CONSERVATIVE SURVEY OF MACHINE LEARNING AND DATA ANALYTICS

Oluyinka I Omotosho[1*], Kehinde D. Adebisi[2], Deborah O. Fenwa[3]

[1,3]Department of Cyber Security Science, Ladoke Akintola University of Technology, Ogbomosho - Nigeria*
[2]Department of Computer Science, Ladoke Akintola University of Technology, Ogbomosho - Nigeria

Corresponding Author:
oiomotosho@lautech.edu.ng

Abstract

Computing power has been increasing exponentially, meaning that processing power can be harnessed to solve more complex tasks. Two fields that have emerged alongside this rapid growth are data analytics, machine learning. Data analytic and Machine learning algorithms are two terms used interchangeably in the world of big data and statistical analysis, the line that divide these two related terms is so tiny that most data analyst forget about the existence of such line dividing and providing blur differences between data analytic and machine learning algorithms. In this study, the underlying the difference between these two related terms and their common and different applications is studied in a concise manner. Their impact in decision making for firms, organizations and cooperate bodies, their approaches to problem solutions, as well as their limitations.

Keywords: *Data Analytics, Machine Learning, Big Data*.

I INTRODUCTION

The world today is data driven, organizing data into information has contributed extensively to the development of the world and bring evolution to science and technology, various field of studies including Finance, Art, Political, Medicine, Research and Computing use data largely for proper development and future enhancement of these field of studies. The world exists in the stream of data, where data is the cheapest thing around us, Human will continue to depend on data for decision making, solving problems and interacting with our environment.

The use of data is endless and it calls for a need to understand how data will be organize, categorize, structure and analyse in an efficient way that continue to contribute to the development of societies, organizations, firms and individuals. Several Texts, articles etc. exist on machine learning, however, this review work is to dissect these techniques in a straightforward manner in order expedite decision on the specific applicability and associated tools.

Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. It is a broad term that encompasses many diverse types of data analysis. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system. For example, manufacturing companies often record the runtime, downtime, and work queue for various machines and then analyze the data to better plan the workloads, so the machines operate closer to peak capacity. Data analytics can do much more than point out bottlenecks in production. Gaming companies use data analytics to set reward schedules for players that keep the majority of players active in the game. Content companies use many of the same data analytics to keep you clicking, watching, or re-organizing content to get another view or another click.

The term Machine Learning often confused with Artificial Intelligence, but it is the sub-domain of AI. Machine learning is a predictive modelling that enables computer to handle data and perform analysis. With huge amount of data stored it's impossible to extract data from the datasets and interpret the pattern, in that case, we use machine learning algorithms to find statistical regularities and other patterns from the data. Nowadays machine learning is in very high demand and many industries from medical to the military apply these algorithms to extract the relevant insights from the datasets.

The method of data analysis that automates the process of analytical model building by learning from the data is machine learning. It is a sub-domain of artificial intelligence used to get insights from the past or live data and make decisions via predictive modelling techniques. The primary goal is to train or allow the computer to learn automatically with less human assistance and take the decisions accordingly. A huge amount of data can be analyzed using machine learning algorithms. These algorithms are faster and accurate, requires additional time and resource to get trained properly. We can combine these algorithms with AI to make it more effective in processing a very large amount of data.

This study aims to provide to Identify the line that distinguish data analytic and machine learning algorithms, provide a broad and useful information for researchers, students and readers carrying out research or findings on data analytic and machine learning algorithms.

II DATA ANALYTICS

Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements. Data analytics is the pursuit of extracting meaning from raw data using specialized computer systems. These systems transform, organize, and model the data to draw conclusions and identify patterns. While data analytics can be simple, today the term is most often used to describe the analysis of large volumes of data and/or high-velocity data, which presents unique computational and data-handling challenges. The era of big data drastically changed the requirements for extracting meaning from business data.

Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system. For example, manufacturing companies often record the runtime, downtime, and work queue for various machines and then analyze the data to better plan the workloads so the machines operate closer to peak capacity. Data analytics can do much more than point out bottlenecks in production. Gaming companies use data analytics to set reward schedules for players that keep the majority of players active in the game. Content companies use many of the same data analytics to keep you clicking, watching, or re-organizing content to get another view or another click.

Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data. A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new—and better—products and services.

Data Analysis Procedural Process

1. *Data Requirement Gathering:* This determining the purpose of the analysis, which and type of data to use.
2. *Data Collection:* Guided by the identified requirements, it's time to collect the data from determined sources. Sources include case studies, surveys, interviews, questionnaires, direct observation, and focus groups. It also include organizing the collected data for analysis.
3. *Data Cleaning:* Not all of the data collected will be useful. This process includes removal of white spaces, duplicate records, and basic errors. Data cleaning is mandatory before proceeding to analysis.
4. *Data Analysis:* data analysis software and other tools are used to assist interpret, understand the data and arrive at conclusions. Data analysis tools include Excel, Python, R, Looker, Rapid Miner, Chartio, Metabase, Redash, and Microsoft Power BI.
5. *Data Interpretation:* Once results is generated, one need to interpret them and come up with the best courses of action based on findings.
6. *Data Visualization:* Data visualization is a fancy way of graphically showing information in a way that people can read and understand it. It includes the use charts, graphs, maps, bullet points, or a host of other methods. Visualization helps derive valuable insights by assisting to compare datasets and observe relationships.

Types of Data Analytics

Data analytics is broken down into four basic types.

1. ***Descriptive analytics*:** This describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
2. ***Diagnostic analytics:*** This focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that latest marketing campaign impact sales?
3. ***Predictive analytics:*** This moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
4. ***Prescriptive analytics:*** This suggests a course of action. If the likelihood of a hot summer is measured as an average of these five weather models is above 58%, we should add an evening shift to the brewery and rent an additional tank to increase output.

Data Analytic Techniques

Some analytical methods or techniques that can be can used to process data and extract information by data Analyst are;

- ***Regression analysis:*** entails analyzing the relationship between dependent variables to determine how a change in one may affect the change in another.
- ***Factor analysis:*** entails taking a large data set and shrinking it to a smaller data set. The goal of this maneuver is to attempt to discover hidden trends that would otherwise have been more difficult to see.
- ***Cohort analysis:*** is the process of breaking a data set into groups of similar data, often broken into a customer demographic. This allows data analysts and other users of data analytics to further dive into the numbers relating to a specific subset of data.
- ***Monte Carlo simulations:*** model the probability of different outcomes happening. Often used for risk mitigation and loss prevention, these simulations incorporate multiple values and variables and often have greater forecasting capabilities than other data analytics approaches.
- ***Time series analysis:*** tracks data over time and solidifies the relationship between the value of a data point and the occurrence of the data point. This data analysis technique is usually used to spot cyclical trends or to project financial forecasts.

Data Analytic Tools

In addition to a broad range of mathematical and statistical approaches to crunching numbers, data analytics has rapidly evolved in technological capabilities. Today, data analysts have a broad range of software tools to help acquire data, store information, process data, and report findings.

- Data analytics has always had loose ties to spreadsheets and Microsoft Excel. Now, data analysts also often interact with raw programming languages to transform and manipulate databases. Open-source languages such as Python are often utilized. More specific tools for data analytics like R can be used for statistical analysis or graphical modeling.
- Data analysts also have help when reporting or communicating findings. Both Tableau and Power BI are data visualization and analysis tools to compile information, perform data analytics, and distribute results via dashboards and reports.
- Other tools are also emerging to assist data analysts. SAS is an analytics platform that can assist with data mining, while Apache Spark is an open-source platform useful for processing large sets of data. Data analysts now have a broad range of technological capabilities to further enhance the value they deliver to their company.

Data Analytics versus Data Analysis

The difference between data analysis and data analytics is that data analytics is a broader term of which data analysis forms a subcomponent. Data analysis refers to the process of compiling and analyzing data to support decision making, whereas data analytics also includes the tools and techniques use to do so.

III MACHINE LEARNING ALROGITHMS

Machine learning is a predictive modelling that enables computer to handle data and perform analysis. With huge amount of data stored it is impossible to extract data from the dataset and interpret the pattern, in that case, machine learning algorithms is used to find statistical regularities and other patterns from the data. Nowadays machine learning is in very high demand and many industries from medical to the military apply these algorithms to extract the relevant insights from the datasets.

