International Journal of Engineering and Techniques - Volume 10 Issue 4, August 2024

RESEARCH ARTICLE OPEN ACCESS

Machine Learning Algorithms for solving Generic Real Life Problems - A Comparative Study

Sandeepa Oraon[1], Samreen Anjum[2], Suman Kumari[3], Siba Mitra[4]

1(Student, CSE Dept., Birla Institute of Technology, Mesra, Lalpur
Email: sandeepa2004oraon@gmail.com)
2 (Student, CSE Dept., Birla Institute of Technology, Mesra, Lalpur
Email: samreenanj2@gmail.com)
3 (Student, CSE Dept., Birla Institute of Technology, Mesra, Lalpur
Email: sinhasuman0305@gmail.com)
4 (Assistant Professor, CSE Dept., Birla Institute of Technology, Mesra, Lalpur
Email: sibamitra@bitmesra.ac.in) *corresponding author

Abstract:

Nowadays most of the technological research work is being conducted to make human life easier and for doing that one of the promising field of research is machine learning. Machine learning is used for training a computer with huge amount of data and preparing it to perform decision making. There are various algorithms in the area of machine learning that are used for training and testing a machine learning model. In this research work a thorough theoretical analysis of multiple machine learning algorithms namely SVM, KNN, Naïve Bayes, Decision Tree and Neural Network are carried out. A thorough related work exhibiting the current state of technology is demonstrated in this article. Moreover this work also presents a comparative study cum analysis of the aforesaid algorithms. The parameters considered over here are the functionality of the algorithm, handling huge data sets along with the missing value problems, over-fitting and outlier detection issues and how different algorithms handle continuous and discrete nature of the data.

Keywords —**ML algorithms, SVM, Naïve Bayes, Decision Tree, KNN, Neural Network, Linear regression.**

I. INTRODUCTION

Machine learning (ML) is one of the technology that is growing very fast and it has lots of applications in the field of information technology. The use of machine learning is done for intelligent data analysis in order to make some automatic decision making. Nowadays the increasing availability of data had made it possible to make a computer learn and help human to take meaningful decisions in medical, agricultural, educational and business fields. The development of a Machine Learning model is done in several phases known as ML lifecycle [14]. The machine learning process involves handling of huge volumes of data, analyzing them, and finally establish relationships with various parameters and factors. One of the most crucial phases in ML lifecycle is selecting the most suitable algorithm. The choice of our algorithm has a great impact on the performance of our model.

ML algorithms are used for generating computational models that can predict or forecast based on any type of data without any programming. Different algorithms use different method of creating model. No two algorithms give the same score when their performance is measured using different metrics. To select the most suitable algorithm, the developer must have an understanding of all the ML algorithms. By analyzing the characteristics of the dataset, we can decide which algorithm is most suitable to develop the model.

ML algorithms can be of three types, [7] supervised, unsupervised and reinforcement. Supervised learning algorithms work upon datasets with both input and output variables while unsupervised learning is the opposite that is it works on unlabeled data. Semi-supervised learning algorithms is the intermediate between the above two algorithms in the sense that use both kinds of datasets mentioned above. Reinforcement learning algorithms trains a model by awarding for correct prediction and penalizing for incorrect prediction. The aim of this paper to give a basic understanding of the different Supervised ML models- Decision Trees, Neural network, SVM, Naive Bayes, KNN, and Linear Regression [3, 5, 10, 7]. The characteristics of these algorithms are first described in detail following which these algorithms are analyzed using some significant parameters. These parameters include the tasks performed by the algorithms, capability to handle missing values, large datasets, anomalous data, outliers and whether they are prone to over fitting.

Various researchers have done different types of comparative studies on many ML algorithms to show how different algorithms function. The article [8] presents a research work on application of different ML algorithms on same data set to predict breast cancer prognosis. The researchers in [12] also have done a research work to present a comparative study and proposing a model for adaptive traffic management. They have used around eight ML algorithms on traffic signal data of Karnataka for their prediction. Sriram et al have proposed [15] an ML model for Parkinson Disease detection using voice data of the patients, which is an example of handling continuous data and fitting ML algorithm in it. The reviewers in [2] have presented a review work in their research article which researches and reviews on the effect of ML algorithms in real-life healthcare diagnosis interventions. In the text book mentioned in [7] presents a vivid study on all the machine learning algorithms are presented, along with the mathematical explanation

By having such comparison, we can figure out the plus points and negative points of an algorithm for a particular dataset. In the next section a literature survey about the current topic is presented.

