Volume: 10 Issue: VI Month of publication: June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44888

A Review on Machine Learning Tools and Techniques

M Ram Kishore[1], V Bhuvaneshwari[2], Aniket Kumar T[3], P Harish Reddy[4], A V V Sai Pranav[5]
[1, 2, 3, 4, 5] Department of Computer Science and Engineering, GITAM UNIVERSITY, Vizag, Andhra Pradesh , 530045, India.

Abstract: Data is evolving as the fuel for the new economy and the future economy. Having the right data saves time. For major businesses, a significant portion of their work is spent gathering data, sorting it, and then analyzing it in various business contexts to derive insights that are beneficial to the firm. The key challenge or the key opportunity to any organization is to be able to convert the data available into intelligent action. The best way to take advantage of this data is to convert into intelligence, specifically intelligent actions. All these cannot be performed effectively when done manually. Using machine learning, it is possible to build a system that continuously learns from historical data, build models which are fed with input data, train those models, and then predict the future when it receives new data. This whole task of building intelligent systems can be boiled down to building better models that can make precise predictions. There are several tools available that can be utilized to construct better models for different algorithms, which simplifies the process of building models seamlessly and easier to comprehend. Building a machine learning system requires a number of steps, including handling various data types, analyzing and preprocessing them, creating neural networks, training the model, evaluating it, and making iterative changes to the model to improve its performance on both training and test datasets. In this paper we will introduce you to various tools that can be used at different phases of developing a strong machine learning model. Additionally, we'll provide case-studies of a few applications, guide you through the tasks the application performs, and explain which machine learning techniques are employed to meet the application's primary goals.

Keywords: Machine learning, Data Science, Data Analytics, Machine learning tools, Machine learning applications

I. INTRODUCTION

Data is the prerequisite to work with anything and everything. The data that has been processed, organized, structured, or presented in a systematic way makes it usable which gives information regarding a particular aspect. It paves the way to be able to set performance targets, locate benchmarks, and build baselines because of the fact that data allows us to measure. Through the interpretation of vast amounts of data from real time in a methodical approach to make strategic decisions and by analysis and study of hidden patterns in the data, we can easily come up with predictive insights for any specific problem statement. This methodology when inherited into machine learning can help derive the functional form of a model or the parameters of an algorithm which inturn helps the model to learn by itself rather than depending on an external agent to hardcode its functionalities. Making use of this approach in businesses can help companies grow on a large scale by making decisions using evidence-based data so as to figure out the unknown trends regarding upcoming business prospects, better serve customers, drive sales, enhance operations and much more. This could aid businesses grow watchfully and plan carefully to pursue business objectives in the long run. According to a recent poll conducted by the Mckinsey Global Institute, Data-driven businesses are 23 times more likely to attract customers, 19 times more certainly be profitable, and six times more likely to retain customers.

There is certainly not a lack of data available but there is a lack of knowledge and techniques on how to use the data wisely so that utmost information can be extracted out of the data available. BARC research surveyed a variety of businesses and discovered that those using data- driven approaches in their business saw an 8 percent increase in their profit and a 10 percent decrease in costs incurred. Thus , feeding this abundantly available data to the machines to aid them learn by themselves by unveiling the hidden patterns in the data and learn through experiences seems more promising than using a model- driven learning because data is the most valuable asset anyone can make use to create tremendous impact. Alongside feeding data to the machine, it is equally important to know the various methodologies and tools that can be used to handle the data so as to extract the most out of it. Choosing the right tools and being able to implement the data driven strategy wisely is the most crucial step because a misstep can lead to providing completely unsatisfactory results[1].

The field of machine learning deals with the study of how to construct a computer algorithm that can mostly improve automatically based on the experience that gained and the past used data rather than hard coding the algorithm.The term machine learning was first used in 1959, But machine learning started emerging as an area of AI(Artificial Intelligence) in the 1990s because algorithms that adopt the ideas from different fields of mathematics like statistics, probability are more of dynamic than fixed, rule based algorithms, which require human effort. Machine learning borrows ideas from many different disciplines, so it is an interdisciplinary and multidisciplinary field. The majority of fields associated with machine learning are mathematics, computer science, statistics, deep learning, artificial intelligence, data science, data mining, and Natural language processing. The most accepted definition of the term machine learning is :“A computer program or a machine that learns from experience E that gained by doing some task T and has computed performance measure P, if its performance on T, is measured as P, improves with experience E.” this was the definition used by Tom M Mitchell in the book “Machine Learning” . In our current world, we use machine learning in every domain and in every aspect of life, from recommending a toy to a toddler to predicting falls in older adults.ML is not only restricted for the above mentioned tasks it is used for predicting the climate change, Financial Frauds, Speech recognition , Sentiment analysis and many more. Furthermore, machine learning is a multidisciplinary area that has grown through time and continues to evolve. It has obviously grown rapidly since the 1990s, with the discovery of support random forests, vector machines, Long Short Term Memory (LSTMs) networks, and the advancement of machine and deep learning frameworks like scikit-learn, TensorFlow, PyTorch, and Theano[2].

II. OVERVIEW OF MACHINE LEARNING

British mathematician Clive Humby quoted “Data is the new oil”.Nowadays data-driven decisions are ruling the world that only rule-based decisions. Machine learning algorithms are completely data-driven, historical data is mandatory for developing an ML model. Machine learning models formulate and fit a mathematical function for the provided data. So in future, if any new data comes then the model tries to fit that data point into the formulated mathematical function and predicts the output. The calculated performance of an ML model is measured based on the capability of the model to predict the correct outcome when a new dataset is given as input. One has to perform different tasks like collection of historical data and data integration, Exploratory analysis, Data preprocessing, feature selection, selection of model and parameters, training a model, performance analysis on the model, Deployment while developing any ML model. The below Figure 1 shows the general workflow of an ML model. There are wide varieties of ML algorithms like linear regression, decision trees, naive Bayes, and many more, which are used based on the use case. ML models typically try to solve the problems which fall under the categories of Regression, classification, and Clustering[3].

