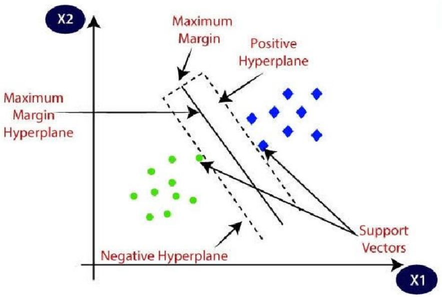


Disadvantages:

- **Model interpretability**: Random Forest is considered a black box model, meaning it is not as interpretable as simpler models like linear regression. It can be challenging to understand the underlying decision-making process of the ensemble.
- **Computational complexity**: Random Forest can be computationally expensive, especially when dealing with large datasets or a high number of trees. Training and prediction times can be slower compared to simpler models.
- **Overfitting**: While Random Forest is less prone to overfitting than individual decision trees, it can still over-fit if the number of trees in the ensemble is too high or if the trees are allowed to grow too deep. Careful tuning of hyperparameters, such as the maximum depth of trees, is required to avoid overfitting

E. SVM

Support Vector Machine (SVM) is a form of supervised learning technique. It is used for both regression and classification purposes however, most of the time it is used in classification problems. Support Vector Machine is a fast and dependable classification method that excels when given a small amount of data[9] The main ideology behind SVM is to create a hyperplane and to classify the dataset given. To isolate the two classes of data points, there are numerous Conceivable Hyperplanes that could be picked. Our objective is to find the plane with the greatest margin, i.e.the greatest separation between data points of the two classes. Expanding the margin enables the future data points to be classified with much more precession.[1]Hyperplanes are those that help in classifying the data. Data points or vectors that fall on either side if the hyperplane can be credited to different classes. If we have two independent features then our hyperplane will be three dimensional. If we have one independent feature then we will have a simple one dimensional hyperplane below is the figure from [3] that shows how hyperplane is built and can be used to classify the data points. H represents the hyperplane. H1 and H2 are the lines drawn parallel to hyperplane such that the distance between these two i.e. the margin is maximum



DOI: 10.48175/IJARSCT-10256

Above fig that shows how hyperplane is built and can be used to classify the data points.

An algorithm was utilised to provide a variety of useful rules to find intrusions. They succeeded to record 99.60% as accuracy rate based on this methodology

Advantages of SVM:

- **Effective in high-dimensional spaces:** SVM performs well even when the number of dimensions is greater than the number of samples. This is known as the "curse of dimensionality" problem, and SVM addresses it through the use of hyperplanes to separate classes.
- **Versatility in kernel selection**: SVM allows the use of different kernel functions, such as linear, polynomial, radial basis function (RBF), and sigmoid. This flexibility enables SVM to handle various types of data distributions and capture complex relationships.
- **Robust against overfitting:** SVM uses a regularization parameter (C) to control the trade-off between maximizing the margin and minimizing the training errors. This helps in preventing overfitting and improves generalization performance.

Disadvantages of SVM:

- **Computationally intensive:** SVM can be computationally expensive, especially when dealing with large datasets. Training an SVM model can take a significant amount of time and memory, particularly for non-linear kernels and high-dimensional data.
- **Difficult to interpret**: SVM produces a black-box model, meaning it can be challenging to interpret the learned relationships between features and the target variable. Understanding the contribution of individual features to the decision boundary is not straightforward.

Sensitivity to parameter tuning: SVM has several parameters, such as the choice of kernel, kernel-specific parameters, and the regularization parameter (C). Selecting appropriate values for these parameters can be crucial for achieving good performance. Improper parameter tuning can lead to suboptimal results or even poor generalization.

II. CONCLUSION

This review paper has provided a comprehensive analysis of various machine learning algorithms and their applications And Advantages and Disadvantages. A Concepts and principles underlying machine learning, including supervised, unsupervised, and reinforcement learning. It then delved into an extensive discussion of popular machine learning algorithms such as linear regression, decision trees, support vector machines, random forests Among Others.

Each algorithm was critically evaluated based on its strengths, weaknesses, and suitability for different types of data and problem domains. The review also highlighted the achieving optimal algorithm performance.

REFERENCES

[1]. C. R. Srinivasan, B. Rajesh, P. Saikalyan, K. Premsagar, and E. S. Yadav, “A review on the different types of INetwork (Intrusion network),” J. Adv. Res. Dyn. Control Syst., vol. 11, no. 1, pp. 154–158, 2019.

[2]. "A Hybrid Unsupervised Clustering-Based Anomaly Detection Method", Guo Pu, Lijuan Wang, Jun Shen, and Fang Dong, TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007-0214 02/11 pp146–153DOI:10.26599/TST.2019.9010051Volume 26, Number 2, April 2021.

[3]. "Anomaly Detection Using Machine Learning Approaches", Mausumi Das Nath, TapalinaBhattasali. St. Xavier’s College (Autonomous), Kolkata, India, m.dasnath@sxccal.edu, tapalina@sxccal.edu.Azerbaijan Journal of High Performance Computing, Vol 3, Issue 2, 2020

[4]. "Anomaly Detection for Cybersecurity of the Substations", Chee-Wooi Ten, Member, IEEE, Junho Hong, Student Member, IEEE, and Chen-Ching Liu, Fellow, IEEE. IEEE TRANSACTIONS ON SMART GRID, VOL. 2, NO. 4, DECEMBER 2011.

[5]. W Hu, Y Liao, VR Vemuri,(2003) “Robust anomaly detection using support vector machine”. academia.edu (2003).

DOI: 10.48175/IJARSCT-10256

[6]. J. Joyia, R. M. Liaqat, A. Farooq, and S. Rehman, “Internet of Medical Things (IOMT): Applications, benefits and future challenges in healthcare do- main,” J. Commun., vol. 12, no. 4, pp. 240–247, 2017.

[7]. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, “INetwork for smart cities,” IEEE Internet Things J., vol. 1, no. 1, pp. 22–32, Feb. 2014.

[8]. B. Karbab, M. Debbabi, A. Derhab, and D. Mouheb, “Android malware detection using Machine learning on API method sequences,” Dec. 2017, arXiv:1712.08996. [Online]. Available: https://arxiv.org/abs/1712.08996

[9]. S. Jabbar, K. R. Malik, M. Ahmad, O. Aldabbas, M. Asif, S. Khalid, K. Han, and S. H. Ahmed, “A methodology of real-time data fusion for localized big data analytics,” IEEE Access, vol. 6, pp. 24510–24520, 2018.

[10]. F. Ullah, J. Wang, M. Farhan, M. Habib, and S. Khalid, “Software plagiarism detection in multiprogramming languages using machine learning approach,” Concurrency Comput.,Pract. Exper., to be published.

[11]. D.-K. Chae, J. Ha, S.-W. Kim, B. Kang, and E. G. Im, “Software plagiarism detection: A graph-based approach,” in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage., Nov. 2013, pp. 1577–1580.

[12]. Y. Akbulut and O. Do¨nmez, “Predictors of digital piracy among Turkish un- dergraduate students,” Telematics Inform., vol. 35, no. 5, pp. 1324–1334, 2018

DOI: 10.48175/IJARSCT-10256