IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 12, Issue 1, January. 2022, || Series -I || PP 28-32

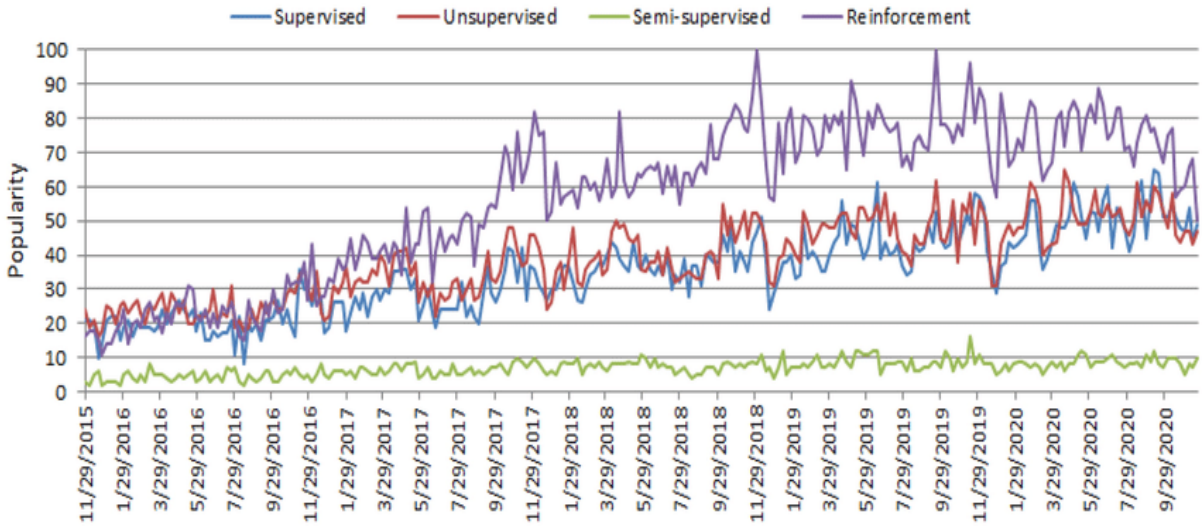
Real World Applications and Research Directions for Machine Learning: Challenges and Defies

Gunjan Divakar[1],Naina Tyagi[2], Praful Saxena[3], Vikas Verma[4]
[1, 2]*(BCA Student,CCSIT, iNurture_Teerthankar Mahaveer University ,U.P.,India)*
[3, 4]*(Assistant Professor, CCSIT, iNurture_Teerthankar Mahaveer University ,U.P.,India)*
Received 11 January 2022; Accepted 27 January 2022

Abstract: The digital world offers a lot of data in this era of the Fourth Industrial Revolution. Intelligently analysing these data and implementing suitable smart and automated applications at scale. Machine learning refers to a set of algorithms, scientific investigations, and statistical models that allow computers to accomplish tasks without having to be explicitly programmed. Many of the applications we use on a daily basis involve learning algorithms. The main advantages of using machine learning are; once an algorithm learns what to do with data, it can do its work automatically. In this paper, we present a comprehensive view on machine learning algorithms that can applied to Enhance the intelligence, the techniques and their applicability thus, this study's key contribution is explaining the principles of different machine learning techniques application domains such as cyber security systems, smart cities, health-care, e-commerce, agriculture and many more as for decision makers in various real world situation and application area from technical point of view.
Key Word: **ML. AI, Big Data**

I. INTRODUCTION

We live in the age of data, where everything around us is connected to data. For instance, the current electronic world has a wealth of various kinds of data[1,2]. Various types of machine learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning exist in the area. To built a data driven automated and intelligent cyber security system, the relevant cyber security data can be used. Artificial intelligence(AI) , particularly, machine learning(ML) have grown rapidly in recent years in the context of data analysis and computing that typically allows the applications to function in an intelligent manner. "Industry 4.0" is typically the ongoing automation of conventional manufacturing using new smart technologies such as machine learning automation. The popularity of these approaches to learning is increasing day-by-day, which is shown in fig.1



The popularity indication values for these learning types are low in 2015 and are increasing day by day . These statistics motivate us to study on machine learning in this paper which can play an important role in the real-world through industry 4.0 automation[3,4]. In general, machine learning solution depend on the nature and characteristics of data and the performance of the learning algorithms, classification analysis, regression, data clustering, feature engineering and dimensionality reduction, association rule learning or reinforcement learning