II. RELATED WORK

In this section a thorough review of the current state of the technology is presented. According to [1], machine learning (ML) is needed for performing any kind of analysis. Machine learning can be said as new trend of deep learning. Their paper mainly focuses on making the good machine learning system. Researchers mentioned in their paper that machine learning have to go through four steps- feature extraction, algorithm selection of machine learning, training and evaluating the model, and applying trained model in real world problem. To build a good machine learning model one must need a data preparation capability, algorithms, scalability, automation, and ensemble modeling. They have divided machine learning into two categories one is shallow learning and other one is deep learning. However, they have also mentioned some challenges in the field of ML such as availability of suitable data, data bias, limitation of resources, privacy of data, and evaluating the problems.

In the research work by Sarker et al. in [4] have provided a detailed overview of various ML algorithms that can be used to make an application smarter and more capable. Researchers have discussed popular area where ML is used and how useful those techniques are in real-world such as cyber security, healthcare, e-commerce, agriculture, etc. are mentioned in the research work. Researchers have discussed various ML techniques and their ability to solve real-world problem and highlighting the need of both quality data and efficient algorithm to build successful models. ML algorithm require to train with real-world data for decision- making. Some challenges are in this area are pointed out and several potential research opportunities are suggested that could lead to better solution in various applications.

Research work presented at [13] shows global interest in AI and ML and the significant growth in ML usage over the past five years due to the advancement in ML technology, increased research, and new national policies in countries like Italy, China, USA, Israel, UK, and the middle east. They have also addressed the challenges and regulations for using ML technology. In their paper they have discussed how ML can solve many day-to-day problems in real-life also. They have emphasized that a successful ML technique relies on both quality data and the performance of the algorithm.

As per [9] ML has become a topic in Computer science, especially during COVID-19. Researchers working in AI and ML field have been working hard to improve the accuracy of ML algorithms and enhance machine intelligence. The main goal of their paper is to provide the survey of ML and deep learning applications across different domains. Their survey discussed the development, key features, and difference between ML and deep learning. They have mentioned how ML and deep learning have impact on our everyday life, from product recommendations in online shopping to updating pictures on social media. Researchers have reviewed existing and ongoing application of ML and deep learning in several area. As per their study they have concluded that ML uses algorithm to analyze and interpret data, learn from it, and make the best possible decisions based on that learning.

Machine learning algorithms are useful even in medical fields. Models for predicting disease can be trained using various algorithms. The research done by researchers in [16] have applied ML methods that could detect disease. They have used algorithms like SVM, Decision Tree, KNN, Naïve Bayes classifier among these Naïve Bayes gives the highest accuracy. They have a front end which takes the input from user and display the outcome calculated by model using Naïve Bayes algorithm since it has the highest accuracy. Their study conducts an in-depth analysis and classification of various ML models utilized in the diagnosis and identification of thyroid illness. ML plays a crucial role in accurately identifying thyroid conditions, enabling early detection and effective treatment planning. Their study aims to highlight their efficiency, strengths, and area for improvements in the context of thyroid illness diagnosis.

The research demonstrated in [14] machine focuses mainly on life cycle of machine learning. According to them by following steps of life cycle of a machine learning, is structured way to building a working model which predicts the correct value which includes developing, deploying, maintaining, and monitoring the machine learning model. Machine learning life cycle starts with collection of relevant data which serves as the foundation of entire project, following data collection, the pre-processing stage involves cleaning and organizing the data to ensure its quality and suitability for analysis. Next is data wrangling and transformation, then model training and testing. The final stage involves the deployment and maintenance of the model. Their research documents each step of the machine learning life cycle. By offering a detailed overview, their study aims to valuable resource for developing robust machine learning technique for decision-making in various applications.

III. MACHINE LEARNING ALGORITHMS

The algorithms form the basic foundation of any ML models that are used in the field of healthcare, natural language processing, fraud detection system and many other areas. The selection of an algorithm, for a particular application area, depends on the testing and evaluation of the data set already available in hand. These algorithms are typically categorised as supervised machine learning algorithms, unsupervised ML algorithms and reinforcement machine learning algorithms. In this research work the main target area is supervised machine learning algorithms. Supervised machine learning algorithms uses labelled data sets for its action. The main focus is to identify the nature of mapping in between input data and output label. Hence it performs prediction on new and unseen data. In the following sub sections various machine learning algorithms viz. Support Vector Machine (SVM) algorithm, K-Nearest Neighbours (KNN) algorithm, Naïve Bayes algorithm, Linear Regression algorithm, Decision Tree algorithm and Neural Network (NN) algorithm.

A. Support Vector Machine Algorithm

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression task. It mainly helps to identify the correct category of the new data. It works by finding the optimal hyper-plane in an n-dimensional space that can separate the data points in different classes in the feature space.

The prediction of output is done by comparing the given data to different categories and output depends on which side they fall. SVM works by creating a hyper-plane which divides the labelled data into different classes in an n-dimensional space and ensures there is a large gap between the categories. The hyper-plane is determined by dividing data into two parts, each part contains similar type of data. A dimension of a hyper-plane is one less than of the original space. The figure 1 illustrates the idea very clearly.

