Analysis of Machine Learning Techniques -A Survey

M.Balamurugan
Dept of CSE
Sri Sairam Engineering college
Chenna,India,
Research Scholar
Vels Institute of Science,Technology & Advanced Studies
balamurugan.cse@sairam.edu.in

Dr.Karthika.R.A
Dept of CSE
Vels Institue of Science,Technology & Advanced Studies
Chennai,India,
karthika.se@velsuniv.ac.in

Article Info
Page Number: 3472-3481
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract

A wide variety of strategies have been developed to help machine learning systems make sense of their broad array of possibilities. Supervised, unsupervised, and reinforcement learning are only a few of the many types of machine learning. Various strategies are put to the test on a variety of data sets. We shall compare the performance of supervised techniques in this study.

I. Introduction

With the aid of a new technique known as machine learning, computers are now capable of learning on their own. Various strategies are used by machine learning to develop mathematical models and predict future events. Automated tagging on social media sites like Facebook or email are just a few examples of tasks that artificial intelligence can do. Using historical data, a machine learning system may be programmed to generate predictions about future datasets, which it can then do on its own volition. For a model to accurately predict output, the amount of data required is directly proportional to the accuracy with which that data can be predicted.

To avoid having to write code to solve a complex problem requiring predictions, we may just feed data to generic algorithms, and the machine will generate reasoning based on the input and anticipate the conclusion. Machine learning has brought a fresh way of thinking about this problem. As time goes on, machine learning becomes more and more significant. Machine learning has become a need in many industries because of its capacity to do tasks that are beyond the capabilities of a human person. Computer technologies and machine learning must be used to speed up our processes since we cannot access that much data manually.

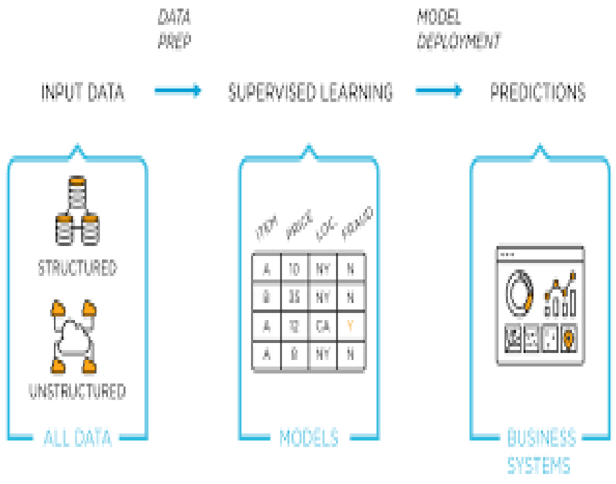
With a lot of data, you can build powerful machine learning algorithms that can do all the work for you by themselves. The cost function can tell us how much data the algorithm can process. Approaches based on machine learning may allow us to save both time and money. The value of machine learning may be shown by looking at how it has been put to use. Just a few of the current applications of machine learning include self-driving cars, cyber-fraud detection and Facebook friend recommendations. It is possible for companies like Netflix and Amazon to better anticipate what their consumers desire by using machine learning algorithms that are trained on vast amounts of data.

II. Machine learning techniques

Machine learning may be broadly categorized into three types. Three methods of learning exist: teaching by a teacher, self-study, and reinforcement by a teacher.

Training a machine learning system is done by feeding it data that has been tagged, and it predicts the output based on what it has learned so far. Separate methods for unsupervised and supervised learning are used in supervised learning. Classification, regression, and reciprocal connections are all discussed.

We put a set of data through the model after it has been trained and processed to see whether it can accurately predict the result. If this occurs, the system makes the necessary adjustments to its model. The goal of supervised learning is to establish a connection between the data that is fed in and the results that are produced. Learning under supervision is similar to how kids learn in class since it depends on constant supervision. Many reviews are presented in literature by many researchers with respect to ML and IoT in different domain.[5-10]. This analysis will surely enable the researchers with the idea of ML technique in different applications. [11-13]



ITEM	PRICE	LOC	FRAUD
A	10	NY	N
B	35	NY	N
A	12	CA	Y
A	8	NY	N

Fig.1. Supervised Learning

Unsupervised learning:

A machine is considered to be using unsupervised learning when it picks up new skills on its own. As a result, algorithms have to work with unlabelled data and function without human intervention. Working with unsupervised data, the goal is to generate new features or a group of objects with similar patterns.