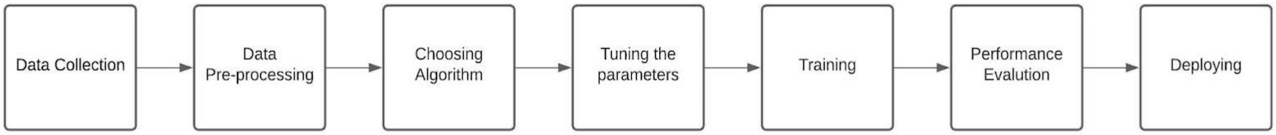

Fig. 1 Workflow of a Machine learning project

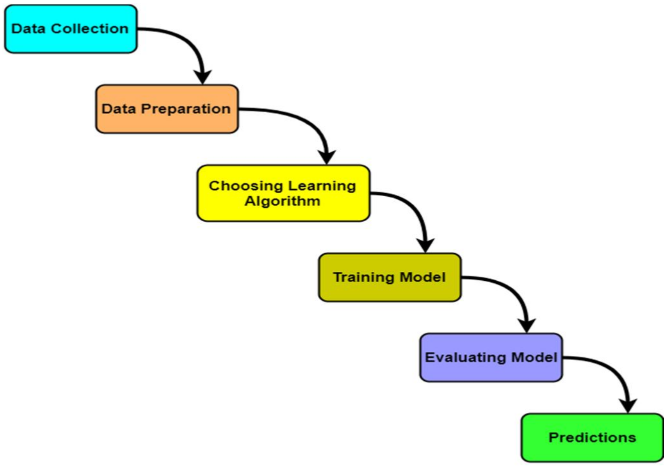

Fig. 2 Machine Learning Workflow

A. Data Collection

In this stage, we attempt to collect data from various sources in relation to our problem statement. It is important to collect large-quantity and high-quality data. Having large-quantity and high-quality data will help our machine learning model perform better and provide more accurate results.

B. Data Pre-Processing

In this stage, we ensure that our data has been refined and simplified, so that the trained machine learning model can simply handle, manage and understand data, and we also make sure that our data is balanced. we must also divide our dataset into the training and testing datasets.

Whatever data we divide into the train set is used to train our machine learning model, while whatever data we divide into the test set is used to evaluate our machine learning model.

C. Choosing Algorithm

The next stage in machine learning is algorithm or model selection. In the machine learning area, several machine learning models have previously been produced by various machine learning researchers, and these various machine learning models are suited for different types of data. So, based on the sort of data we have, we must choose the best model for our problem.

D. Tuning the Parameters

The majority of machine learning algorithms need some initial human interaction in order to choose the better acceptable values or data for certain parameters. So determine the best parameters for an algorithm in order to improve its performance in a certain working environment.

E. Training

In this stage, whatever selected data we have split into the training set must be passed to our machine learning model so that our machine learning model can understand or gain insight into the data.

F. Performance Evaluation

In this stage, we evaluate our trained machine learning model using a test set of data picked during the data pre-processing stage to determine how accurate the model is performing.

G. Deploying

In this stage, we expose our Machine Learning model to a real-world environment to solve our problem in real time.

III. TYPES OF DATA

Data generated by various sources will be in various forms:

A. Structured Data

Structured data is such data that cling to a predefined structure or a data model. As it has a predefined format it is easy to understand for humans as well as for computers. Generally, structured data is stored in well-defined schemes like tabular format. Relational database tables are the best examples, for instance, student information tables, customer tables, sales transaction tables are the best examples[4].

B. Unstructured Data

Unstructured data is defined as the data that is not well arranged or organized with a predefined structure. The unstructured data is very difficult to understand by humans and computers too. So, it is also called “Quantitative Data” as it cannot be analyzed using the traditional methods used for structured data. Generally unstructured data is stored in the “NoSQL” databases. Most of the data in the world is in the unstructured format and data mining techniques are used to structure the data from it. Some of the examples of Unstructured data are Media records, Webpages, Business documents, etc[5].

C. Semi Structured Data

Semistructured data is similar to structured data. But the only difference is it does not follow the tabular structure of the data model. Instead, it will use “tags”, which separate the data and make sure to create the hierarchy of the records within the data. Some of the examples of Semi Structured data are XML, JSO[6].

D. Qualitative data:

Qualitative data describes the clarity of the data. This data is easy to understand and it is categorical and unique to one individual.Some of the examples are Name, Grades, Citizenship etc.Qualitative data is divided into two types[7].

1) Nominal: Nominal data represents the “labeled” or “named” data which is segregated into various groups to avoid overlapping. Here, the data is not evaluated or mentioned, it is just assigned to various groups. The nominal data cannot be quantitative(i.e. measured), it is considered to be discrete. Some of the examples are hair color, gender, country, etc[8].

2) Ordinal: Ordinal data is categorical data with a set of range to it. A scale is a prerequisite for the data and the data should lie within the scale. Some of the examples are Age scale, Grades etc.

