Big Data Analytics with Machine Learning Models

Manish Kumar*, Dr. (Prof.) Deva Prakash**

*(Research Scholar, Department of Mathematics & Computer Science, Magadh University, Bodhgaya, Bihar
**(Associate Professor, Head of Department, Department of Mathematics, S.M.D. College, Punpun, Bihar)

Abstract
Data that is too big, moving too quickly, or complex to process using conventional techniques is referred to as "Big Data." Formerly, the 3Vs were employed in Big Data, but today the 5Vs — volume, velocity, variety, veracity, and value — are used. While there is no doubt that these huge data have great potential. The artificial intelligence technique of information discovery for thoughtful decision-making is called machine learning. The most popular technologies for study in various analytics and computations today are Big Data analytics and machine learning. Although data preparation for Big Data is a topic in and of itself, large data may benefit from machine learning. Bigdata may facilitate the creation and fine-tuning of incremental/online/stream-oriented ML algorithms; in particular, it may be worthwhile to consider models that have already been created for drifting data. In most cases, learning is created by performing in-depth computations on pre-existing datasets to produce a learning model. Since data sizes are growing daily and a typical system cannot manage very large dataset calculations, the discovered model needs to be adjusted accordingly. Machine learning makes utilization of Big Data approaches. We are all aware that large data sets are ideal for machine learning, and here is where Big Data comes into play. Big Data is used to glean hidden knowledge or important insights from massive data sets. In a nutshell, we may assert that machine learning would be useless without large data. Big Data analytics and machine learning are crucial for classification and prediction in many businesses, including those in the health, education, agriculture, manufacturing, banking, and other industries. Data must be provided to machine learning models as input, and sometimes the more comprehensive the data, the better the model's output. To create the required result in such a scenario, huge data is provided as an input to a machine learning model. One of the input sources for the machine learning model could be Big Data. Machine learning is a technique used in artificial intelligence to find information that may be used to make wise decisions. This article discusses machine learning methodologies, key Big Data technologies, and a few machine learning applications in Big Data. It also explains machine learning algorithms in Big Data analytics, and machine learning challenges us to make decisions where there is no known "right path" for the specific problem based on previous lessons. It also enumerates some of the most widely used tools for analyzing and modeling Big Data.

Keywords: Big Data, Machine Learning, Big Data Analytics, Machine Learning Algorithms, Information Technology, Stream processing, Apache Foundation

1. Introduction

In a wide range of applications, including computer vision, audio processing, natural language comprehension, neuroscience, healthcare, and Internet of Things, machine learning (ML) systems have had a profoundly positive social impact. The focus of ML is on developing a framework for a process that gets better over time [1]. When it comes to certain activities and performance measures, the problem of learning from the past is referred to as a machine learning (ML) challenge. Users using ML techniques can extract hidden structure from large data sets and make predictions. With competent learning methods, abundant and rich data, and efficient computing environments, ML thrives. Figure 1 depicts the development of machine learning.

1.1 Big Data

In traditional data, large and complicated problems are tackled by a single computer system using centralized database design. Processing a lot of data requires centralized architecture, which is expensive and inefficient. Big Data is built on the distributed database design, which divides a large block of data into multiple smaller pieces to solve the problem. Then, numerous computers connected to a particular computer network compute the answer to a problem. In order to solve a problem, the computers converse with one another [2]. Comparing the distributed database to the centralized database, the distributed database offers better computation, reduced costs, and improved performance. This is so because distributed database systems, which use microprocessors, are more cost-effective than centralized architecture, which is based on mainframes. Additionally, compared to the centralized database system used to manage conventional data, the distributed database has higher processing capacity.

A large amount of data that cannot be processed or stored by any conventional data storage or processing devices is referred to as "Big Data." Data is produced on an enormous scale, and many multinational corporations use it to process and analyze it in order to find insights and enhance the operations of several organizations.

Big Data is the term used to describe the enormous volumes of data that these smart gadgets will produce. Big Data is defined by the authors of [3] using the following 5Vs:

Volume: The term "Big Data" inherently refers to an immense magnitude.

Volume contains a tremendous amount of info.

The magnitude of the data is a very important factor in determining its value.

The term "Big Data" is really used to describe data that is exceedingly huge in volume. This means that the amount of data will determine whether or not a given set of data may be called a Big Data.

Consequently, when working with Big Data, it is vital to take into account a certain "Volume."

Example: In 2018, global mobile data traffic amounted to 19.01 exabytes per month. By 2022, mobile data traffic is expected to reach 77.5 exabytes per month worldwide at a compound

annual growth rate of 46 percent.

Velocity: Velocity is the rapid rate of data collection.

• In Big Data velocity, information comes from a variety of sources, including devices, networks, social media, mobile phones, etc.
• A vast and constant flow of data is present. The speed at which data is created and processed to satisfy demands influences the data's potential.
• Sampling data can be useful in addressing problems like "velocity."
Example: Google receives more than 3.5 billion searches per day. Additionally, the number of Facebook users is rising by around 22% yearly.

Variety: Structured, semi-structured, and unstructured data are all referred to as well as diverse sources of data.

• The emergence of data from fresh sources, both inside and outside of a company, is Essentially what is meant by variety. It might be organized, somewhat organized, or unorganized. The data collection methods may include photographs, excel files, and graphs [4].

There are commonly three different types of Big Data. These are as follows:

(a). Structured Data: Structured data is data that follows a pre-established data model and is easy to analyze. A tabular format with relationships between the various rows and columns is what structured data follows.
Excel spreadsheets and SQL databases are typical instances of structured data. Each of them has structured, sortable rows and columns. A data model, which is a representation of how data can be stored, processed, and accessed, is necessary for the existence of structured data. Each field is discrete and can be accessed independently or in conjunction with data from other fields thanks to a data model. Because it is feasible to swiftly aggregate data from many areas in the database, structured data is incredibly powerful.