It is impossible to predict what will happen if you learn anything on your own without supervision. When confronted with a massive amount of data, the computer makes an effort to extract any useful information. Clustering and Association are two of the algorithms included in this class.

Reinforcement learning: It is a teaching method centred on providing students with feedback, in which they are praised for good work and chastised for mistakes. Agents learn and improve their performance via the use of these feedbacks. In systems that employ reinforcement learning, the agent's interactions and learning are always occurring. The ultimate purpose of an agent is to maximise reward points in order to increase its overall performance.

When a robot learns to move its limbs on its own, it is an excellent illustration of reinforcement learning.

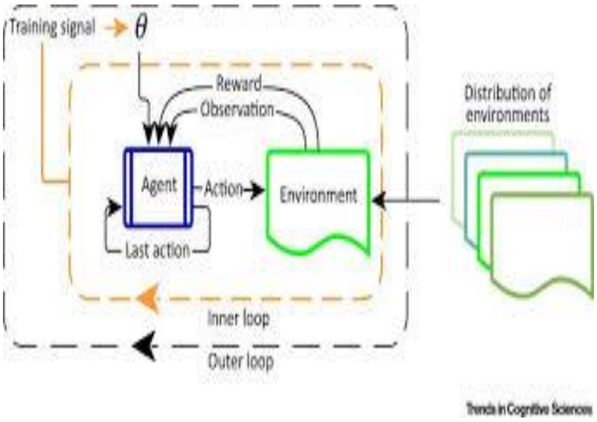


Fig.2. Reinforcement Learning

A. Supervised Learning Techniques

1. Support Vector Machine Algorithm: Supervised Learning [1] methods such as SVM are widely utilized in classification and regression issues. SVM is one of the most widely used algorithms. The majority of the time, it is used in Machine Learning to help with Classification problems. So that we may more easily categorize new data points in the future, the SVM approach generates the best line or decision boundary to split n-dimensional space into classes. You may think of this optimum choice line as a hyperplane. The hyperplane is constructed using extreme points and vectors selected using SVM. A SVM is an algorithm that analyses extreme instances, which are called "support vectors. "Facial recognition, picture classification, text classification, and other tasks are all possible using SVM.

SVM can be of two types:

Linear SVM: For datasets that can be separated into two different classes by drawing a single straight line, an SVM classifier called the Linear SVM is utilized. Non-linear SVM: In the case of non-linearly separated data, the classifier utilized is known as a Non-linear SVM classifier, which indicates that a dataset cannot be categorized by drawing a straight line across the points.

Hyperplane: Decision boundaries in n-dimensional space may be divided into numerous lines, but we must choose the one that is most useful for classifying data points. We call this optimal boundary the hyperplane of SVM.

This indicates that if there are just two features in the dataset, the hyperplane will be straight and the size of the hyperplane will change accordingly. It is a two-dimensional plane if all three properties are present. Hyperplanes with maximum margins show that the largest distance between data points is as big as it can be.

Support Vectors:

It is the collection of data points or vectors that are close to the hyperplane and impact its location that is referred to as a support vector. As a result of this, these vectors are referred to as Support vectors.

SVM is applied on the Kaggle dataset of heart disease and the Test Accuracy 86.89%

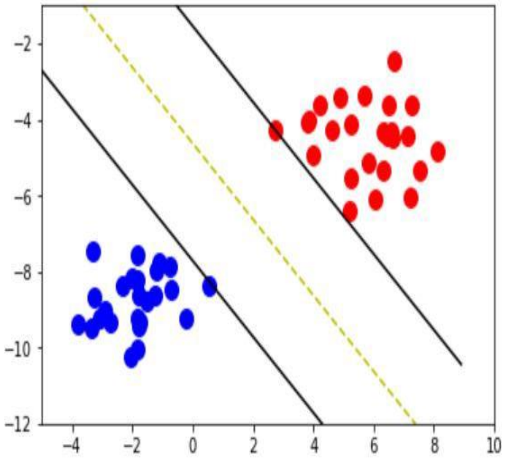


Fig.3. Support Vector Machine

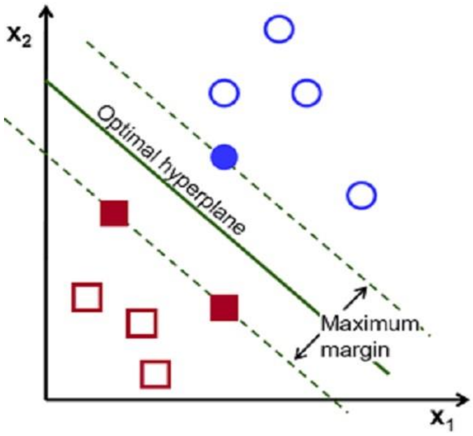


Fig.4. Support Vector Machine

2. Random Forest Algorithm:

The supervised learning approach includes the widely used Random Forest algorithm [4]. This method may be used to situations requiring classification as well as regression analysis in machine learning. An ensemble of classifiers is used to combine and enhance the model's performance, and this approach is based on the notion of ensemble learning.