E. Quantitative data:

Quantitative data represents the data with a certain amount of measurement. Generally this data contains units to measure. For example, meters, seconds etc. This data is also called “numerical data”. Quantitative data is classified into two types:

1) Continuous: Continuous data describes the infinite values between a given range.Some of the examples are time, age, temperature, length, etc.

2) Discrete: Discrete data represents the values in the data set that are specific and countable. Some of the examples are the number of students in class, number of horses in the field , etc.

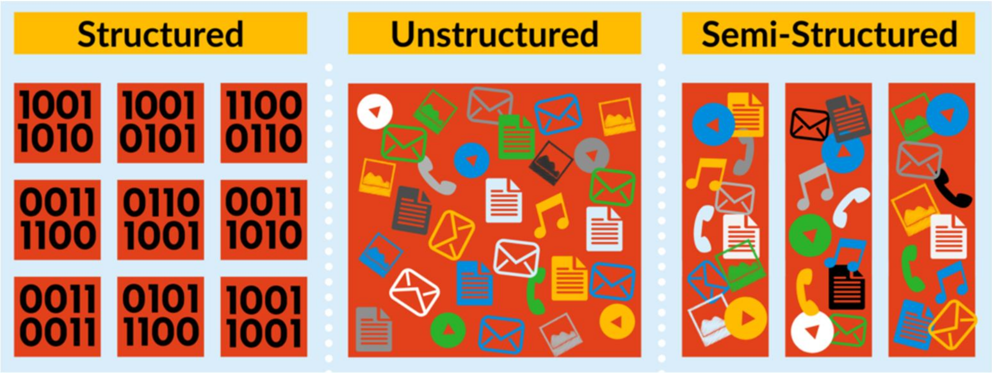

Fig. 3 Various Data Forms

IV. MACHINE LEARNING TYPES:

A. Supervised learning:

Supervised learning is defined as a sort of machine learning technique wherein the main machine trains on its own based on data that is perfectly labeled and categorized. Basically, a machine needs to find a mathematical function that can map the input to an output based on data available in the training dataset. Supervised machine learning can also be implemented with the help of the following tasks[9].

1) Regression: Regression is defined as one of the supervised learning techniques that works on continuous data values. Here, the main task is to generate an equation that best satisfies the training data. Based on the inferred equation, the Machine learning model then attempts to predict the outcome for test data[10]. Popularly used regression algorithms are:

- Lasso Regression
- Ridge Regression
- Linear Regression
- Decision tree algorithm
- Gaussian Process Regression

2) *Classification:* Classification is also a supervised machine learning technique. Unlike Regression, the technique of classification works on discrete data variables, i.e. values that are specific and countable. As the name suggests, the classification algorithm makes an effort to classify or categorize the data provided to it into classes or groups by assigning a class label. In classification, the model trains on all the features of the training data and uses this knowledge to predict class labels for a given test data[11]. Popularly used classification algorithms are:

- Naive Bayes
- Support Vector Machines (SVM)
- Linear Discriminant Analysis
- Logistic Regression
- K-Nearest Neighbors

B. Unsupervised Learning :

Unsupervised learning is a machine learning technique employed to work with unlabeled datasets. With unsupervised learning, it becomes easy to identify patterns, groups, and features of an unlabeled dataset[12]. Clustering is one of the most often used unsupervised learning strategies.

1) *Clustering:* as the name suggests, is an unsupervised machine learning technique in which identifies groups in the given data by identifying similarities of the data points in the dataset. The technique of clustering tries to ensure that there is less intracluster distance and high intercluster distance. The term Intracluster distance refers to distance between any data points within a given cluster and inter cluster distance is defined as the distance between given data points of two different types of clusters[13]. Prominently used algorithms in clustering are:

- K-Means
- K-Medoids
- Mean-Shift
- DBSCAN
- Gaussian Mixture Modeling
- Agglomerative hierarchical

C. Reinforcement learning:

Reinforcement learning is defined as a sort of machine learning technique that learns how to solve problems or computations via trial and error method.

By using Reinforcement learning we can overcome the limitations of deep learning to solve multi-step problems. In Reinforcement learning, the process begins with a given agent that interacts with its surrounding environment. Here the agent is aiming to accomplish a multi step goal it has, within the given environment.The environment has a state, which the agent always can observe.

The agent observes the environment state based on what it senses, sees, feels, or hears through any other ways. The agent can perform activities that modify the state of the environment.

Finally, as the agent gets closer to its goal, it receives reward signals. These reward signals are used by the agent to evaluate which acts were successful and which were not.

We continue this state, action, and reward cycle until the agent learns how to operate efficiently inside the environment through trial and error. The agent's purpose is to learn in any situation how to always take a proper action, in any state of the environment, that will get closer to its particular goal[14].

A self-driving automobile, for example, may attempt to drive on real-world roadways. Its mission is to transport from house to business place while avoiding any obstacles in its way. The state contains the position of the automobile, the road conditions, and the location of other cars.

An agent has the ability to drive forward, backward, turn right, or turn left. All of these agent activities alter the car's current position on the road in relation to its surroundings.

Furthermore, if these activities collide, they may have an impact on other cars and obstacles. We may award a point for each unit of distance traveled that brings the agent closer to the goal. And also deduct points for each time violating traffic regulations, deviating from the path, or colliding with obstacles[15].