(b). Semi- Structured Data: Semi-structured data is a type of structured data that does not adhere to the formal organization of data models linked to relational databases or other types of data tables, but still contains tags or other markers to enforce hierarchies of records and fields within the data and to separate semantic elements. As a result, another name for it is self-describing structure. JSON and XML are two types of semi-structured data that serve as examples. Between structured and unstructured data, this third category was created since semi-structured data is much simpler to examine than unstructured data. The capacity to "read" and process either JSON or XML is a feature of many Big Data solutions and technologies. When opposed to unstructured data, this makes it simpler to examine structured data. Example- Comma Separated Value (CSV)

(c). Unstructured Data: Unstructured data is information that is either not arranged in a predefined way or does not have an established data model. Unstructured data can also include facts like dates, numbers, and figures but is often text-heavy. In contrast to data stored in organized databases, this produces anomalies and ambiguities that make it challenging to understand using conventional algorithms. A couple more common types of unstructured data include audio and video files.

Veracity: In order to apply the analytics method, agricultural data must be exact and correct.

- It relates to data inconsistencies and ambiguity; that is, readily available data can occasionally become disorganized, and quality and accuracy are challenging to manage.
- Because there are so many different data dimensions arising from several dissimilar data kinds and sources, Big Data is also unpredictable.
- As an illustration, more data can be confusing, but less information can only transmit partial or incorrect information.

Value: Value, the fifth V, is taken into consideration after the four others. The majority of data that has no value is useless to the organization until you can make it valuable.

• Data by itself is useless and unimportant; information must be extracted from it by transforming it into something worthwhile. Therefore, you might say that Value! is the most crucial of the five virtues.

1.2 A SUMMARY OF MACHINE LEARNING

1.2.1. History:

- 1642, Blaise Pascal creates the mechanical addition, subtraction, multiplication, and division machine.
- 1679, Gottfried Wilhelm Leibniz created the binary coding method.
- Charles Babbage develops the concept for a general-purpose computer that could be programmed using punched cards in 1834.
- 1842, Ada Lovelace, the first programmer, presents a series of processes for resolving mathematical issues using Charles Babbage's hypothetical punch-card machine.
- 1847 - George Boole develops Boolean logic, a type of algebra in which all values are convertible into the true or false binary values.
- 1936 - Alan Turing, an English cryptanalyst and logician, puts forth the idea of a universal machine that could decipher and carry out a set of instructions. His published demonstration is regarded as the cornerstone of computer science.
- 1952 - Arthur Samuel develops a software to aid an IBM computer in improving its checkers performance over time.

- 1959 - The first artificial neural network, MADALINE, is used to solve a real-world issue, the elimination of echoes from phone lines.
- In one week in 1985, an artificial neural network developed by Terry Sejnowski and Charles Rosenberg taught itself how to speak 20,000 words.
- In 1997, Garry Kasparov was defeated by IBM's Deep Blue chess program.
- 1999 - An intelligent workstation based on a CAD prototype evaluated 22,000 mammograms and identified cancer 52% more correctly than radiologists.
- Geoffrey Hinton, a computer scientist, coined the phrase "deep learning" in 2006 to refer to the study of neural networks.
- 2012 - A Google unsupervised neural network acquired a 74.8% accuracy rate when trained to identify cats in footage on YouTube.
- 2014 - A chatbot that persuaded 33% of the human judges that it was a Ukrainian kid called Eugene Goostman passed the Turing Test.
- In 2014, Google's AlphaGo, a computer program, beats the world champion human in Go, the most challenging board game.
- June 2016 - DeepMind's artificial intelligence system LipNet successfully recognizes lip-read words in videos with a 93.4% accuracy rate.
- As of 2019, Amazon holds a 70% market share for virtual assistants in the United States.

1.2.2. Types of Machine Learning:

The way in which a prediction-making algorithm learns to improve its accuracy is a common way to classify traditional machine learning.

There are four fundamental strategies:

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement learning

1.2.3. What exactly is it?

A branch of artificial intelligence called "machine learning" developed from the theories of "pattern recognition" and "computational learning." Machine learning is a field of research that allows computers to learn without being explicitly taught, according to Arthur Lee Samuel. Basically, artificial intelligence is a branch of computer science that "learns" from data without the assistance of a human. But this point of view is flawed. A.I. and "Neural Networks that can replicate Human brains (as of now, that is not achievable)", Self 4 Driving Cars, and other things come to mind when the term "machine learning" is used. However, machine learning is much more than that. In the section below, we explore certain aspects of modern computing that are often not expected and some that are Machine Learning is in action.

1.2.4. The ML Expected Contribution:

We'll start with several scenarios in which machine learning could be useful.

1. Speech Recognition (or, in more technical words, Natural Language Processing) - On Windows devices, Cortana may be spoken to. How, though, can it comprehend what you say? The discipline of natural language processing, or N.L.P., emerges. It focuses on the linguistic study of how humans and machines communicate. You guessed it: Machine Learning Algorithms and Systems, including Hidden Markov Models as one example.
2. Computer Vision - A branch of artificial intelligence that focuses on how a machine may (probably) understand the real world In other words, computer vision encompasses all techniques for character, pattern, and facial recognition. Again, the core of computer vision is machine learning, which has a vast variety of algorithms.
3. Google's autonomous vehicle: We may probably guess what motivates it. More wonderful machine learning.

1.2.5. The Unexpected Machine Learning Factor:

Let's look at several locations where the average person would not immediately associate machine learning:

(i). Amazon's Product Suggestions - Have you ever pondered why Amazon always offers a suggestion that simply begs you to open your wallet wider? Recommender Systems are a class of machine learning algorithms that are operating in the background. It gains knowledge about each user's preferences and tailors its recommendations to them.