The method of data analysis that automates the process of analytical model building by learning from the data is machine learning. It is a sub-domain of artificial intelligence used to get insights from the past or live data and make decisions via predictive modelling techniques. Computational learning theory is a branch of statistics that deals in the performance and computational analysis of machine learning algorithms.

The primary goal is to train or allow the computer to learn automatically with less human assistance and take the decisions accordingly. A huge amount of data can be analyzed using machine learning algorithms. These algorithms are faster and accurate, requires additional time and resource to get trained properly. We can combine these algorithms with AI to make it more effective in processing a very large amount of data.

Types of Machine Learning Algorithm.

Machine learning algorithms are often divided into two categories i.e. supervised and unsupervised learning, but there are other algorithms too that can play a major role in machine learning. All are captured as follows:

Supervised Learning

Supervised learning algorithms are those that need external assistance. As the name suggests these algorithms are designed to learn by example, it is like a supervisor or teacher to train the whole process. The set of inputs with pre-defined outputs is provided as training data and the algorithm will search for the patterns in the data. After training, the algorithm is implemented to the new set of data with new inputs that will be classified as depicted in Figure 1.. The objective of the supervised learning algorithm is to predict the outputs for new data.

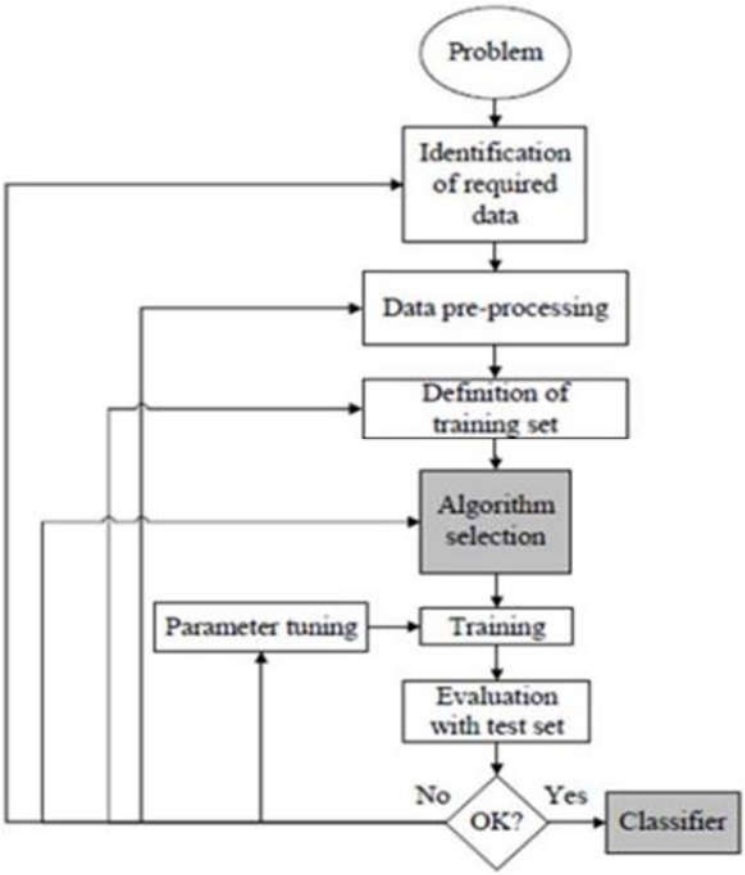

Fig 1: Workflow of Supervised machine learning algorithm

Supervised Learning Algorithms

1. ***Decision Tree,*** *also referred to as a classification tree or a reduction tree:* uses a tree-like model to specify sequences of decisions and consequences. The decision tree uses predictive models to predict the behavior that has not been tested. I.e. if an organization wants to switch from an analog controller to a digital controller, this model can be used to test the performance change. In a decision tree predictive model is used to map the observations about an item to a conclusion

about its output value. The response predictions can be achieved by making a decision tree with testing points and branches as Depicted in Figure 2.