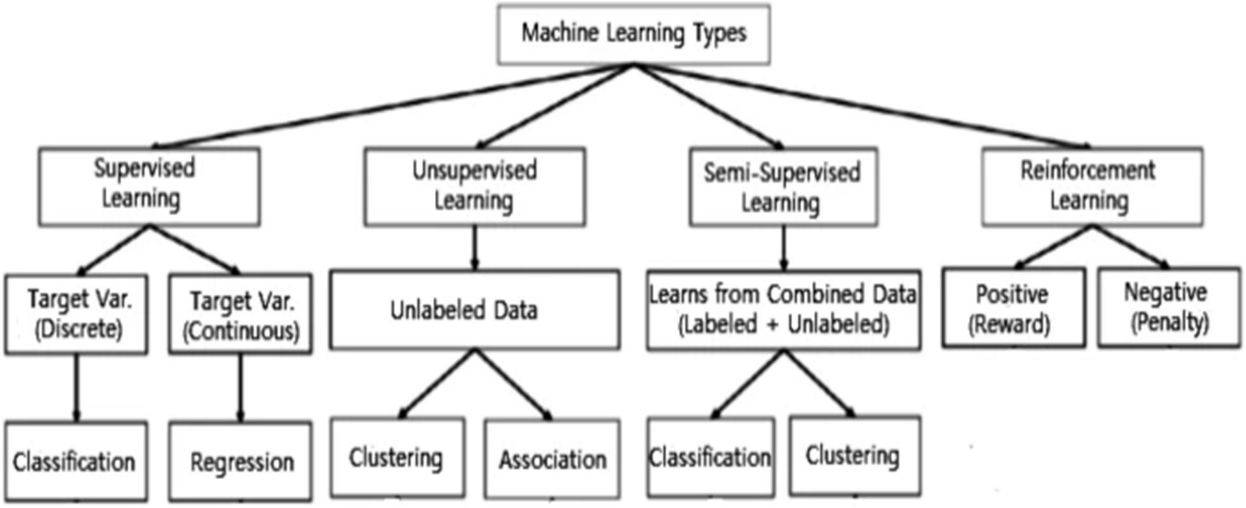

Fig. 4 Types of Machine learning

TABLE I
Tools Used for Model Development in Python

S.No	Tool Name	Open Source	Developer	Platform	Language	Description	Features
1.	Pandas	Yes	Wes McKinney	Cross-Platform	Python	The name "Pandas" is derived from "panel data." It is an open-source foundation python software library for data analysis and processing. It is used for manipulating time series and numerical tables[16].	i. fancy indexing mechanism and integrated indexing for data manipulation. ii. Supports different file formats like CSV, MS excel,SQL,Fast HDF5 formats. iii. Pliable in reshaping and pivoting datasets. iv. intelligent label based slicing v. Automatic label-based alignment and manipulates messy data.
2.	SKLearn	Yes	David Cournapeau	Cross platform	Python	Scikit-Learn, also known as sklearn, is a Python numerical and scientific library built to work with NumPy and Scipy. Scikit-learn makes use of NumPy and is widely used for high-speed linear algebra and array operations. It's a significant and simple method for analysing predictive data. NumPy, SciPy, and Matplotlib were used to create this.	i. SKlearn has many pre-cleaned data sets which can be used directly by importing ii. SKLearn integrates Matplotlib, plotly, Numpy, Scipy, and many more. iii. Sklearn features the following: a.clustering, b.regression, c.classification, d.dimensionality reduction algorithms and many more.

3.	Tensor Flow	Yes	Google Brain Team	Cross platform	Python	Tensorflow is a flexible ecosystem of tools, libraries, and a large community of resources that makes building and deploying machine learning applications simple for academics and developers. Classification, Understanding,Perception, Prediction, Discovering, and Creation are all done with TensorFlow. Tensorflow constructs models using different data flow graphs and also enables developers or programmers to build large scale neural networks with several various layers[17].	i. By using tensor flow one can create a large-scale deep learning model with many layers. ii. Tensorflow has the capability to run on multiple CPUs and GPUs. iii. It is an easy model building as well as robust and very powerful experimentation for researchers. iv.Tensorflow has packages for programming languages like C#, Haskell, Julia, R, Scala,Rust, MATLAB and many more.
4.	Keras	Yes	François Chollet	Cross-platform	Python	Keras means Python Deep Learning Library.Keras is basically used to provide an interface for artificial neural networks like tensorflow library. It's the most popular deep learning framework on the market. Keras is the best framework for quickly experimenting with deep neural networks.	i. keras uses a global state to implement functional-model building API ii. Keras also uses that global state to uniquify auto generated layer names. iii. (Keras allows to distribute the computing power of training a deep learning models on clusters of TPU(Tensor Processing Unit) and GPU(Graphic Processing Unit)) iv. Keras simplify the coding necessary for deep neural network code. v. It has plenty of tools to work on image and text data. vi. Keras allows users to produce deep models on smartphones,on the web and even on java virtual machines.
5.	PyTorch	Yes	Facebook's AI Research lab (FAIR)	Cross platform	Python	Pytorch is basically an optimized version for torch library which is used for deep learning.Pytorch uses CPUs and GPUs.The main motto of pytorch is computer vision and natural language processing.The interface for pytorch is more enhanced and dynamic with python and pytorch also has c++	i. Easy,efficient and interoperability in order to leverage the rich ecosystem of python libraries as part of user programs. ii. Writing programs in pytorch is more pythonic, in turn it will be easy for data scientists iii. Pytorch uses tensor