(ii). YouTube/Netflix - They function the same as before!

(iii). Data mining and Big Data - For many, this might not come as much of a surprise. But analyzing and learning from data on a wider scale only manifests itself as data mining and Big Data. And machine learning is always lurking close when the goal is to extract knowledge from data.

(iv). The stock market, housing finance, and real estate- All employ a lot of machine learning systems, namely "Regression Techniques," to better appraise the market. These techniques range from the poor task of estimating the price of a house to the analysis and prediction of stock market movements.

2. BIG DATA HANDSHAKES MACHINE LEARNING:

Learning a target function (f) that optimally maps input variables (X) to an output variable is how machine learning methods are characterized (Y).

$$Y = f(X)$$

Given fresh samples of the input variables, we would want to predict the future (Y) in this generic learning problem (X). The shape and appearance of the function (f) are unknown. If we did, we wouldn't have to utilize machine learning techniques to learn it from data; instead, we could apply it right away. Learning the mapping Y = f(X) to create predictions of Y for fresh X is the most popular sort of machine learning.

Making the most precise forecasts is the fundamental goal of this process, which is also known as predictive modeling or predictive analytics. Machine learning will make it feasible to extract relevant information from Big Data, even if we use it to store and manipulate large amounts of data. We can extract effective patterns using this machine learning.

The effectiveness of machine-learning algorithms increases as training dataset size increases. Therefore, when Big Data and machine learning are combined, we get twice as much: the algorithms help us keep up with the constant inflow of data, while the amount and diversity of the same data feed the algorithms and foster their development. Let's examine the potential operation of this integration process: We may anticipate specified and evaluated outcomes from feeding massive data to a machine-learning system, such as hidden patterns and analytics that can help with predictive modeling. These algorithms may automate formerly human-centered activities for some businesses. But more often than not, the firm will examine the algorithm's results and look for insightful information that might direct corporate activities. People are once again involved in this scene. While AI and data analytics are powered by computers that perform much better than people, they are limited in their ability to make certain decisions. Many traits that are unique to people—such as critical thinking, intention, and the capacity for holistic approaches—have yet to be replicated by computers. The value of algorithm-generated outcomes declines in the absence of the correct data, and algorithmic ideas may affect business choices in the absence of an expert to understand the output.

3. WHERE AND HOW TO USE MACHINE LEARNING IN BIG DATA

For the collection, analysis, and integration of data, machine learning offers effective and automated solutions. Machine learning ingests efficiency into processing and combines enormous volumes of data independent of its source in cooperation with cloud computing excellence.

Every aspect of a Big Data operation, may use machine learning techniques. such as

- Data Segmentation
- Data Analytics
- Data Labeling
- Personalizing Recommendations
- Predicting Trends
- Exploring Customer Behavior
- Aiding Decision-making
- Decoding patterns
- Simulation

All of these steps work together to extract the big picture from the Big Data, which includes insights and patterns that are then classified and presented in an intelligible way. Machine learning and Big Data are combined in an endless cycle.

As information enters and exits the system, the algorithms developed for specific goals are continuously checked and improved.

4. BIG DATA APPLICATIONS FOR MACHINE LEARNING

Machine learning is employed in many different applications nowadays. The recommendation system that runs Facebook's News Feed is perhaps one of the most well-known applications of machine learning. Facebook employs machine learning to individually tailor each user's feed.

The recommendation engine will start to display more of that group's activity sooner in the feed if a member regularly pauses to read the posts in that group. The engine is working behind the scenes to reinforce recognized patterns in the member's online activity. Machine learning, or ML, is simply the ability of a machine to learn. Therefore, based on your prior searches, the Amazon or Facebook website code will have discovered the kinds of goods you are interested in. As a result, the website will place such things in front of you as soon as you check in. And it could even pair them with a tempting offer. Such tactics aim to encourage impulsive internet purchasing. According to what I've heard, Amazon is creating or has already deployed a system that allows them to estimate the demand for specific commodities in a given location and manage the stockpiles of such things.

Today's technology has made machine learning a buzzword, and it is developing extremely quickly. Without realizing it, we use machine learning every day in applications like Google Maps, Google Assistant, Alexa, etc. The following list of the top machine learning real-world applications includes:

- Image Recognition
- Speech Recognition
- Traffic prediction
- Product recommendations
- Self-driving cars
- Email Spam and Malware Filtering
- Virtual Personal Assistant
- Online Fraud Detection

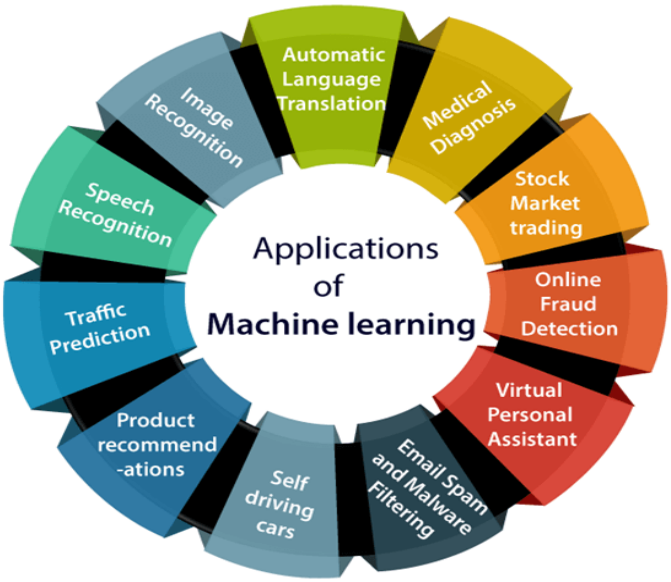

Fig. 1: Applications of Machine learning
(Pic credit- https://www.javatpoint.com/applications-of-machine-learning)

5. MACHINE LEARNING MODELS

5.1 Parametric & Non-Parametric Machine Learning Algorithms

5.1.1. KNN Algorithm: KNN is a nonparametric supervised learning technique that uses training sets to segment data points into given categories. In simple classifications, the word collects information from all educational cases and similarities based on the new case. Look at the training for the most similar (neighbor) K cases and predict the new instance (x) by summarizing the output variables for these K cases. Classification is the class value mode (or most commonly). A flow diagram of the KNN algorithm is shown in Figure 2.