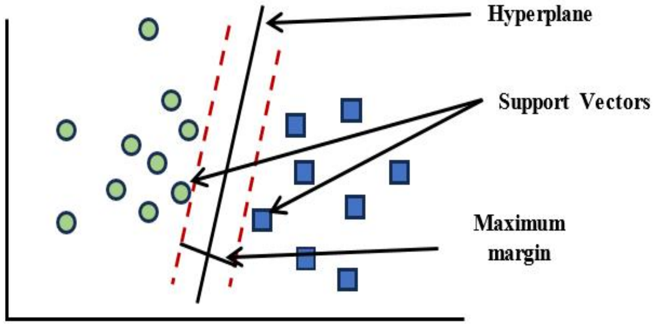


Fig. 1. Support Vector Machine (SVM)

SVM transforms the non-linear data as shown in figure 2 (a) into a higher-dimension space. This transformation allows the algorithm to find a linear decision boundary in the new space. After transformation, the SVM identifies the optimal hyper-plane that separates the classes with maximum margins, which shows that the data is now linearly separable as shown in figure 2 (b) below.

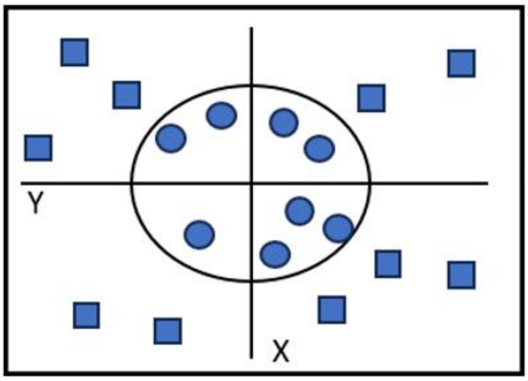


Fig. 2 (a) Non-linear Data Relationships

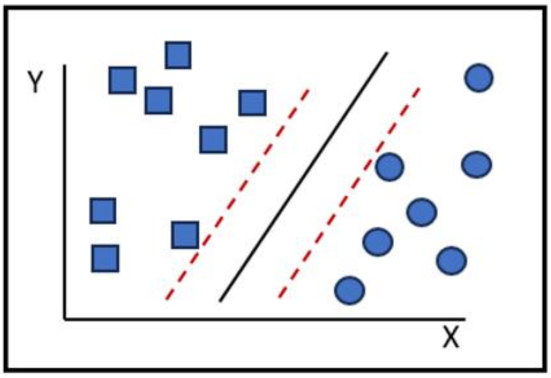


Fig. 2 (b) Linear Data Relationships after application of SVM

In figure 2 (a), we see two types of data points, represented by square and circle. These points are not linearly separable, means there is not a straight line between square and circle that can separate the squares from circle. The arrangement of the data points shows a non-linear relationship, which makes it difficult for a linear classifier to distinguish between the two classes. In figure 2 (b), the same data as figure 2 (a) are shown but this time they are separated by a straight line or hyper-plane and two parallel support vectors to hyper-plane are drawn which only touches the nearest data point from the hyper-plane. The support vectors lie closest to the decision boundary and are crucial as they define the margin, determine the position and orientation of the hyper plane.

B. K-Nearest Neighbor (KNN) Algorithm

KNN algorithm is one of the simplest type of machine learning algorithm, which uses supervised learning process. KNN algorithm can be used both for regression and classification problem, but very commonly used for classification problems. In this technique it stores the original data source initially and whenever it gets any new data point it classifies the same into a category that is near to similar to the new data.

The algorithm works step wise by starting with finding out the k number of neighbours of a data point, then it computes the Euclidean distance with these k neighbors belonging to different categories. Now the numbers of neighbors are calculated for each category and the eventually the new data point will belong to that particular category for which it has the highest neighbor count. This algorithm works well if there are missing values in the data set. The figure 3 below shows that a yellow square is considered as a new data point, which has to be classified in either category A or category B. However, after applying KNN the new data point was found to have 3 neighbors from category A and 2 neighbors from category B as mentioned in figure 4. Therefore according to KNN now the new data point will be categorised as a member matching with category A.

The intrinsic steps of KNN are data pre-processing, fitting algorithm to the training set, then prediction of the test result and finally evaluate the accuracy. KNN algorithm is found to be robust with noisy training data and also comfortable to handle huge data set. One major limitation of KNN is its computation time for finding out the neighbor distances.