						interface[18].	computing via GPU with strong acceleration which makes pytorch faster. iv. Type based automation differentiation system is developed to build a deep neural network.
6.	NLTK	Yes	Steven Bird, Edward Loper, Ewan Klein (Team NLTK)	Cross platform		NLTK is used to build python programs which work with data of language used by humans.This platform also reinforce research as well teaching in Natural Language Processing for human language which is composed in python.It has a user manual kind of book, explains the fundamental concepts of behind the language processing tasks of the toolkit.	i.NLTK is a community driven project. So it has huge community support. ii.NLTK also contributes the following: a.Classification, b.tokenization, c.stemming, d.tagging, e.parsing, f.semantic reasoning functionalities.
7.	SciPy	Yes	Travis Oliphant, Pearu Peterson, Eric Jones	Cross-platform	Python	SciPy is used for mathematical,engineering, scientific as well as technical computing, the SciPy ecosystem includes a variety of data management and computation tools, as well as efficient research and high-performance computation. SciPy library currently has a BSD license.SciPy uses high-level Python commands to manipulate and visualize data[19].	i. SciPy connects with databases quickly and flawlessly. ii. SciPy mainly has these outstanding features: a. Linear algebra operations. b. Interpolation c. Optimization and fit d. Statistics and random numbers. e. Numerical Integration f. Fast Fourier Transforms g. Signal Processing h. Image manipulation
8.	OpenCV	Yes	Gary Bradsky	Cross-platform	C++, Python, Java	Opencv stands for open source computer vision , which is basically designed to work out for the problem of computer vision.Opencv main motto is to provide real-time computer vision.Opencv is also used for video capture and analysis.Opencv free for use under the open-source Apache2 License. Because OpenCV is written in C++, the primary interface is written in C++. However, OpenCV also supports Python, Java, and other programming languages[20].	The main features that made opencv more popular are as follows: i. General object recognition, ii. Edge detection, iii. Feature matching for object recognition, iv. Color filtering, v. Subtracting the background from images, vi. Gesture recognition vii. Mobile robotics viii. Motion tracking and many more

9.	Numpy	Yes	Travis Oliphant	Cross-platform	Python	NumPy is a very essential and primitive package needed for scientific computing.NumPy basically comes up with an encapsulated N-Dimensional array of homogeneous data types in a N Dimensional array object.Numpy arrays have a fixed size at the time of creation and elements required to be of same data type which these makes more efficient with scientific computing.Numpy contains pre-compiled C code which is optimized.	i. Numpy mainly has the following features which made numpy so popular: ii. High-performance iii. Integrating code from C/C++ languages iv. Multidimensional container v. v.Broadcasting functions vi. Work with varied databases vii. Additional Linear Algebra functions.
10.	Google Auto ML	No	Google	Cloud		Google auto ML is a cloud platform which is used to design ML models in the google cloud and then we can integrate the model in our application.We create models by using GUI.By using google auto ML we can create the ML models even without having the knowledge in any programming language and ML models.	i. AutoML Natural language(this is used for text and documents) ii. AutoML Tables iii. AutoMLTranslation iv. AutoMLVideo Intelligence v. AutoML Vision and much more
11.	Teachable Machine	No	Google	Web-based		By using a teachable machine we can create sophisticated models which can recognize images,sounds and poses.No coding is required for creating the models everything is done by using GUI and then we can integrate the model once it is developed.	We can train the following in teachable Machine i. Images ii. Sounds iii. Poses
12.	AWS Machine Learning	No	AWS	Cloud		Amazon Machine Learning is an Amazon web services product which is used to find patterns in the data by using some sophisticated algorithms , developing mathematical and predictive models on the data.AWS provides all the different models which one can use on their data for getting different insights.Mainly AWS provides models which are highly useful for a business in order to manage their customers.	i. Intelligent Contact Center ii. Personalization iii. Intelligent document processing iv. Intelligent Search v. Fraud Detection vi. Media Intelligence vii. Business Forecasting viii. and much more

TABLE II
Tools Used For Data visualization

S.No	Tool Name	Open Source	Developer	Platform	Language	Description	Features
1.	Matplotlib	Yes	Michael Droettboom	Cross-Platform	Python	Matplotlib is a Python package that is used to visually depict data. Matplotlib contains different methods which are used to represent the data in different graphical forms.By representing the data graphically we can gain more knowledge about the data than seeing the data as it is.Matplotlib contains an Object Oriented API which is used for the process of embedding the plots into the apps which are developed by using GUI toolkits like wxpython, Tkinter, GKorQt. [16]	By using Matplotlib we can draw the following i. Histogram ii. Scatter plot iii. Stream plot iv. Line plot v. Image plot vi. 3D plot vii. Barcharts, pie charts, tables, polar plots And many more[21].
2.	Seaborn			cross-platform	Python	Seaborn is a python library it is used to develop the statistical graphs from the data.It is developed on the top of matplotlib.seaborn has many options to customize the plot which is drawn by using matplotlib.seaborn has many default options for an attractive plot.Seaborn is like a statistical plotting frontend for matplotlib[21].	i. Updates plots ii. closely integrated with pandas data structures. iii. KDEplots iv.Pairplot v.Violin plot vi.distplot vii.swarmplot
3.	Tableau	No	Christian Chabot, Pat Hanrahan, Chris Stolte Andrew Beers		Python	Tableau is a data visualization software which mainly focuses on business Intelligence. Tableau won CODie award for “Best Business Intelligence or Knowledge Management Solution” in 2008.Tableau has different software products like Tableau desktop, prep builder, public, mobile, server, Vizable, reader, CRM, Online[22].	i. Tableau Dashboards which are used to combine different reports into one. ii. Collaboration and sharing. iii. Cloud storage support. iv. Third party integration. v. Support all the advanced visualization techniques. vi. Drag and drop interface.
4.	Plotly	Yes	Alex Johnson, Jack Parmer, Matthew Sundquist and Chris Parmer.	Cross-Platform	Python Language	The Plotly Python package is an open-source,interactive plotting framework that supports many different chart types for a variety of statistical,scientific, geographic, financial,and 3D applications.It offers enterprise products like Dash Enterprise’ Chart Studio Cloud, Chart Studio Enterprise, Data visualization libraries, Fig Converters[23].	By using Plotly we can draw the following. i. Scatter plot ii. Dot plot iii. Filled area plot iv. Box plot v. Tree plot vi. Quiver plot