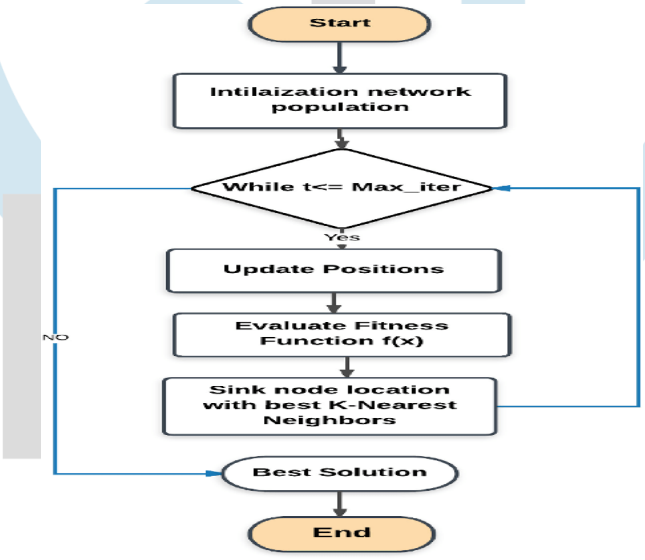

Fig.2: Flow chart for KNN algorithm

5.1.2. Support Vector Machines: SVM divides the given data into decision surface. Decision surface is further dividing the data into hyper plane of two classes. Training points defines the supporting vector which defines the hyper plane. Probably, a hyper plane with the greatest distance to the closest learning data point typically has better margins and larger errors because of the larger margins, the generalization of classifiers is weak. The flow chart for SVM is given in the figure 4, it shows the steps involved in SVM algorithm.

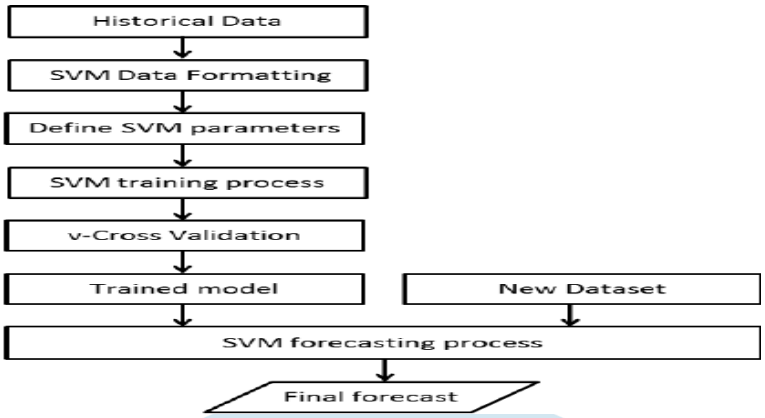

Fig.3: Flow chart for SVM algorithm

5.1.3. Decision Tree: Decision tree is a predictive model which works by checking condition at every level of the tree and proceeds towards bottom of the tree where various decisions are listed. The condition depends on the application and the outcome might be in terms of decision. There are various types of Decision tree algorithms such as C4.5, CART and ID3 algorithm.

The ID3 algorithm starts with the original set S as the root node. Each iteration of the algorithm finds each unused attribute in set S and computes the entropy H (S) (or information gain IG (S)) of this attribute. Then select the attribute with the smallest entropy value (or the largest information gain). The set S is then divided into the selected attributes. The algorithm is continually reproduced in each subset, considering only those attributes that were not previously selected. The figure 4 shows the flow diagram of the ID3 algorithm.

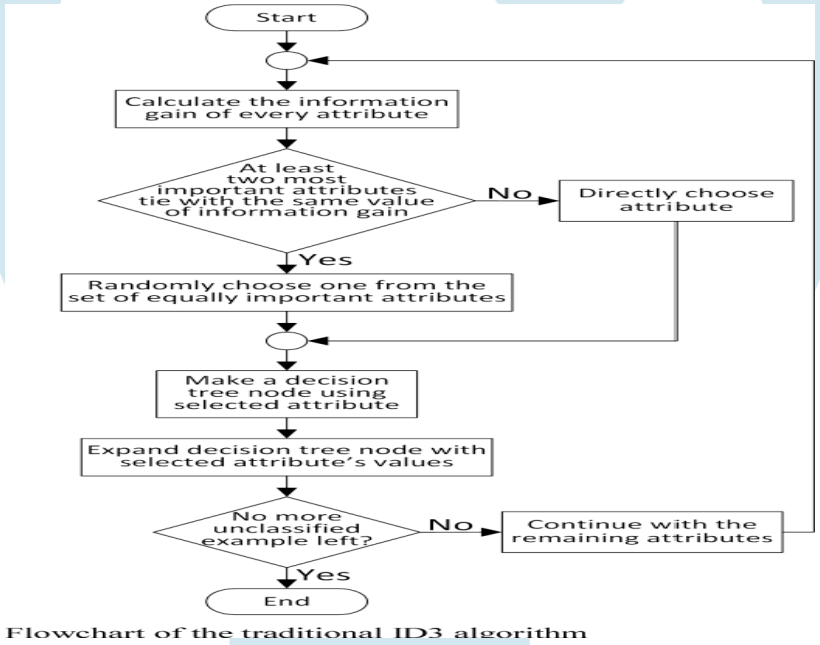

Fig.4: Flow chart for ID3 algorithm

5.1.4. Naive Bayes Classifier: Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. This is also known as Maximum A Posteriori (MAP).