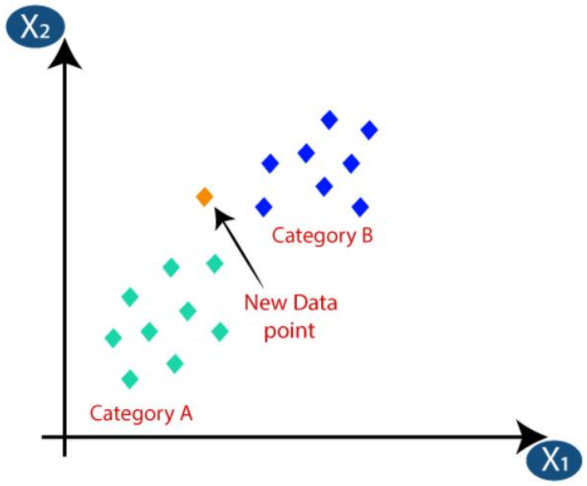


Fig. 3. New data point before application of KNN

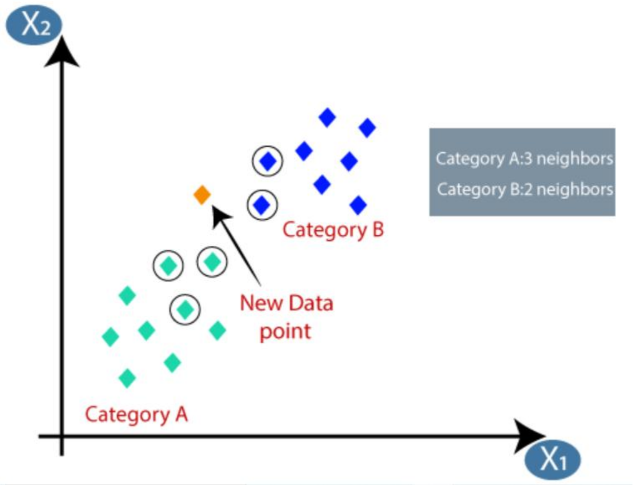


Fig. 4. Nearest Neighbor count using KNN

C. Naïve Bayes Algorithm

The Naïve Bayes algorithm for machine learning is used for solving classification problems and is propounded on the basis of Bayes theorem. It is simple and effective in making quick decisions when the text data is used in a voluminous data set. This algorithm has manifold applications in the field of health care, credit scoring system, spam filtering and many other real time classifications. The main objective is to perform prediction using the posterior probability and likelihood probability using the Bayes theorem.

The general functionality of Naïve Bayes algorithms performs some tasks to give the output. It starts with conversion of the given data set and its given features into frequency table, followed by that the likelihood table is generated by computing the probabilities of a given features. Finally the Bayes theorem is applied on it to generate the posterior probability. For doing these activities some more tasks are done on the given data set and it starts with pre-processing the data, followed by fitting the algorithm to the training set and finally prediction of the test result is performed. However the accuracy of the output is checked by using the confusion matrix. Though, Naïve Bayes algorithm can be used for both binary class and multi class classification but it generates better results with multi class classification data. Moreover it is very good for text classification data. Here one limitation is there as this technique considers mainly independent features therefore it cannot develop relationships among the features.

The Naïve Bayes works using certain assumptions that all features are equally important and no feature is related to the other, there should not be missing data problem, and continuous features have normal distribution whereas discrete features exhibit multinomial distribution. Now some of the advantages of the Naïve Bayes algorithm are that it is easy to implement in the training data set and it is efficient computationally also. It is effective with huge number of features and it comfortably works with very limited training data. Therefore it can be considered as probabilistic classifier, as it learns the probability of each and every object and its features, and it mainly employs Bayes theorem for that.

D. Linear Regression Algorithm

The main technique used in linear regression ML algorithm is to identify patterns and relationships within data. This model understand the relation between the outcome that we want to predict and the predictors the one we want to make that prediction. It helps to find out the dependency of the changes in the predictors affect the outcome, and making informed decision or predictions by modelling the relationship between a dependent variable and one or more independent variable as a straight line. The corresponding straight line is used for prediction of the value of the dependent variable based on the values of the independent variables. It is specifically designed to predict outcomes for continuous variables. Linear regression is a supervised learning technique used to predict numerical values. It works by finding a straight-line relationship between an input variable and an output variable. The main aim is to determine the best values for the different features so that the difference between the predicted outcomes and the actual results is as small as possible. When you plot the data points on a graph, linear regression tries to draw a line that is as close as possible to all the points. This line shows the general trend of the data, indicating how the outcome changes when the predictor changes. The goal is to make the distance between the data points and the line as small as possible as shown in figure 5 below. In figure 5 the stars represents the individual data points, showing the relationship between predictor (height) and outcome (weight). The red line is the regression line which best fits the data points. It shows the general trend to show changes with predictor. The main target of linear regression is to position the red line in such way that it is as close as possible to all stars, minimizing the distance between the points and the line. This allows us to make the prediction about outcome based on the predictor.