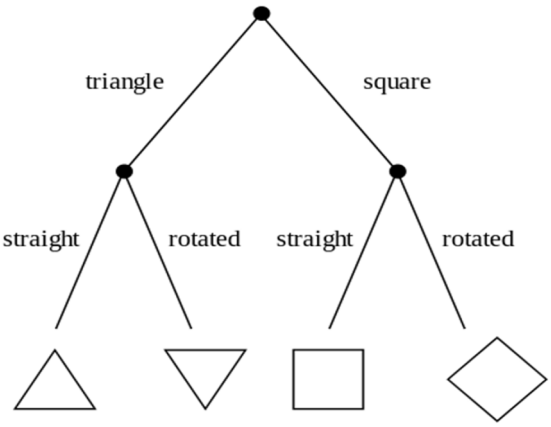

Fig 2: Simple Decision Tree

2. ***Naive Bayes:*** This algorithm is based on Bayes Theorem which follows probabilistic technique and mainly used for clustering and classification purpose. But it works well for natural language processing problems. Naïve Bayes creates trees based on the conditional probability of happening and these trees are known as Bayesian Network.

3. ***Support Vector Machine*****:** It is another widely used machine learning technique and mainly used for both regression and classification tasks. It works on the principle of margin calculation. The main objective of this algorithm is to find an N-dimensional hyperplane. In this technique, we choose hyperplane to separate the two classes of data points and find a plane which has the maximum margin. (i.e. margin is the maximum distance between the data points of both the classes) and these decision boundaries of hyperplane is to classify the data points. Data points which falls on either side of the boundaries are considered as the different classes.

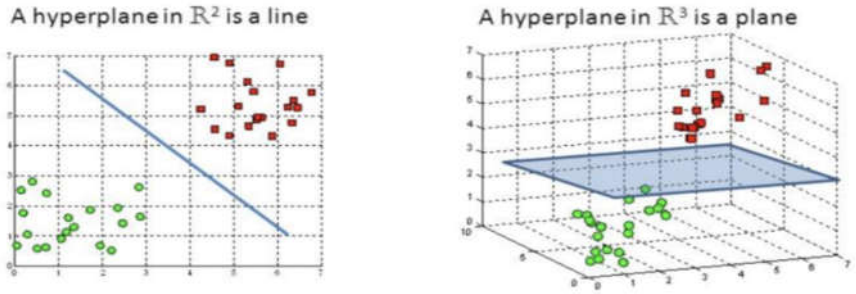

Fig 3. Hyperplane in 2D and 3D Space

Unsupervised Learning

K-Like supervised learning algorithms, unsupervised learning algorithms do not need a supervisor or teacher to train the whole process. It uses unlabelled datasets and works on its own and discovers hidden insights or patterns that were previously not discovered. The more complex task can be performed using unsupervised learning as compared to supervised learning. Unsupervised learning problems are further categorized into clustering and feature reduction problems. Clustering is well known unsupervised learning problem. Means Clustering and Principal Component Analysis are two main algorithms for clustering and feature reduction comes under unsupervised learning.

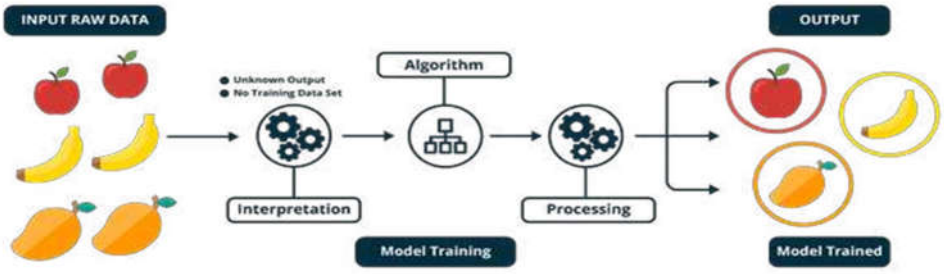

Fig. 4: Example of Unsupervised Learning

1 ***K-Means Clustering:*** A collection of similar data objects is known as clusters and the process of organizing or collecting those data objects into groups is clustering. The objects having same properties grouped together and put in the same

cluster. This algorithm creates K distinct clusters; the centre of the cluster contains the mean of the values in a particular cluster. Figure 5 shows the un-clustered data and clustered data.