Naive Bayes classifier assumes that all the features are unrelated to each other. Presence or absence of a feature does not influence the presence or absence of any other feature. We can use Wikipedia example for explaining the logic i.e., A fruit may be considered to be an apple if it is red, round, and about 4″ in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

In real datasets, we test a hypothesis given multiple evidence(feature). So, calculations become complicated. To simplify the work, the feature independence approach is used to 'uncouple' multiple evidence and treat each as an independent one.

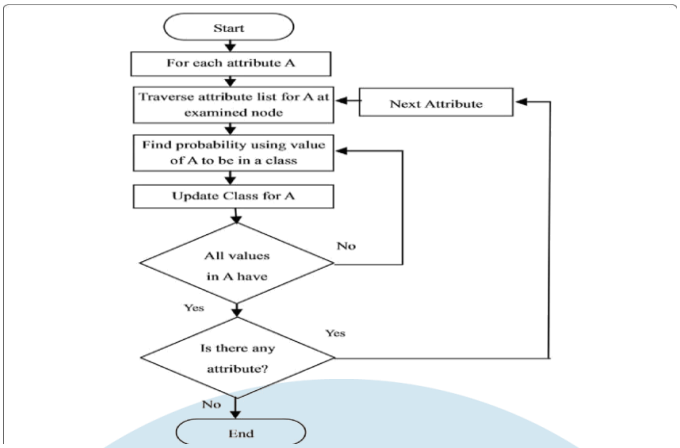

Fig.5: Flow chart for Naive Bayes algorithm

5.1.5. The Random Forest Classifier: Random Forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction (see figure below).

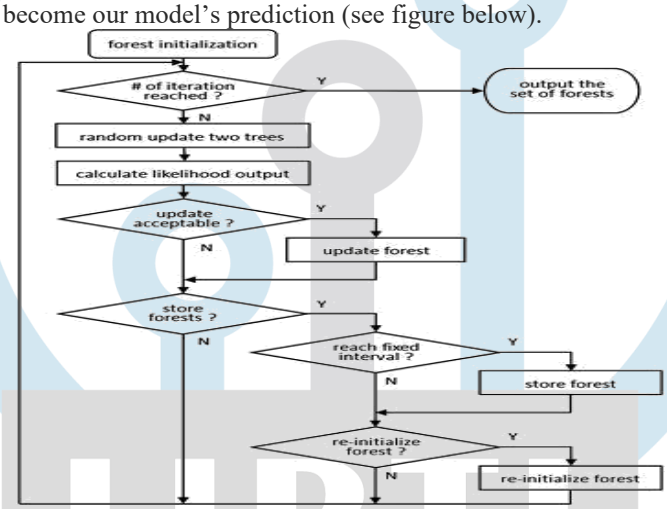

Fig.6: Flow chart for Random Forest Classifier

5.1.6. Decision Tree*:* The **decision tree** method is a powerful and popular predictive machine learning technique that is used for both classification and regression. So, it is also known as **Classification and Regression Trees** (**CART**). The algorithm of decision tree models works by repeatedly partitioning the data into multiple sub-spaces, so that the outcomes in each final sub-space is as homogeneous as possible. This approach is technically called recursive partitioning.
The produced result consists of a set of rules used for predicting the outcome variable, which can be either:

- a continuous variable, for regression trees
- a categorical variable, for classification trees

The decision rules generated by the CART predictive model are generally visualized as a binary tree.

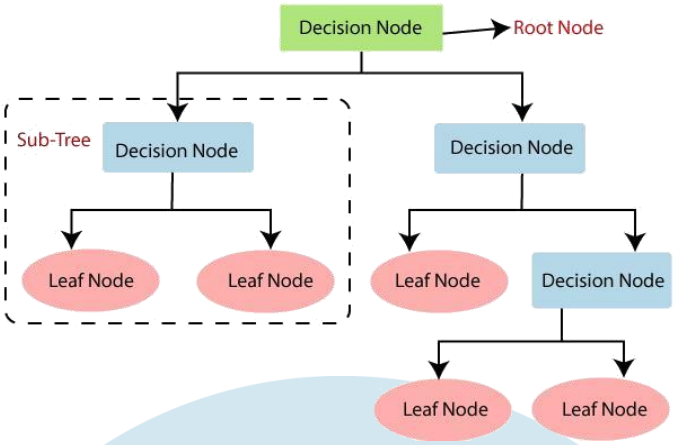

Fig.7: Flowchart of Decision Tree

5.1.7. Linear Discriminant Analysis: Linear Discriminant Analysis or LDA is a dimensionality reduction technique. It is used as a pre-processing step in Machine Learning and applications of pattern classification. The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs. Discriminant analysis is used to predict the probability of belonging to a given class (or category) based on one or multiple predictor variables. It works with continuous and/or categorical predictor variables.
Previously, we have described the logistic regression for two-class classification problems, that is when the outcome variable has two possible values (0/1, no/yes, negative/positive).
Compared to logistic regression, the discriminant analysis is more suitable for predicting the category of an observation in the situation where the outcome variable

contains more than two classes. Additionally, it's more stable than the logistic regression for multi-class classification problems.
Note that, both logistic regression and discriminant analysis can be used for binary classification tasks.
The following discriminant analysis methods will be described:

- **Linear discriminant analysis** (**LDA**): Uses linear combinations of predictors to predict the class of a given observation. Assumes that the predictor variables (p) are normally distributed and the classes have identical variances (for univariate analysis, p = 1) or identical covariance matrices (for multivariate analysis, p > 1).