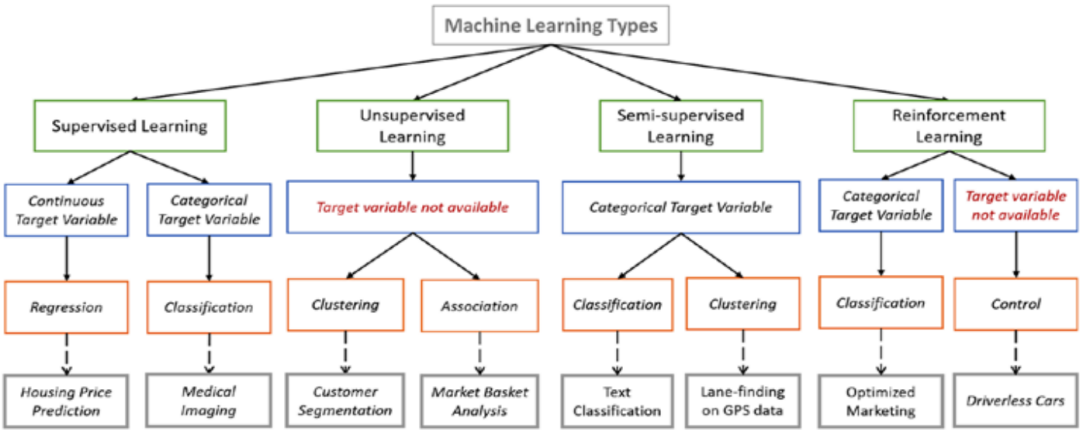
techniques exist to effectively build data driven systems. The purpose of this paper is , therefore, to provide a basic guide for those academia and industry people who want to study research, and develop data driven automated and intelligent systems in the relevant areas based on machine learning techniques.
We briefly discussed and explain different machine learning algorithms in the subsequent section followed by which various real-world application areas based on machine learning algorithms are discussed and summarized.
We highlight several research issues and potential future directions, and the final section concludes this paper.

II. TYPES OF REAL WORLD DATA

The availability of data is considered as, construct a machine learning model or data driven real world systems. Data can be various forms such as Structured, semi-structured or unstructured. In the following, we briefly discuss these types of data.
Structured: conforms to a data model following a standard order, which is highly organized and easily accessed, and used by an entity or a computer program. In well-defined schemes such as relational database, structured data are typically stored i.e. in tabular format.
Unstructured: There is no pre-defined format or organization for unstructured data , making it much more difficult to capture, process and analyze, mostly containing text and multimedia material. Types of business documents can be considered as unstructured data.
Semi-Structured: It does have certain organizational properties that make it easier to analyze. HTML, XML, JSON, No SQL database etc, are some examples of semi –structured data.
Metadata: "data about data" that data are simply the material that can classify, measure, or even document something relative to an organization's data properties. Its more significance for data users.
In the area of machine learning and data science, researchers use various widely used datasets. Different types of machine learning techniques can be discussed in the following.

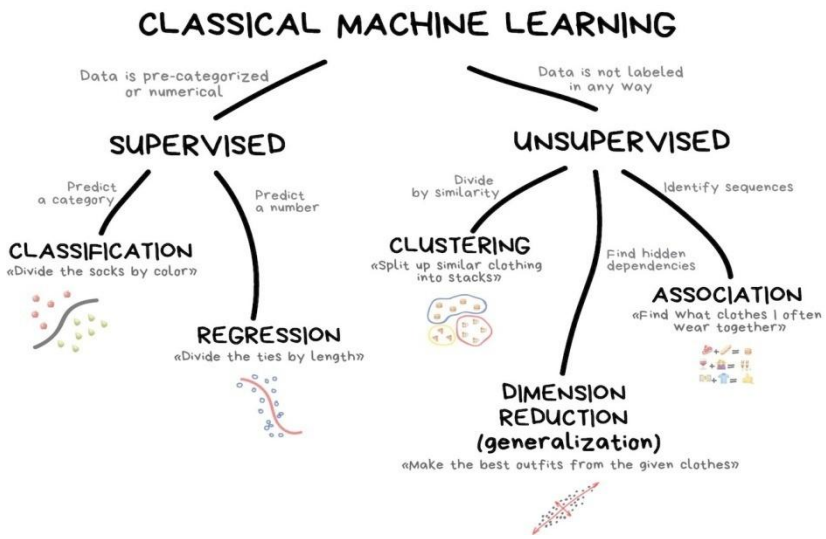
III. TYPES OF MACHINE LEARNING TECHNIQUES