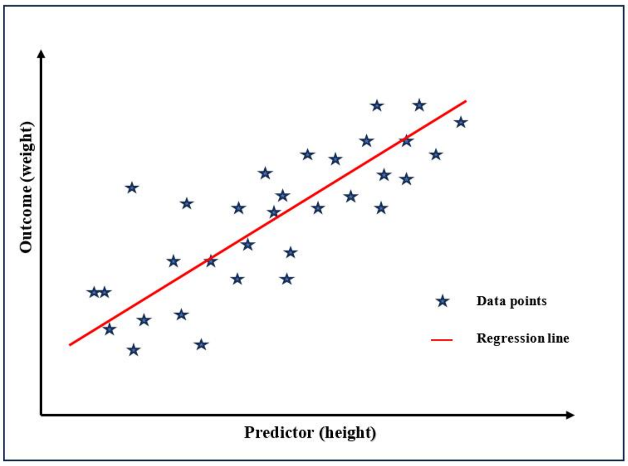


Fig. 5: Representation of Linear regression

Several difficulties may arise when using this algorithm such as the data not representing a linear model, or the presence of noise or outliers. However, this algorithm is quite easy to use. There are two types of Linear Regression algorithm can be classified into two type namely simple linear regression and multiple linear regression. If one use one independent variable to forecast the value of numerical dependent variable, then this type of regression is referred to as simple linear regression. The multiple linear regression is a method that is used to predict the value of a numerical dependent variable by using two or more independent variable.

E. Decision Tree Algorithm

A Decision tree is used supervised machine learning algorithm which can be used for both classification and mainly for regression task. For regression task, it predicts the value linked with the leaf node in continuous target value. For classification, it uses multiple decision trees to improve classification and prevent overfitting in categorical target value. It can automatically select relevant feature and handle the missing values which reduced the need for feature engineering and pre-processing either by not considering missing values or treating them differently from other categories.

In the figure 6 some samples are shown in a distributed 2D plane. A Decision Tree is a tree-like structure of decisions and the possible outcome of decisions. A decision tree consists nodes, branches, and leaves as shown in figure 7 where root and each node is labelled with questions. The branch of each node represents possible outcomes of the questions. Each leaf node represents answer to the problem. A test function is implemented at each decision node at same level. One decision node is taken based on the outcomes of decision node and process will be repeated until a leaf node is hit, which will be the output of a given problem. It uses ‘divide and conquer’ method to divide the problem into tree-like structure where each leaf nodes hold a class label. Figure 6, show a 2D plot where there are 12 shapes, where there are 7 spheres from which 2 are blue sphere and 5 are green spheres and 5 rectangles. In fig. 7 which is indicated decision tree, in a root node at 0th level, where the number of spheres is greater than number of rectangles. At 1st level, two possibilities, with one decision node and one output node. In figure 7, oval nodes are the decision node and quadrilateral nodes are the leaf and output nodes. According to fig. 7, at root node, number of spheres are compared with the number of rectangles, number of spheres is greater than the number of rectangles so the, so the decision node that is next node, will be left node. Again there will be comparison between blue sphere and green sphere, the number of green spheres is greater than the number of blue spheres. Since green sphere is maximum in number, the next decision node will be again left node which is also a leaf node, so the final output is the leaf node which is green coloured circle.

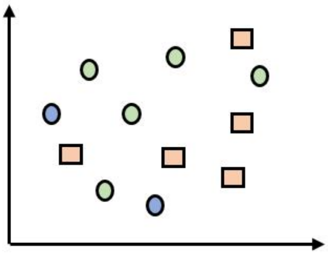


Fig. 6 Data Representation in 2D plane

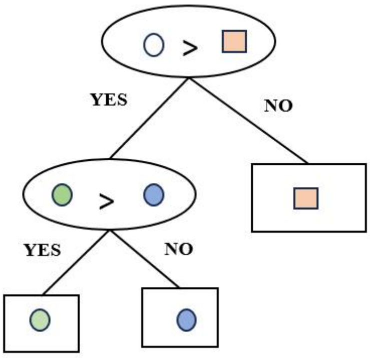


Fig. 7 Decision Tree

F. Neural Network

Neural Network consists of interconnected nodes or neurons and edge that join them together and which is designed to processes and learns from data. It is inspired by the structure and functioning of human brain. Each neuron applies a transformation to its input which involves a combination of weights and an activation function and then passes the result to the neurons in the subsequent layer. This architecture allows neural network to be applied in various task like classification, regression, clustering, and reinforcement learning. Neural can easily adapt the new situations and are useful when the connection between inputs and outputs are not well defined. Neural network consists of multiple layers- Input layer, one or more Hidden layer, and Output layer. All the node in one layer is usually connected to every other node in the next layer. The input layer represents the features of data and is responsible to receive the raw input data. In the hidden layer, computations are performed, input are calculated by multiplying them by weights, adding then up, and passing them through activation function. The output layer provides the final prediction or outcome of the neural network by calculating the results in hidden layers.