- **Quadratic discriminant analysis** (**QDA**): More flexible than LDA. Here, there is no assumption that the covariance matrix of classes is the same.

- **Mixture discriminant analysis** (**MDA**): Each class is assumed to be a Gaussian mixture of subclasses.

- **Flexible Discriminant Analysis** (**FDA**): Non-linear combinations of predictors is used such as splines.

- **Regularized discriminant analysis** (**RDA**): Regularization (or shrinkage) improves the estimate of the covariance matrices in situations where the number of predictors is larger than the number of samples in the training data. This leads to an improvement of the discriminant analysis.

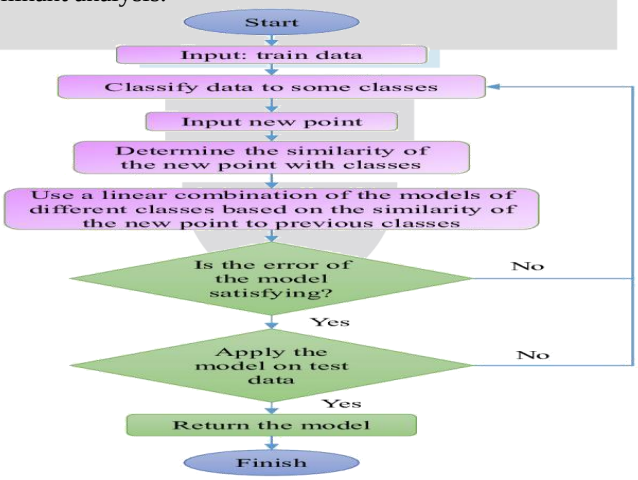

Fig.8: Flowchart of LDA

Once the model has been trained efficiently it is tested on the Testing dataset which is different from the Training data in sample values.

6. STEPS OF MODEL IMPLEMENTATION

We are developing real-time model to predict the crop which has to be sown by the farmer.

- The first step involves analysis on historical data to create a machine learning model
- The second phase use machine learning models to make crop predictions for the farmer.

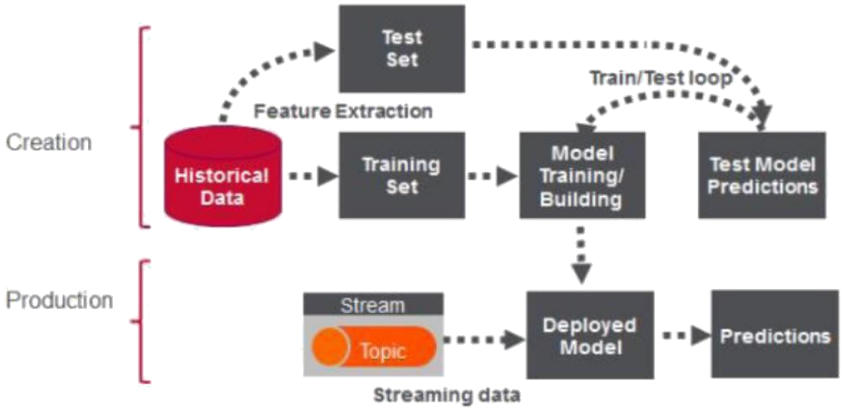

Fig.9: Evaluating System

Classification can also be a machine learning paradigm that involves obtaining a function that will separate the data into categories, or classes, with a training set and testing set of observations. This function is then employed to identify which model to deploy for crop prediction.

7. PROGRAMMIMNG LANGUAGE FOR MODEL CREATION

7.1. INTRODCUTION OF R PROGRAMMING LANGUAGE [5]

R is an open-source programming language that is widely used as a statistical software and data analysis tool. R generally comes with the Command-line interface. R is available across widely used platforms like Windows, Linux, and macOS. Also, the R programming language is the latest cutting-edge tool.

It was designed by **Ross Ihaka and Robert Gentleman** at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R programming language is an implementation of the S programming language. It also combines with lexical scoping semantics inspired by Scheme. Moreover, the project conceives in 1992, with an initial version released in 1995 and a stable beta version in 2000.

7.1.1. Why R Programming Language?

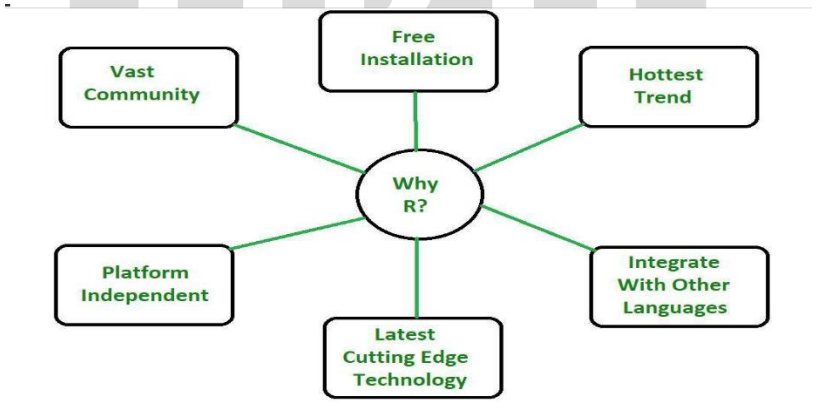

Fig10: Why R

R programming is used as a leading tool for machine learning, statistics, and data analysis. Objects, functions, and packages can easily be created by R. It's a platform-independent language. This means it can be applied to all operating system.

It's an open-source free language. That means anyone can install it in any organization without purchasing a license.

7.1.2. Features of R Programming Language

7.3.1. Statistical Features of R:

- **Basic Statistics:** The most common basic statistics terms are the mean, mode, and median. These are all known as "Measures of Central Tendency." So, using the R language we can measure central tendency very easily.
- **Static graphics:** R is rich with facilities for creating and developing interesting static graphics. R contains

functionality for many plot types including graphic maps, mosaic plots, biplots, and the list goes on.