Mainly divided into four categories; Supervised Learning, unsupervised learning, Smi-supervised learning, and Reinforcement learning as shown in fig.2



Supervised : Supervised learning is typically the task of machine learning to learn a function that maps an input to an output based on sample input-output pairs. The most common supervised task are "classification" that separates the data and "regression" that fits the data , predicting the class label or sentiment of a piece of text, like a tweet or a product review i.e. text classification
Unsupervised: Unsupervised learning analyzes unlabeled datasets without the need for human interference. i.e. data-driven process and widely used for identifying meaningful trends and structures, groupings in results, and exploratory purpose. The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionally reduction, finding association rules anomlay detection, etc
Semi-Supervised: Semi-supervised learning can be defined as a hybridization of the above-mentioned supervised and unsupervised methods, as it operates on both labeled and unlabeled data. In real world, labeled data are could be rare in several contexts and unlabeled data are numerous, where semi-supervised learning is useful.
Reinforcement: Enables software agents and machines to automatically evaluate the optimal behavior in a particular context or environment to improve its efficiency. It is powerful tool for training AI modela that can

help increase automation or optimize the operational efficiency such as robotics, autonomous driving tasks, manufacturing and supply chain logistics.



IV. MACHINE LEARNING ALGORITHMS AND THEIR TASKS

In this section we discuss various machine learning algorithms that include classification analysis, regression analysis, data clustering, association rule learning, feature engineering for dimensionality reduction, as well as deep learning methods.

Classification Analysis: Regarded as a supervised method in machine learning, referring to a problem of predictive modeling as well, where a class label is predicted. To predict the class of given data points, it can be carried out on structured or unstructured data. For example; spam detection such as "spam" and "not spam" in email services providers can be classification problem.

Many classification algorithms have been proposed in the machine learning and data science literature. In the following, we summarized the most common and popular methods that are used widely in various application areas.

Naïve Bayes(NB): The naïve bayes algorithm is based on the bayes's theorem with the assumption of independence between each pair of features. It works well and can be used for both binary and multi-class categories in many real world situations, such as document or text classification, spam filtering, etc. The key benefit is that, compare to more sophisticated approaches, it needs a small amount of training data to estimate the necessary parameters and quickly.

Linear Discriminant Analysis(LDA): Linear discriminant analysis is a linear decision boundary classifier created by fitting class conditional densities to data and applying Bayes's rule. The standard LDA model usually suits each classes with a Gaussian density, assuming that all classes with a Gaussian density, assuming that all classes share the same covariance matrix. LDA is closely related to ANOVA and regression analysis.

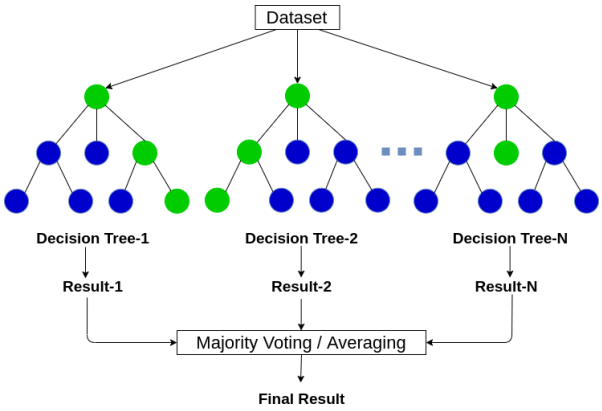
Logistic regression (LR): Logistics regression typically uses a logistic function to estimate the probabilities, which is also referred to as the mathematically defined sigmoid function. It can over fit high dimensional datasets and work well when the datasets can be separated linearly. The assumption of linearity between the dependent and independent variables is considered as a major drawback of logistic regression.

K-nearest neighbors: "instance-based learning" or non-generalizing learning, also know as a "lazy learning" algorithm. It does not focus on constructing a general internal model; instead , it stores all instances corresponding to training data in n-dimensional space.KNN uses data and classifies new data points based on similarity measures. It is quite robust to noisy training data, and accuracy depends on the data quality. The biggest issue with KNN is to choose optimal number of neighbors to be considered. KNN can be used both for classification as well as regression.