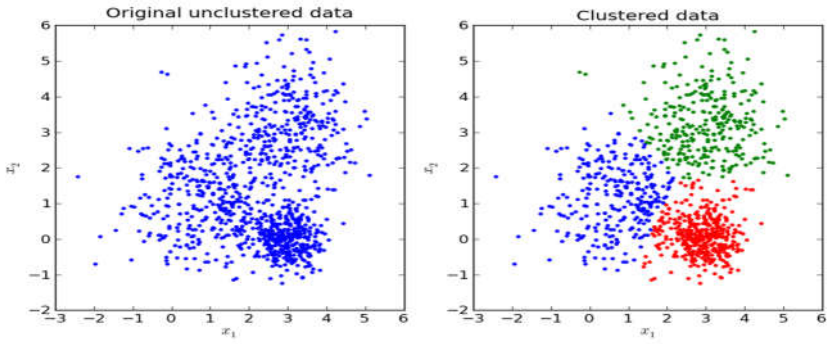

Fig 5. K-Means Clustering.

Semi-Supervised Learning

The semi-supervised learning algorithms are the mixture of supervised and unsupervised learning techniques. This algorithm can be trained using the dataset that contains both labeled and unlabelled data, but mostly unlabelled data is preferred. This technique is used when there are no enough labeled data to produce an accurate result and neither resources to get labeled data.

1. **Generative Models:** Generative models is one of the oldest technique used in semi-supervised learning. It processes a large amount of training data and data reduction is performed digitally and generates new data instances. These models are implemented on neural networks and generate new reduced data instances.

2. **Self-Training:** In self-training models, the classifier is trained using a small amount of labeled data and after training it is provided with the unlabeled data to make the predictions. The classifier starts learning itself when this process is repeated many times, hence known as the self-training technique.

3. **Transductive Support Vector Machine**: It is an extension of SVM and widely used for treating partially labeled data. But labeled and unlabelled data both can be considered. It can be used to label the unlabelled data and the margin between them is maximum.

Neural Network Learning

It is also known as artificial neural network or ANN. The concept of neural network is derived from biology where neurons are a cell-like structure in a brain. To fully understand the working of neural networks we have to know how neurons work in our brain. A neuron consists of mainly 4 parts viz. dendrites, nucleus, axon, and soma.

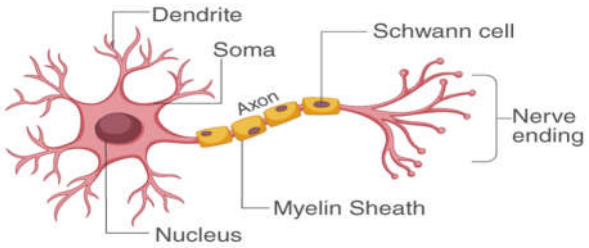

Fig 6. The Structure of Neuron

An electrical signal is received by the dendrites and processed by the soma. Axon carries the output of the processed signal to the dendrites terminals and then it is transferred to the next neuron. The heart of the neuron is a nucleus. The electrical impulse travels around the brain through neurons and this interconnection of neurons is known as a neural network.

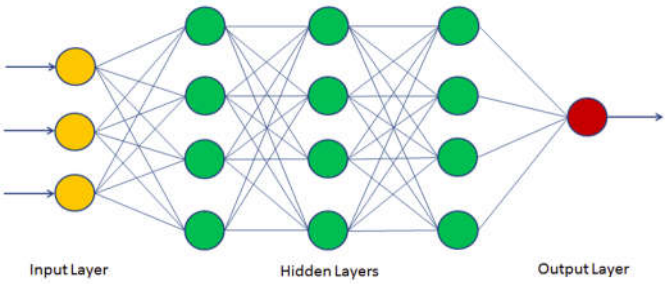

Fig 7: Structure of Neural Network.

An artificial neural network works like our brain. There are mainly three layers viz. input layer, hidden layer, and output layer. The input layer takes input and hidden layer process that input and output layer then sends the processed output.

Categories of Neural Network

The neural network can be divided into three basic categories: supervised, unsupervised, and reinforced.