- **Probability distributions:** Probability distributions play a vital role in statistics and by using R, we can easily handle various types of probability distribution such as Binomial Distribution, Normal Distribution, Chi-squared Distribution and many more.

7.1.3. Programming Features of R:

- **R Packages:** One of the major features of R is it has a wide availability of libraries. R has CRAN (Comprehensive R Archive Network), which is a repository holding more than 10, 0000 packages.
- **Distributed Computing:** Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Two new packages **ddR and multidplyr** used for distributed programming in R were released in November 2015.

7.1.4. Programming in R:

Since R is much similar to other widely used languages syntactically, it is easier to code and learn in R. Programs can be written in R in any of the widely used IDE like **R Studio, Rattle, Tinn-R**, etc.

7.1.5. Advantages of R:

- R is the most comprehensive statistical analysis package. As new technology and concepts often appear first in R.
- As R programming language is an open source. Thus, we can run R anywhere and at any time.
- R programming language is suitable for GNU/Linux and Windows operating system.
- R programming is cross-platform which runs on any operating system.
- In R, everyone is welcome to provide new packages, bug fixes, and code enhancements.

7.1.6. Disadvantages of R:

- In the R programming language, the standard of some packages is less than perfect.
- Although, R commands give little pressure to memory management. So, R programming language may consume all available memory.
- In R basically, nobody to complain if something doesn't work.

7..1.7. Applications of R:

- We use R for Data Science. It gives us a broad variety of libraries related to statistics. It also provides the environment for statistical computing and design.
- R is used by many quantitative analysts as its programming tool. Thus, it helps in data importing and cleaning.
- R is the most prevalent language. So many data analysts and research programmers use it. Hence, it is used as a fundamental tool for finance.
- Tech giants like Google, Facebook, bing, Accenture, Wipro and many more using R nowadays.

7.2. INTRODCUTION OF PYTHON PROGRAMMING LANGUAGE

Python is a high-level, general-purpose, interpreted programming language that primarily promotes code readability. Professional programmers and developers from a range of industries, including Web Development and Machine Learning, heavily utilize it. Python has its own set of benefits and drawbacks, just like any other programming language you must have heard of, read about, or maybe used for a variety of purposes. [7]

In 1989, Guido Rossum developed the object-oriented programming language Python. It is perfectly suited for the quick prototyping of sophisticated applications. It is extendable to C or C++ and offers interfaces to numerous OS system functions and libraries. The Python programming language is used by several huge corporations, such as NASA, Google, YouTube, BitTorrent, etc. [6]

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991 [8].

7.2.1. ADVANTAGES OF PYTHON [9]

7.2.1.1. Simple to use and easy to understand

Since the grammar of the Python programming language is similar to that of the English language, anyone may read and understand its codes with ease. This is a language that is simple to learn and easy to pick up. This is only one of the advantages Python has over other programming languages like C, C++, or Java. Additionally, Python uses comparatively less lines of code than other programming languages with bigger code blocks to accomplish the same operations and tasks.

7.2.1.2. Increases output

Another advantage of Python is that it is a very productive language. Additionally, Python programmers may simply concentrate on problem-solving because to Python's simplicity.

7.2.1.3. Translated Language

Python can run the code directly, one line at a time, as it is an interpreted language. Additionally, if there is a mistake, it indicates the error that happened rather than continue with the execution.

7.2.1.4. Open-source

Python can be easily distributed since it is offered to users for free and has an open-source license. Python allows you to download any source code, edit it, and then share your customized version with others. This function might be useful, especially if you want to reuse code and save time while creating original apps. Additionally, you may modify its functionality and utilize an earlier code version for development.

7.2.1.5. Easy to Transport

Most programming languages require changes to the code in order to run a particular program on several platforms, including the ones you have learnt to read, write, and use like C, C++, etc. It's not the case with Python programming, though.

7.2.1.6. Large-scale libraries

The extensive standard library of Python contains all the required functions you need for any given activity. Python is now independent of outside libraries thanks to this. However, if you do want to utilize any external libraries, you may quickly import a number of packages from the vast Python Package Index (PyPi), which contains more than 200,000 packages, using the Python package manager (pip).

7.2.1.7. Integration with Other Programming Languages is Simple

Another distinguishing characteristic of Python is that it enables cross-platform development by integrating not just with libraries like Jython and Cython but also with other programming languages like Java, C, and C++. As a result, Python is stronger in comparison. There is no perfect programming language, hence it is typically not a good idea to use only one.

7.2.2. Disadvantages of Python

Despite the fact that Python's benefits much outweigh its drawbacks, there are a few drawbacks you should be aware of.

7.2.2.1. Slow Speed

Unfortunately, low speed strengths can occasionally result in disadvantages. Here is an example of one. Python is an interpreted language with dynamic typing, which contributes to its sluggish performance by requiring line-by-line execution of the code. Python's slow execution speed is mostly caused by its dynamic nature, which necessitates some additional work. This is among the causes for not using Python when a program's speed is an important consideration.

7.2.2.2. Inefficient use of memory

Python must make certain compromises in order to provide developers and programmers
with some simplicity. A drawback of this language is that it consumes a lot of memory, which is problematic when designing an app that prioritizes for memory optimization.

7.2.2.3. Lack of expertise in mobile device programming

Instead of utilizing it for client-side or mobile apps, developers often employ Python for server-side development. This is due to Python's poor memory efficiency and sluggish processing speed when compared to other programming languages.

7.2.2.4. Database Layer with a difficult user interface

Python makes program creation remarkably stress-free and simple, but it falls short when it comes to database interaction. Compared to more widely used technologies like ODBC and JDBC, it features a basic and immature database layer. Most businesses want to connect with complicated data smoothly, which makes using Python challenging for them.