Support Vector Machine(SVM): Another common techniques that can be used for classification, regression, or another tasks is a support vector machine. A support vector machine construct a hyper-plane or set of hyper-planes, the hyper-plane , which has the greatest distance from the nearest training data points in any class. It is effective in high-dimensional spaces and can behave differently based on different mathematical functions known as the kernel. Linear, polynomial, radial basis function , sigmoid, etc. are the popular kernel functions used in SVM classifier however, when the data set contain more noise such as overlapping target classes, SVM does not perform well.

Decision Tree(DT): Is a well known non-parametric supervised learning method, DT learning methods are used for both classification and regression tasks.

Random Forest(RF): A random forest classifier is well known as an ensemble classification techniques that is used in the field of machine learning and data science in various application areas. This method uses "parallel ensembling" which fits several decision tree classifiers in parallel as shown in fig.3



Adoptive Boosting(Adaboost): Is an ensemble learning process that employs an iterative approach to improve poor classifiers by learning from their errors. And also known as "meta-learning". AdaBoost uses "sequential ensembling" . It creates a powerful classifier by combining many poorly performing classifiers to obtain a good classifier of high accuracy. Adaboost is best used to boost the performance of decision trees, base estimator , on binary classification problems.

Extreme gradient boosting(XGBoost): Gradient boosting like random forests above, is a ensemble learning algorithm that generates a final model based on a series of individual models, typically decision trees. The gradient is used to minimize the loss function, similar to how neural network use gradient decent. XGBoost is fast to interpret and can handle large-sized datasets well.

V. APPLICATIONS OF MACHINE LEARNING

In the current age of the fourth industrial Revolution, machine learning becomes popular in various application areas, because of its learning capabilities from the past and making intelligent decisions .In the following we have ten popular application areas of machine learning technology are; predictive analytics and intelligent decision-making, cyber security and threat intelligence, Internet of things(IOT) and smart cities, Traffic prediction and the transportation, Healthcare and COVID-19 pandemic, E-commerce and product recommendations, NLP and sentiment analysis, Image speech and pattern recognition, Sustainable agriculture and User behavior analytics and context-aware smart phone.

VI. CHALLENGES AND RESEARCH DIRECTIONS

Our study on machine learning algorithms for intelligent data analysis and applications opens several research issues in the area. Thus , in this section, we summarize the challenges faced and the potential research opportunities and future directions.

The effectiveness and the efficiency of a machine learning- based solution depend on the nature and characteristics of the data, and the performance of the learning algorithms To collect the data in the relevant domain, such as cyber security, IOT , healthcare, agriculture and etc. Thus, collecting useful data for the target machine learning-based applications. The in-depth investigation of data collection methods is needed while working on the real-world data. Moreover, the historical data may contain many ambiguous values, missing values, outliers, and meaningless data.

To analyze the data and extract insights, there exist many machine learning algorithms, the ultimate success of a machine learning-based solution and corresponding applications mainly depends on both the data and the learning algorithms. If the data are bad to learn, such as non-representative, poor-quality, irrelevant features, or insufficient quantity for training, then produce lower accuracy. Therefore, effectively processing the data and handling the diverse learning algorithms are important, for a machine learning- based solution and eventually building intelligent application.

VII. CONCLUSION

In this paper, we have conducted a comprehensive overview of machine learning algorithms for intelligent data analysis and applications. According to our goal, we have briefly discussed how various types of machine learning methods can be used for making solution to various real world issues. A successful machine learning model depends on both the data and the performance of the learning algorithms. The sophisticated learning algorithms then need to be trained through the collected real-world data and knowledge related to the target application before the system can assist with intelligent decision making. Finally, we have summarized and discussed the challenges faced and the potential research opportunities and future direction in the area. Therefore, the challenges that are identifies create promising research opportunities in the field which must be addressed with effective solutions in various application areas. Overall, we believe that our study on machine learning-based solution opens up a promising direction and can be used as reference guide for potential research and applications for both academia and industry professionals as well as for decision – makers from technical point of view.

REFERENCES

[1]. Cao L. Data science: a comprehensive overview. ACM Comput Surv (CSUR). 2017;50(3):43.
[2]. Sarker IH, Hoque MM, MdK Uddin, Tawfeeq A. Mobile data science and intelligent apps: concepts, ai-based modeling and research directions. Mob Netw Appl, pages 1–19, 2020.
[3]. Han J, Pei J, Kamber M. Data mining: concepts and techniques,Amsterdam: Elsevier; 2011.
[4]. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2005.

Gunjan Divakar, et. al. "Real World Applications and Research Directions for Machine Learning: Challenges and Defies." *IOSR Journal of Engineering (IOSRJEN)*, 12(01), 2022, pp. 28-32.