The process of calculating result in the hidden layers continues iteratively until the output layer is reached. The figure 8 demonstrate the work-flow of Neural Network. For classification task, the output is calculated on categorical value and might represent probabilities across different classes. For regression task, the output is calculated based on continuous value and it might produce a continuous value. In clustering task, the goal is to group similar data points together without predefined label. Neural network learns to capture the underlying structure of the data and can identify and group similar data points based on the learned features. For Reinforcement learning, neural network can be used to train agents to make optimal decision through interactions to learn policies and value functions that maximize cumulative rewards.

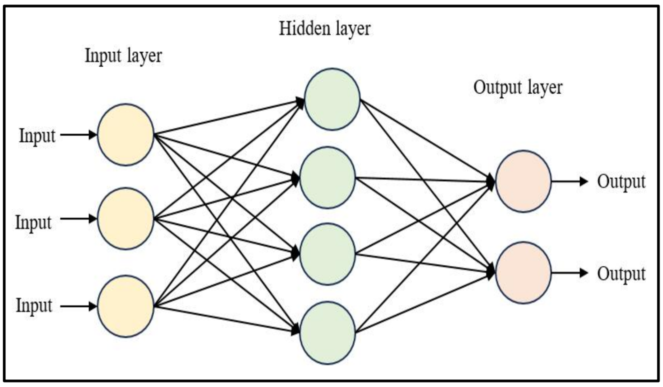
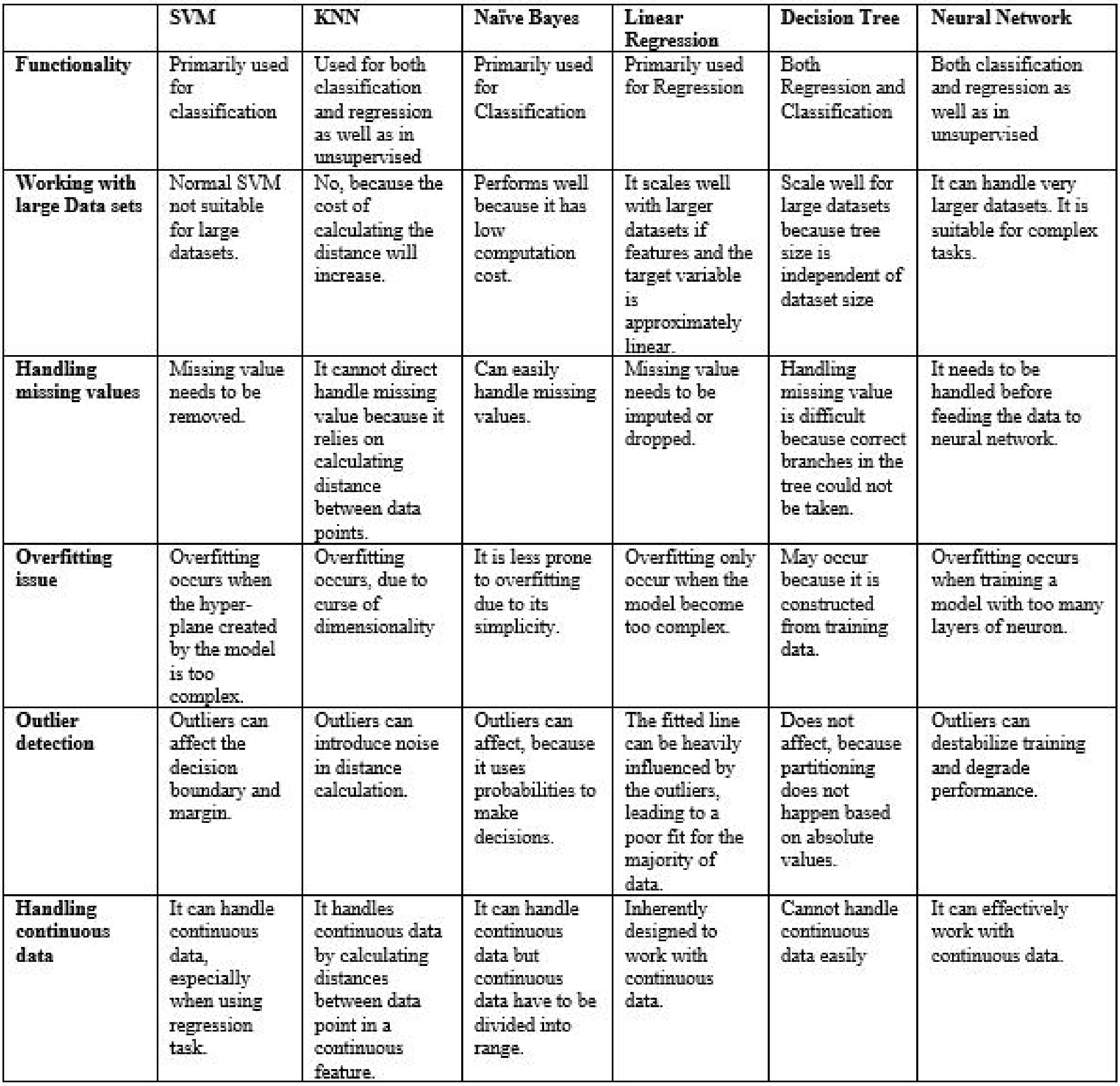