V. MACHINE LEARNING APPLICATIONS

A. Amazon Go

Amazon Go is the world's most advanced, human-free retailer, employing technology such as machine learning and deep learning algorithms, as well as several cameras from various angles. A consumer is spotted and recognised using the CV algorithm and face recognition. It detects if an item has been selected or returned to the shelf using a sensor fusion technique. When a consumer leaves the store, the complete amount of his purchases will be automatically deducted from his credit card, eliminating the need for the customer to stand in line at the bill counter. The programme can anticipate which item will be purchased by the individual based on the customer's frequent recent purchase history using deep learning[24].

B. Google microscope: (cancer cell detection)

Google researchers created an augmented reality microscope that uses machine learning algorithms to detect cancer tissue cells that standard electronic microscopes can't see. Because cancer is difficult to notice, each and every cell must be analyzed to detect it. Almost all pathologists now have access to machine learning due to this technology. To begin, data is input into a computer, which performs billions of computations before detecting cancerous cells. They use numerous example photos of cancer tumor cells to train the machine learning model, and the computer also learns from pathologists' input. This microscope is now used to identify breast cancer; however, researchers are working on detecting other cancers in the future.

C. Tesla

Tesla employs two AI processors to make autopilot more reliable, using raw camera data as input for Deep Neural Networks that give depth prediction and object identification. To produce 3D depiction of the surroundings, a bird's eye view is employed. Tesla has trained the model in a variety of complex simulations for tough conditions. It took 70,000 hours of GPU training (equal to 8 years) to create 48 neural networks that function on their own[25].

D. Netflix Recommendation

The Netflix recommendation system predicts which episodes a user will watch based on viewing history, how previous content titles are rated, genre, actors, year of release, categories, time and date a viewer watches a show, how long a viewer watches a show, which device viewer watches the show on, the scenes that are repeatedly watched, and so on[26].Consider an hour-long episode of "Stranger Things," which comprises over 86000 video frames. All 86000 frames are analyzed by the AVA in order to determine the best thumbnails that may be shown. Based on their involvement with prior material, different viewers experience changes in thumbnails of the content on a frequent basis. A viewer who enjoys humor will see a thumbnail portraying a comedy scene from the film, whereas a viewer who enjoys horror will see a thumbnail depicting a horror moment from the film. The thumbnails may also vary depending on the viewer's location. The graph depicts the most popular thumbnails for the same material in various areas. Different thumbnails are shown for the same movie according to users liking.

E. Spotify Recommendation

Spotify filters out music based on the preferences of a comparable set of users in order to propose songs and playlists to a new user with similar tastes. Collaborative filtering is used to accomplish this. As a result, they began their investigation into user music recommendations. Spotify makes use of a technology known as "Collaborative Filtering." This is one of the first and most basic techniques of providing consumers with music suggestions.

Spotify began to utilize NLP in conjunction with collaborative filtering to improve their recommendation engine because collaborative filtering only works on a small number of members and is ineffective at promoting freshly released music to users. The NLP model gathers information from social media, news, blogs, and other sources to determine what is hot in the music industry. New songs and albums are recommended to a user based on insights from various sources and collaborative filtering. If the material from the sources is bogus, these models might be prejudiced and fully propose the wrong music to a user. Spotify built a Convolutional Neural Network supplied with technical characteristics of an audio file such as the Beats Per Minute(BPM), loudness, amplitude, and many others as an enhancement to the current audio model. Spotify now only suggests a song or playlist to a user if it passes the Collaborative Filtering, NLP audio model, and Convolutional Neural Network tests[27].

VI. MACHINE LEARNING TECHNIQUES

A. Analysis Based On Machine Learning For Behavioral Distinction (Depression And Anxiety Depression)

The researchers wanted to explore if they could tell the difference between depression and anxiety by looking for a specific pattern of biased reaction to emotional cues in each condition. Test battery was created to help us achieve our goal. The battery focuses on four common types of bias: Attention biases, memory biases, self-interpretation biases, and Expectancy Biases. Even in the face of significant scoring noise and instrumental, these approaches allow the identification of complicated nonlinear high dimensional specific interactions that may influence the output predictions. ML approaches based on the decision tree algorithms that were applied specifically. They were created to be more sensitive enough to classify subjects into four different categories: low anxiety level and high depression level [HD], low depression level and high anxiety level [HA], low anxiety level and depression level [LAD], and high anxiety level and depression level [HAD]. Using the replies on the self-report questionnaires, an asymptomatic profile was created for each individual participant by labeling depression and anxiety levels. Based on the behavioral tasks, the machine learning algorithm was taught to infer each of the participant symptomatic given profiles[28].

ML algorithms analyzed the information and predicted the outcome. Based on their success in the exercises, each participant is assigned to a group. In a two-group model, the HD and HA groups had 72% and 71% specificity and sensitivity, and 68% and 74% classification accuracy respectively. These categorization accuracy rates were all higher than the chance rate. The value of integrating behavioral data with machine learning technologies in the field of mental diagnoses is demonstrated by these findings[29].