You may use an algorithm called a "Random Forest" to classify a dataset by using many decision trees on various subsets to get a more accurate prediction. It takes projections from several trees, rather than relying on a single one, and then calculates the final outcome based on those predictions that received the most votes.

Overfitting may be avoided by having a larger number of trees to choose from when making a model.

If you have N trees to combine into a random forest initially, then you can forecast the outcome of each one of them using Random Forest. The working procedure is shown in the following diagram and steps:

The first step is to choose K random data points from the training set.

Step two is to construct the decision trees using the provided data points (Subsets). Decide the number of decision trees you wish to generate. The category that obtains the most votes from each decision tree should be allocated to new data points.

RF APPLICATIONS

Random forest is most often utilized in four sectors:

In the financial industry, this technique is most often used to determine the risk of a loan.

Medicine: Using this strategy, trends and risks of sickness may be discovered.

Method: In the future, this system might help us locate areas with similar land use. Using this method, you can spot emerging market tendencies.

Benefits of Using Random Forests

Classification and regression problems may be handled with Random Forest.Data sets with great dimensionality are no problem for this software. It improves the model's accuracy and avoids overfitting.

Disadvantages of Random Forest

Because of this, random forests may be used for both classifying data and predicting future outcomes.

Random Forest is applied on the Kaggle dataset of heart disease and the Test Accuracy of Random Forest: 88.52%.

3.Decision Tree Algorithm:

Both classification and regression issues may be solved using supervised learning techniques like Decision Tree, The majority of the time, nevertheless, it is used to sort data. The dataset's properties are represented by internal nodes, while decision rules are represented by branches, with each leaf node representing a result. This is a classifier based on a tree.

Nodes in the decision tree are called "leaves," and the Decision Node and Leaf Node are two examples. Instead of having branches to follow them, leaf nodes are the direct effect of these decisions.[3]

Decisions or tests are made based on the dataset's characteristics.It is a visual depiction of all the potential answers to a particular issue or choice.The "decision tree" gets its name from the fact that it extends outward from the root node like a tree.To create a tree, programmers utilize the CART algorithm, an acronym for Classification and Regression Tree algorithm.According to the answer (yes/no) to a query, a decision tree separates itself into subtrees.

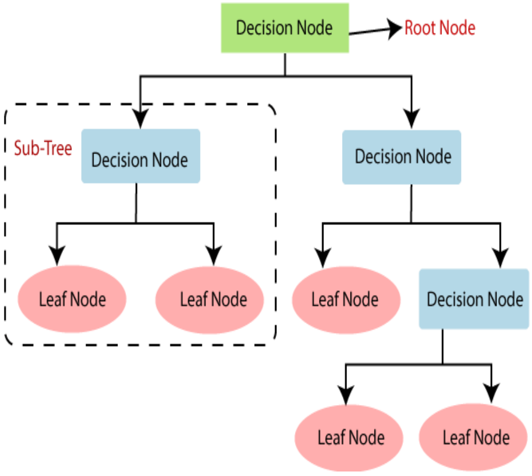


Fig.5. Decision Tree

Test Accuracy of Decision Tree Algorithm for dataset of heart dataset is 69%

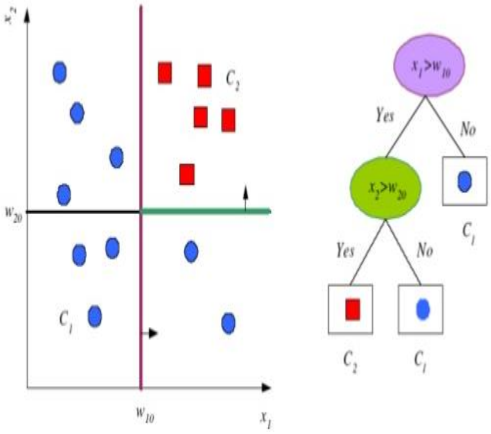


Fig. 6. Accuracy for Decision Tree

4.Naive Bayes Algorithm:

Text categorization using a large, multi-dimensional training set makes extensive use of this technique.