Fig. 8: Neural Network

IV. COMPARATIVE STUDY

In this section of the article a comparative analysis of the afore-mentioned algorithms is presented. The theoretical analysis is mentioned textually and the tabular representation is given in the following Table I. The study includes SVM algorithm, KNN algorithm, Naïve Bayes algorithm, Linear Regression algorithm, Decision Tree algorithm and Neural Network algorithm. There were several parameters for the comparative analysis for this research work and they are namely, voluminous data set issues, missing value problems, overfitting problems, outlier detection problem and handling continuous data. The parameters are defined and discussed hereafter. Nowadays, data is growing very fast nowadays and as a result the data sets are growing by multiplying themselves. Now, to handle such volume of data and generate simplified decisions it needs good amount of knowledge processing. The ML algorithms used for training computers should be efficient enough to handle complex and voluminous data and elicit simple model out of it. There are situations when some of the variables may miss some values and as a result missing value problems arise and samples with missing value can affect the computation.

TABLE I
COMPARATIVE STUDY OF DIFFERENT MACHINE LEARNING ALGORITHMS



	SVM	KNN	Naïve Bayes	Linear Regression	Decision Tree	Neural Network
Functionality	Primarily used for classification	Used for both classification and regression as well as in unsupervised	Primarily used for Classification	Primarily used for Regression	Both Regression and Classification	Both classification and regression as well as in unsupervised
Working with large Data sets	Normal SVM not suitable for large datasets.	No, because the cost of calculating the distance will increase.	Performs well because it has low computation cost.	It scales well with larger datasets if features and the target variable is approximately linear.	Scale well for large datasets because tree size is independent of dataset size	It can handle very larger datasets. It is suitable for complex tasks.
Handling missing values	Missing value needs to be removed.	It cannot direct handle missing value because it relies on calculating distance between data points.	Can easily handle missing values.	Missing value needs to be imputed or dropped.	Handling missing value is difficult because correct branches in the tree could not be taken.	It needs to be handled before feeding the data to neural network.
Overfitting issue	Overfitting occurs when the hyper-plane created by the model is too complex.	Overfitting occurs, due to curse of dimensionality	It is less prone to overfitting due to its simplicity.	Overfitting only occur when the model become too complex.	May occur because it is constructed from training data.	Overfitting occurs when training a model with too many layers of neuron.
Outlier detection	Outliers can affect the decision boundary and margin.	Outliers can introduce noise in distance calculation.	Outliers can affect, because it uses probabilities to make decisions.	The fitted line can be heavily influenced by the outliers, leading to a poor fit for the majority of data.	Does not affect, because partitioning does not happen based on absolute values.	Outliers can destabilize training and degrade performance.
Handling continuous data	It can handle continuous data, especially when using regression task.	It handles continuous data by calculating distances between data point in a continuous feature.	It can handle continuous data but continuous data have to be divided into range.	Inherently designed to work with continuous data.	Cannot handle continuous data easily	It can effectively work with continuous data.

Therefore one possible solution is to discard them and get rid of, but sometimes it is not affordable to do so because of shortage of data in the data set. So a possible solution is imputation, where the missing data can be filled up by estimation process. Imputation can be done by averaging for numeric values, or by using most likely value for discrete data or prediction can also be done depending on the type of data.

The problem of overfitting can develop inefficient ML models which can only give good predictions for the training data only and it will start giving wrong decisions when used upon new set of data. It means the algorithm is just fit enough for the training data only. Hence the overfitting can lead to very poor performance of the model. An outlier sample in a data set may exhibit anomalous behaviour in the entire data. The detection of outliers may help in finding out exceptions in the sample and that can help to detect abnormal behaviour in the system. Outliers can happen due to some kind of fault in the system and it is necessary to detect the same and discard from the sample. In any type of class there are typical instance and abnormal instance so outlier detection means finding out the anomalous or abnormal instances and get rid of them, because any typical instance is fit for ML models. When we mention about the quantitative data it can be classified into discrete data and continuous data. And when it comes to handling of continuous data like speech, where it is difficult to segment the data into discrete observations there the performance of the ML model is also very crucial.