B. Image Processing Based On Machine Learning For Satellite Image Analysis

Satellite photos are huge in size, acquired from a long distance, and are affected by noise and many other environmental conditions. As a result, they must be treated and cleaned before researchers can analyze them. Many real time applications, such as navigation, agriculture, land identification, and geographic information systems, rely on satellite pictures. Satellites capture remote sensing photos that are used in agriculture, navigation, defense and other fields. Due to the presence of several differences, the satellite pictures range greatly in terms of color variations, textural contrasts, and they are also highly complicated. Hence applying different processing techniques to the satellite data is rather challenging. In addition, satellite data is captured over great distances and is influenced by the presence of undesired interferences, lowering image quality[30].

Remote sensing is now a constantly increasing technique for gathering satellite data and is critical for analyzing spatiotemporal changes on the earth's surface. Government agencies and Researchers all around the world gather this data in order to study the earth's changes. Global climate change estimation, environmental monitoring as considered under disaster management, land cover change detection, security, and urban and land development, to mention a few applications, all benefit from remote sensing data. A variety of mathematical techniques and algorithms for satellite image processing have been presented[31].

C. Image processing based on machine learning for detecting plant diseases:

Image processing systems may be used for a variety of purposes, including image classification, feature extraction, object tracking, segmentation and reconstruction, to name a few. Image processing is a collection of processes and techniques that are routinely used on digital pictures. The term "vision" is commonly used to describe such procedures. There is, however, a distinction to be made between computer vision and image processing. Image processing techniques, for example, are not confined to the visible region of the electromagnetic spectrum, whereas vision systems are only capable of duplicating visual processes in pictures that are visible to the naked eye. Thus, image processing tools may be used to process any sort of picture, including thermal images, gamma rays, and electromagnetic resonances, to name a few. In any event, image processing and vision techniques require a representation of the physical world that we observe, in the form of a digital picture as the foundation of their operations. A sampling method on an analogue signal from digital capture devices captures digital images[32]. At a lower cost, research is being performed to boost agricultural yield and improve crop quality. As a result, early diagnosis of disease in crops is critical for productivity and loss prevention. Farmers and specialists use traditional disease detection procedures that are continuous, naked, and antiquated since they are time consuming, costly, and hard. Furthermore, because some countries lack adequate agricultural infrastructure, contacting an expert is out of reach. Therefore to address these difficulties, researchers are using Image Processing Techniques to automate the illness diagnosis procedure. Researchers create and use Machine Learning algorithms to quickly and accurately diagnose plant diseases. As a result, worker effort is reduced and productivity is increased. Detection of diseases for a crop is essential to increase the production.Detection techniques which are endorsed by farmers are inefficient, and approaching an expert for detection is merely impractical[33].

VII. CONCLUSIONS

In this paper, we have discussed different types of tools used in Machine learning for training, modeling a machine, and the visualization of data. In addition, we also explored various categories of data, types of machine learning, and techniques used to solve disparate problems. Many famous organizations such as Netflix and Amazon use different tools for their applications to be robust and efficient. There are various tools for different tasks; for instance, we use matplotlib for visualization and scipy library for scientific computations and so on. Nowadays, data is crucial as well as analysis and processing of data are much more important, so we require some tools to do tasks. As data usage gradually increases daily, a data scientist or data science enthusiast needs to manage and handle the data, to use them efficiently. Hence it leads to more demand for tools that make our work faster and better. So, knowing the various machine learning tools and how they are used practically will help build and analyze a machine learning model and its applications.

REFERENCES

[1] O. Obulesu, M. Mahendra and M. ThrilokReddy, "Machine Learning Techniques and Tools: A Survey," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 2018, pp. 605-611, doi: 10.1109/ICIRCA.2018.8597302.

[2] Sundareswaran, Veena & Shankari, T & Sowmiya, Senthil & Varsha, Mundhra. (2020). A SURVEY ON TOOLS USED FOR MACHINE LEARNING.

[3] Yogesh, Singh & Bhatia, Pradeep & Sangwan, Om. (2007). A REVIEW OF STUDIES ON MACHINE LEARNING TECHNIQUES. International Journal of Computer Science and Security.

[4] Kilari, Hemashree & Malkari, Ram Kishore. (2021). Data Forms. 8. 1928.

[5] S. Sato, A. Kayahara and S. Imai, "Unstructured data treatment for big data solutions," 2016 International Symposium on Semiconductor Manufacturing (ISSM), 2016, pp. 1-4, doi: 10.1109/ISSM.2016.7934512.

[6] Sarangam, Ajay. “Semi Structured Data: A Comprehensive Guide In 7 Points.” Semi Structured Data: A Comprehensive Guide In 7 Points, Jigsaw academy, 17 march 2021.

[7] Techtarget. “Qualitative data” What is Qualitative Data? https://www.techtarget.com/searchcio/definition/qualitative-data

[8] Lange, Sacha, Thomas Gabel, and Martin Riedmiller. "Reinforcement Learning: State of the Art." (2011).

[9] Kotsiantis, Sotiris. (2007). Supervised Machine Learning: A Review of Classification Techniques.. Informatica (Slovenia). 31. 249-268.

[10] Bender, Ralf, and Ulrich Grouven. "Logistic regression models used in medical research are poorly presented." BMJ 313.7057 (1996): 628.