NBA makes predictions based on the likelihood of a given item.[2]. Spam filtering, sentiment analysis, and article classification are all instances of the Nave Bayes Algorithm.

Implies that the existence of one trait is unrelated to the occurrence of other features. When a red, spherical, and sweet item is found based on colour, shape, and flavour, it is known as an apple. As a result, each characteristic may stand alone and be used to help identify the fruit for what it is.

Bayes: It is known as Bayes since it is based on the Bayes' Theorem premise.

Bayes' Theorem:

It is sometimes called Bayes' rule or Bayes' law, and it may be used to calculate the likelihood that a hypothesis is correct given certain previous information. The conditional probability has an effect on this.

The formula is:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Accuracy for NBA is 86.89%

5.Regression Analysis in Machine learning:

One or more independent variables are used in regression analysis to represent the connection between a dependent (target) and an independent (predictor) variable. We can better understand how one independent variable affects another when other independent variables are maintained constant using regression analysis. Temperature, age, income, price, and other real-world variables may all be predicted using this model.[3]

One of the most often used methods for discovering correlations between variables is regression, which can then be used to predict an output variable using one or more predictor variables. You may use it to make predictions, forecast and model time series as well as find out how one variable causes another.

Machine learning models may create predictions about the data by generating a graph between variables that best suit the provided data points in Regression. "Regression displays a line or curve that passes through all the datapoints on the target-predictor graph in such a manner that the vertical gap between the data points and the regression line is smallest," as the saying goes. The distance between data points and the line reveals whether a model has captured a strong link or not.

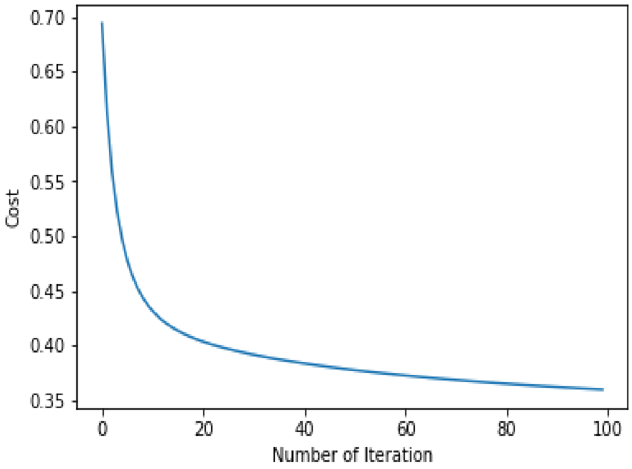


Fig.7. Accuracy graph1 for KNN

Test Accuracy is **86.89%**

6. K-Nearest Neighbor Algorithm

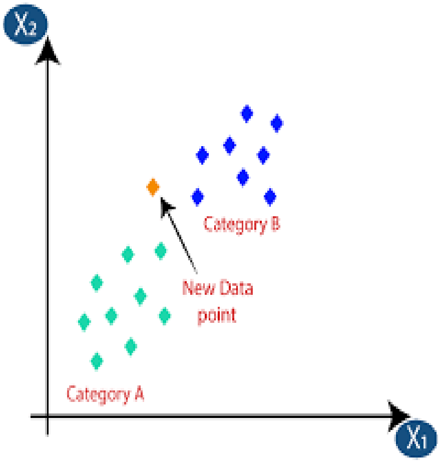


Fig.8. KNN

K-Nearest Neighbor is one of the simplest Machine Learning algorithms.[1]

Using a K-NN method, a new case/data is compared to the current instances and categorized into a most comparable category.

New data points are classified using K-NN algorithm based on their similarities to previously-classified data points, which are stored in a database. By employing K-NN algorithm, fresh data may be quickly categorized into an appropriate category as it arises.

KNN is applied on the Kaggle dataset of heart disease having Accuracy is 88.52%

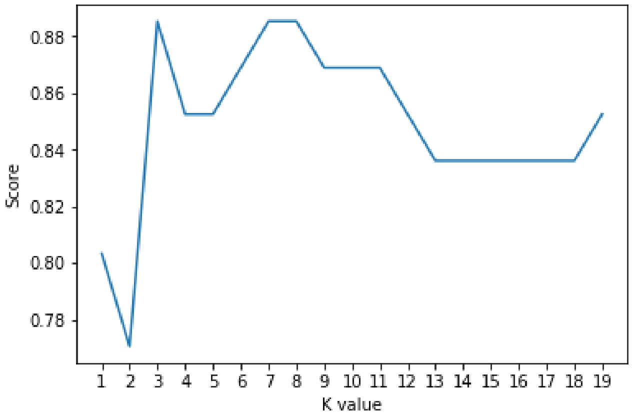


Fig.9. Accuracy graph2 for KNN

III. Results

The data which classified if patients have heart disease or not according to features in it. The data is used to create a model which tries predicting if a patient has this disease or not. The above said techniques are used to train the models and the comparisons of it are analyzed. Python and Kaggle datasets are used.