1. *Supervised Neural Network*: The concept of supervised neural networks is the same as supervised learning algorithms; the output is already known for the set of inputs. The predicted outputs are then compared with the know outputs and the parameters have been changed based on the variations in the outputs. With changed parameters, input is again provided to the neural network to get the desired outputs as depicted in Figure 8.

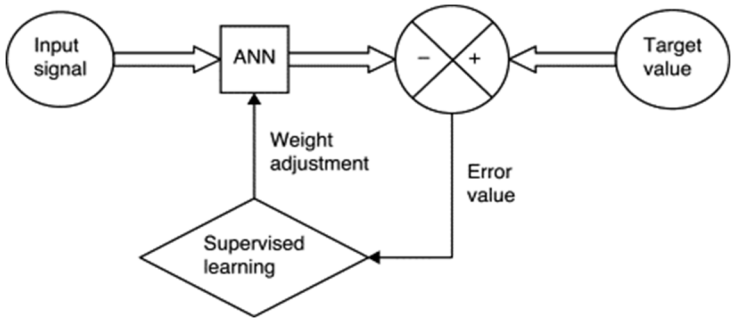

Fig. 8: Supervised Neural Network.

2. *Unsupervised Neural Network*: It is the same as the unsupervised learning algorithm; no idea about the output at the beginning. The network categorizes the data as per the similarities between them. The grouping can be done by the neural network after checking the correlation among the inputs.

3. *Reinforced Neural Network*: It is the same as the reinforcement learning algorithm; agent interacts with the environment as we human do and some reward has been provided on the basis of decision taken by the network. If the decision taken is correct then the connection between the corresponding points is strengthened and weakened if the decision is wrong. In this way, the reinforced neural network works and it has no prior information about the outputs. This functionaly is depicted in Figure 9.

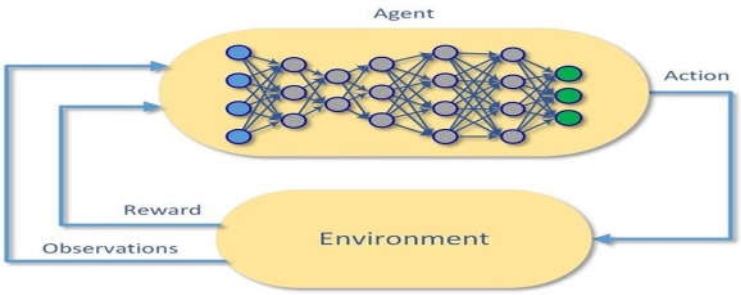

Fig. 9: Reinforced Neural Network

Machine Learning Functionalities

1. Choosing a Model: A machine learning model determines the sought output after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, the model must be suited for numerical or categorical data and choose accordingly.
2. Training the Model: Training is the most important step in machine learning. In training, the prepared data is passed the machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.
3. *Evaluating the Model:* After training the model, one need to check to see how it's performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set earlier splitted. Testing done using same data which is used for training, does not yield accurate measure, as the model is already used to the data, and finds the same patterns in it. This results in disproportionately high accuracy. Rather, when used on testing data, accurate measure of model's performance and speed is guaranteed.
4. *Parameter Tuning:* Once model has been created and evaluated, one need to see if accuracy can be improved in any way. This is done by tuning the parameters present in the model. Parameters are the variables in the model that the programmer generally decides. At a particular value of your parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values
5. *Making Predictions:* In the end, you can use your model on unseen data to make predictions accurately.

APPLICATIONS OF MACHINE LEARNING ALGORITHMS

1. Image Recognition

It is the most common application of machine learning and used to identify persons, objects, places, etc. Traditional algorithms line K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) is used for image classification and recognition.

2. Speech Recognition

While using smartphones we get the option to search by voice option, it comes under speech recognition and very popular machine learning application. It uses a supervised learning algorithm and deep neural network techniques. It can enable the program to process human speech into written text and search accordingly. There is a python library named Speech Recognition that works with the support of APIs provided by Google, IBM, Microsoft, etc.