[11] John R. Hodges & Bruce Miller (2001) The Classification, Genetics and Neuropathology of Frontotemporal Dementia. Introduction to the Special Topic Papers: Part I, Neurocase, 7:1, 31-35, DOI: 10.1093/neucas/7.1.31

[12] Wang, D. “Unsupervised Learning: Foundations of Neural Computation”. AI Magazine, vol. 22, no. 2, June 2001, p. 101, doi:10.1609/aimag.v22i2.1565

[13] Makagonov, P., Alexandrov, M., Gelbukh, A. (2004). Clustering Abstracts Instead of Full Texts. In: Sojka, P., Kopeček, I., Pala, K. (eds) Text, Speech and Dialogue. TSD 2004. Lecture Notes in Computer Science(), vol 3206. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30120-2_17

[14] A. Alharin, T. -N. Doan and M. Sartipi, "Reinforcement Learning Interpretation Methods: A Survey," in IEEE Access, vol. 8, pp. 171058-171077, 2020, doi: 10.1109/ACCESS.2020.3023394.

[15] Lange, Sacha, Thomas Gabel, and Martin Riedmiller. "Reinforcement Learning: State of the Art." (2011).

[16] Lemenkova, Polina, Python Libraries Matplotlib, Seaborn and Pandas for Visualization Geo-spatial Datasets Generated by QGIS (September 25, 2020). Analele stiintifice ale Universitatii "Alexandru Ioan Cuza" din Iasi - seria Geografie, vol. 64(1), pp. 13-32, 2020, Available at SSRN: https://ssrn.com/abstract=3699706

[17] Sharma, Shivani, and Sudhir Kumar Sharma. "A study on machine learning tools." IITM JOURNAL OF MANAGEMENT AND IT 11.1 (2020): 98-102

[18] Morvan, Mario, et al. "PyLightcurve-torch: a transit modeling package for deep learning applications in PyTorch." Publications of the Astronomical Society of the Pacific 133.1021 (2021): 034505

[19] Virtanen, P., Gommers, R., Oliphant, T.E. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020).

[20] https://doi.org/10.1038/s41592-019-0686-2Tuohy, Shane, et al. "Distance determination for an automobile environment using inverse perspective mapping in OpenCV." (2010): 100-105.

[21] Hafeez, Abdul & Sial, Ali. (2021). Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python [HEC Y Cat]. International Journal of Advanced Trends in Computer Science and Engineering. 10. 2770-281. 10.30534/ijatcse/2021/391012021.

[22] Nair, L., Shetty, S. & Shetty, S. (2016). Interactive visual analytics on Big Data: Tableau vs D3.js. Journal of e-Learning and Knowledge Society, 12(4),. Italian e-Learning Association. Retrieved June 24, 2022 from https://www.learntechlib.org/p/173675/

[23] Eugenia Voytik, Isabell Bludau, Sander Willems, Fynn M Hansen, Andreas-David Brunner, Maximilian T Strauss, Matthias Mann, AlphaMap: an open-source Python package for the visual annotation of proteomics data with sequence-specific knowledge, Bioinformatics, Volume 38, Issue 3, 1 February 2022, Pages 849–852, https://doi.org/10.1093/bioinformatics/btab674

[24] K. Wankhede, B. Wukkadada and V. Nadar, "Just Walk-Out Technology and its Challenges: A Case of Amazon Go," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 2018, pp. 254-257, doi: 10.1109/ICIRCA.2018.8597403.

[25] Csongor, Rob. “Tesla Raises the Bar for Self-Driving Carmakers.” NVIDIA Blog, https://blogs.nvidia.com/blog/2019/04/23/tesla-self-driving

[26] Amatriain, X., Basilico, J. (2015). Recommender Systems in Industry: A Netflix Case Study. In: Ricci, F., Rokach, L., Shapira, B. (eds) Recommender Systems Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7637-6_11

[27] Pérez-Marcos, J., López Batista, V. (2018). Recommender System Based on Collaborative Filtering for Spotify's Users. In: , et al. Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017. PAAMS 2017. Advances in Intelligent Systems and Computing, vol 619. Springer, Cham. https://doi.org/10.1007/978-3-319-61578-3_22

[28] Richter, T., Fishbain, B., Markus, A. et al. Using machine learning-based analysis for behavioral differentiation between anxiety and depression. Sci Rep 10, 16381 (2020). https://doi.org/10.1038/s41598-020-72289-9

[29] Sajja, Guna. (2021). Machine Learning based Detection of Depression and Anxiety. International Journal of Computer Applications 183. 20-23. 10.5120/ijca2021921856.

[30] A. Asokan and J. Anitha, "Machine Learning based Image Processing Techniques for Satellite Image Analysis -A Survey," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 119-124, doi: 10.1109/COMITCon.2019.8862452

[31] Ferdous, H., Siraj, T., Setu, S.J., Anwar, M.M., Rahman, M.A. (2021). Machine Learning Approach Towards Satellite Image Classification. In: Kaiser, M.S., Bandyopadhyay, A., Mahmud, M., Ray, K. (eds) Proceedings of International Conference on Trends in Computational and Cognitive Engineering. Advances in Intelligent Systems and Computing, vol 1309. Springer, Singapore.

[32] https://doi.org/10.1007/978-981-33-4673-4_51Oliva, Diego; Hinojosa, Salvador (2020). [Studies in Computational Intelligence] Applications of Hybrid Metaheuristic Algorithms for Image Processing Volume 890 || . 10.1007/978-3-030-40977-7()

[33] B. S. Kusumo, A. Heryana, O. Mahendra and H. F. Pardede, "Machine Learning-based for Automatic Detection of Corn-Plant Diseases Using Image Processing," 2018 International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2018, pp. 93-97, doi: 10.1109/IC3INA.2018.8629507