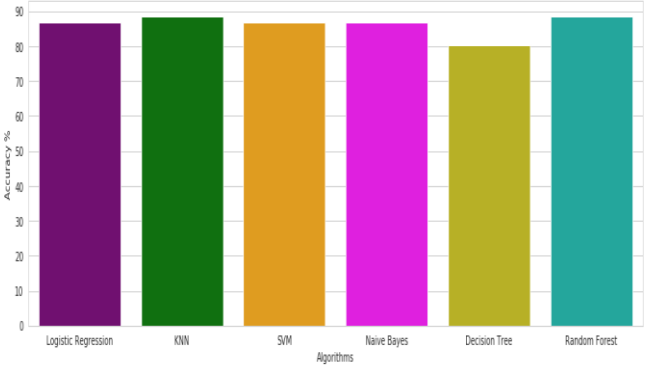


Fig.10. Result values for Algorithms

Random forest and KNN has the highest efficiency of 88.52% followed by the other techniques.

REFERENCES

1. “Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare”, Jian Ping Li; Amin Ul Haq; Salah Ud Din; Jalaluddin Khan; Asif Khan; Abdus Saboo October 2020, IEEE Access,
2. ” Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques”, Pronab Ghosh; Sami Azam; Mirjam Jonkman; Asif Karim; F. M. Javed Mehedi Shamrat; Eva Ignatious; Shahana Shultana; Abhijith Reddy Bee January 2021 IEEE

3. ” HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System”October 2021 IEEE Norma Latif Fitriyani; Muhammad Syafrudin; Ganjar Alfian; Jongtae Rhee.
4. ” Heart Disease Using Prognosis Machine Learning Classification Techniques”,May 2021 IEEE, Mohammed Nowshad Ruhani Chowdhury; Ezaz Ahmed; Md. Abu Dayan Siddik; Akhlak Uz Zaman.
5. Ajitha, P.Sivasangari, A.Gomathi, R.M.Indira, K."Prediction of customer plan using churn analysis for telecom industry",Recent Advances in Computer Science and Communications,Volume 13, Issue 5, 2020, Pages 926-929.
6. "Sivasangari A, Ajitha P, Rajkumar and Poonguzhali,” Emotion recognition system for autism disordered people”, Journal of Ambient Intelligence and Humanized Computing (2019)."
7. Ajitha, P., Lavanya Chowdary, J., Joshika, K., Sivasangari, A., Gomathi, R.M., "Third Vision for Women Using Deep Learning Techniques", 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315196
8. Sivasangari, A., Gomathi, R.M., Ajitha, P., Anandhi (2020), Data fusion in smart transport using convolutional neural network”, Journal of Green Engineering, 2020, 10(10), pp. 8512–8523.
9. A Sivasangari, P Ajitha, RM Gomathi, "Light weight security scheme in wireless body area sensor network using logistic chaotic scheme", International Journal of Networking and Virtual Organisations, 22(4), PP.433-444, 2020
10. Sivasangari A, Bhowal S, Subhashini R "Secure encryption in wireless body sensor networks",Advances in Intelligent Systems and Computing, 2019, 814, pp. 679–686
11. Sindhu K, Subhashini R, Gowri S, Vimali JS, "A Women Safety Portable Hidden camera detector and jammer", Proceedings of the 3rd International Conference on Communication and Electronics Systems, ICCES 2018, 2018, pp. 1187–1189, 8724066.
12. Gowri, S., and J. Jabez. "Novel Methodology of Data Management in Ad Hoc Network Formulated Using Nanosensors for Detection of Industrial Pollutants." In International Conference on Computational Intelligence, Communications, and Business Analytics, pp. 206-216. Springer, Singapore, 2017.
13. Gowri, S. and Divya, G., 2015, February. Automation of garden tools monitored using mobile application. In International Confernce on Innovation Information in Computing Technologies (pp. 1-6). IEEE.