3. Recommender Systems

Machine learning is widely adopted by various e-commerce, entertainment, and OTT companies like Amazon, Flipkart, Netflix, etc. to provide product recommendations to the users. It is based on deep learning and neural network and analyses the user search and purchase patterns. It uses a collaborative filtering technique to filter out the data.

4. Medical Diagnosis

Machine learning is widely used in the field of medical science for disease diagnoses. One of the best examples is a prediction of potential heart failure, an algorithm is designed to scan and identify the patterns in a patient's cardiovascular history and making the analysis using medical reports, and no need for physician to dig through multiple health records.

DATA ANALYTIC AND MACHINE LEARNING OVERLAP

The goal of data analytics is to examine large quantities of data with the purpose of drawing conclusions about the data. Several techniques can be employed, each using similar methods but having a slightly different focus. The methods include, e.g., statistics, data mining, and machine learning. This overlap is shown in Figure 10. The goal of machine learning algorithms on the other hand is to make a machine (computer) learn from data for pattern recognition by following data analytics processes.

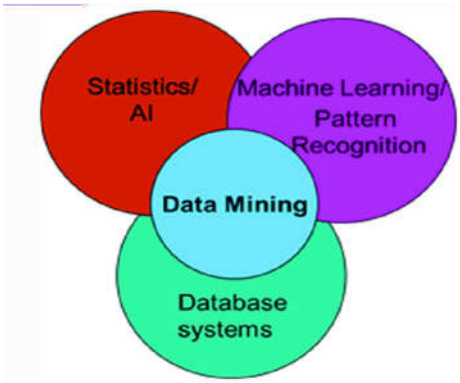

Fig 10: Data Analytic Methods

Sometimes the division between machine learning and data mining is done based on data sets. Data mining is focused on analyzing large databases, whereas in machine learning the focus is on learning patterns from data and the roots of data analysis are in statistics. More significantly, Machine learning and Data Analytics overlaps based on the following factors;

1. *Data-driven*: Each of these areas relies on analyzing huge amounts of data. The more information available, the more effective they are at producing results. It often takes a lot of computer processing power to manage such large data sets.
2. *Insights:* Data analytics, AI, and machine learning can all be used to produce detailed insights in particular areas. By examining data, each can identify patterns, highlight trends, and provide valuable and actionable outcomes.
3. *Predictive models:* These technologies can also help to create forecasts and predictions based on existing data. Again, this process can help organizations of all kinds plan for the future and make informed decisions.

CONCLUSION

In this Paper, the use of data analytics and machine learning algorithms in modern data science community, business organization and technology companies, have been discussed. it is no doubt that these technologies contributes to the development of our environment, it is highly recommended to maximize the efficacy of data analytics and machine learning in effecting strategized decisions that serve as ultimate successive backbone to firms, organizations, Institutions and Individuals.

REFERENCES

[1]. Arpit Kakkar,"Machine Learning Algorithms: Analysis and Application" ISSN: 2395-0072
[2]. Willi Richert et Luis Pedro Coelho "Building Machine Learning Systems with Python"

[3]. Martín Noguerol T, Paulano-Godino F, Martín-Valdivia MT, Menias CO, Luna A (2019)
[4]. Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology. J Am Coll Radiol 16:1239–1247
[5]. Baum J, Laroque C, Oeser B, Skoogh A, Subramaniyan M (2018) Applications of big data analytics and related technologies in maintenance—literature-based research. Machines 6:5
[6]. Doupe P, Faghmous J, Basu S (2019) Machine learning for health services researchers. Value Health 22(7):808–815
[7]. Núñez Reiz A, Armengol de la Hoz MA, Sánchez García M (2019) Big data analysis and machine learning in intensive care units. Big Data Analysis y Machine Learning en medicina intensiva. Med Intensiva 43(7):416-426. https://doi.org/10.1016/j.medin.2018.10.007
[8]. Dash S, Shakyawar SK, Sharma M, Kaushik S (2019) Big data in healthcare: management, analysis and future prospects. J Big Data.
[9]. "https://machinelearningmastery.com/a-tour-of-machine-learning-algoriths/"
[10]. "https://www.expert.ai/blog/machine-learning-definition/"
[11]. "https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590"
[12]. "https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html"