The comparative study given in Table I presents that the SVM, Naïve Bayes and Linear Regression algorithms are mainly used for classification problems, whereas the others can work with both classification and regression problems. All the algorithms can work fine with the large data sets, but in exception to that KNN algorithm is suitable for smaller data set because the computation cost increases with large data sets and SVM algorithm is also good for normal data set. The missing value problem is very well handled by Naïve Bayes algorithm, whereas KNN cannot handle the missing value issue directly but SVM requires all the missing values to be removed during the pre-processing of data. Missing value should be imputed properly or dropped while using Linear Regression but should be handled before feeding the data in Neural Network algorithms. Overfitting problem does not happens with the Naïve Bayes algorithm due to its simplicity otherwise more or less some kind of overfitting happens with all the other algorithms depending upon the complexity of the model. Decision Tree algorithm does not get much affected with the outliers, but the decision making gets affected, influenced or destabilized by the outliers with the use of all others algorithms. The continuous data can be handled by most of the machine learning algorithms discussed here but only Decision Tree cannot handle the same.

However the entire comparative analysis helps to generate an insight about the machine learning algorithms mentioned here.

V. CONCLUSION

The key tasks in different phases of ML lifecycle include specifying the objective, acquisition of dataset, pre-processing of dataset, exploration and analysis of dataset, selection of the most suitable algorithm, using the algorithm to develop a predictive model and finally, the deployment of model. In this paper, the importance of choosing the right algorithm was highlighted, along with the description of some popular supervised ML algorithms and finally these algorithms were compared on the basis of several parameters. The scope of this work can be extended to other ML algorithms. It is possible to have a similar comparison and analysis of other ML learning algorithms that were not listed here. In the present era, new discoveries and breakthroughs are taking place in this field. The issue of model interpretability is arising as a major concern which needs to be addressed through research so that the models that are developed have an assurance that they are safe to use.

REFERENCES

1. *A. Chahal and P. Gulia 2019. Machine Learning and Deep Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 8 (12): 4910-4914.*
2. *A. K. Triantafyllidis and A. Tsanas "Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature" Journal of Medical Internet Research Vol. 21 Issue-4, 2019, doi: 10.2196/12286*
3. *E. Alpaydin "Introduction to Machine Learning" third edition. SBN 978-0-262-02818-9 P (213-216)*
4. *I. H. Sarker Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x*
5. *J. Han| M. Kamber and J. Pei "Data Mining. Concepts and Techniques" The Morgan Kaufmann Series in Data Management Systems, 3rd-Edition, Morgan Kaufmann-2011*
6. *K. Kouroua, T.P. Exarchosa, K.P. Karamouzisc, M.V. and D. I. Fotiadisa (2015) Machine Learning Applications in Cancer Prognosis and Prediction, Computational and Structural Biotechnology Journal, Vol. 13, pp.8–17.*

7. *M. H. Dunham Data Mining Introductory and Advanced Topics, Pearson Education India, 2006*

8. *N. Sangari and Y. Qu "A Comparative Study on Machine Learning Algorithms for Predicting Breast Cancer Prognosis in Improving Clinical Trials" Proceedings of International Conference on Computational Science and Computational Intelligence (CSCI) 2020, pages 813-818, DOI 10.1109/CSCI51800.2020.00152*

9. *N. Sharma, R. Sharma, and N. Jindal, "Machine Learning and Deep Learning Applications-A Vision", June 2021, https://doi.org/10.1016/j.gltp.2021.01.004*

10. *O. M. López, A. M. López and J. Crossa. (2022). Support Vector Machines and Support Vector Regression. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction. P (337-378). https://doi.org/10.1007/978-3-030-89010-0_9*

11. *P. S. Prathyusha, S. Razia, N. V. Krishna, N. SathyaSumana." A Comparative study of machine learning algorithms on Thyroid Disease Prediction," International Journal of Engineering & Technology, 7 (2.8) 2018 pp 315-319.*

12. *R. M. Savithramma, R. Sunathi and H. S. Sudhira "A Comparative Analysis of Machine Learning Algorithms in Design Process of Adaptive Traffic Signal Control System" Journal of Physics: Conference Series 2022, IOP Publishing, doi:10.1088/1742-6596/2161/1/012054*

13. *R. Pugliese, S. Regondi, and R. Marini, Machine learning-based approach: global trends, research directions, and regulatory standpoints, 2021, Pages 19-29, ISSN 2666-7649, https://doi.org/10.1016/j.dsm.2021.12.002.*

14. *S. Anjum, S. Oraon, Saloni, S. Kumari, and S. Mitra "The Life Cycle of a Machine Learning Model" Journal of Emerging Technologies and Innovative Research (JETIR) Vol. 11 Issue 7, 2024, ISSN-2349-5162 pp. 232-238*

15. *T. V. S. Sriram, M. V. Rao, G. V. S. Narayana and DSVGK Kaladhar "Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms" International Journal of Engineering and Innovative Technology (IJEIT) Vol. 3, Issue 3, 2013, pp. 212-215, ISSN: 2277-3754*

16. *Vaishnavi, S. Mitra, and R. Jha "A Comparative Study on Machine Learning Approaches for Diagnosis of Thyroid Disease" AIP Conf. Proc. 3164, 020003 (2024), ICMIA 2023: 1-5*