7.2.2.5. Known to Result in Runtime Errors

The data type of a variable can be changed at any moment because to Python's dynamic capability. A Python variable that originally contained an integer may now contain a string. The result might be runtime errors. As a result, testing must be done numerous times for each application built by developers.

8. CONCLUSION:

A new era of Big Data has begun. As people develop new techniques for manipulating and processing data, the quantity of information and knowledge that can be gleaned from the digital cosmos keeps growing. Big Data lacks a clear definition and a specified structure. The Big Data industry is still developing. There are numerous difficulties in large data analysis, but the study is still in its early stages. This study, which is the result of a joint research effort, provides new Big Data handling technologies and approaches while beginning to examine the conventional view of data analytics. We've found some significant problems with Big Data that Big Data consumers in particular must deal with. If we are to benefit from the advantages of Big Data, we must encourage and support basic research aimed at solving these technological problems. Our upcoming study will focus on creating a more thorough grasp of Big Data difficulties.

Machine learning (ML) is essential to meet the challenges posed by Big Data and expose hidden patterns, facts, and knowledge from large amounts of data with the aim of transforming the latter's capabilities into real motivation for business core leadership and logical research. Making machine learning more declarative will be a key component of machine learning analytics in the future, since this will make it simpler for non-experts to describe and interact with various types of data in various streams. In the future, we will improve and evaluate how well machine learning approaches function for various sorts of issues. Extending machine learning techniques to Big Data, which are effective and extremely scalable in the way they analyze vast volumes of data, is one potential route.

9. SAMPLE CODING IN R

```
# Install the package DataExplorer ----------------------------------------
# Install the package tidyverse -------------------------------------------
# Install the package skimr -----------------------------------------------
# Install the package flextable -------------------------------------------
```

```
# Load the library DataExplore----------------------------------------
library(DataExplorer)
# Load the library tidyverse-----------------------------------------
library(tidyverse)
# Load the library skimr ---------------------------------------------
library(skimr)
# Load the library flextable -----------------------------------------
library(flextable)
# loading or reading the dataset for crop recommendation ----------
dataset_crop_n <- read.csv('Crop_Recomm_n.csv')
#To see every column in a data frame.
glimpse(dataset_crop_n)
output

  Rows: 2,200                                   Columns: 8

$          <int> 90, 85, 60, 74, 78, 69, 69, 94, 89, 68, 91,
I          90, 78, 93, 94, 60, 85, 91, 77~
$          <int> 42, 58, 55, 35, 42, 37, 55, 53, 54, 58,
I          53, 46, 58, 56, 50, 48, 38, 35, 38~
$          <int> 43, 41, 44, 40, 42, 42, 38, 40, 38, 38,
I          40, 42, 44, 36, 37, 39, 41, 39, 36~

   $ temperature <dbl> 20.87974, 21.77046, 23.00446, 26.49110,
20.13017, 23.05805, 22.70884, 20.2~ $ humidity <dbl> 82.00274,
80.31964, 82.32076, 80.15836, 81.60487, 83.37012, 82.63941, 82.8~ $ ph
<dbl> 6.502985, 7.038096, 7.840207, 6.980401, 7.628473, 7.073454,
5.700806, 5.71~ $ rainfall <dbl> 202.9355, 226.6555, 263.9642, 242.8640,
262.7173, 251.0550, 271.3249, 241.~ $ label <chr> "rice", "rice", "rice",
"rice", "rice", "rice", "rice", "rice", "rice", "r~
```

10. SAMPLE CODING IN PYTHON [10]

(To get the value of Pi to n number of decimal places)

```
#!/usr/bin/env python3
# https://github.com/MrBlaise/learnpython/blob/master/Numbers/pi.py
# Find PI to the Nth Digit
# Have the user enter a number 'n'
# and print out PI to the 'n'th digit

def calcPi(limit):  # Generator function
    """
    Prints out the digits of PI
    until it reaches the given limit
    """

    q, r, t, k, n, l = 1, 0, 1, 1, 3, 3

    decimal = limit
    counter = 0

    while counter != decimal + 1:
        if 4 * q + r - t < n * t:
            # yield digit
            yield n
            # insert period after first digit
            if counter == 0:
                yield '.'
            # end
            if decimal == counter:
                print('')
                break
            counter += 1
            nr = 10 * (r - n * t)
```

```
            n = ((10 * (3 * q + r)) // t) - 10 * n
            q *= 10
            r = nr
        else:
            nr = (2 * q + r) * l
            nn = (q * (7 * k) + 2 + (r * l)) // (t * l)
            q *= k
            t *= l
            l += 2
            k += 1
            n = nn
            r = nr

def main():  # Wrapper function

    # Calls CalcPi with the given limit
    pi_digits = calcPi(int(input(
        "Enter the number of decimals to calculate to: ")))

    i = 0

    # Prints the output of calcPi generator function
    # Inserts a newline after every 40th number
    for d in pi_digits:
        print(d, end='')
        i += 1
        if i == 40:
            print("")
            i = 0

if __name__ == '__main__':
    main()
```

REFERENCES:

[1]. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, pp. 255-260, 2015.
[2]. Hey, T., Tansley, S., Tolle, K., editors (2009). The Fourth Paradigm: Data- Intensive Scientific Discovery. Microsoft Research. S.Hasan et a.l
[3]. L. Rabhi, N. Falih, A. Afraites, and B. Bouikhalene, “Big Data Approach and its applications in Various Fields: Review,” Procedia Comput. Sci., vol. 155, pp. 599–605, 2019, doi: 10.1016/j.procs.2019.08.084.
[4]. H. Cadavid, W. Garzón, A. Pérez, G. López, C. Mendivelso, and C. Ramírez, “Towards a smart farming platform: From IOT-based crop sensing to data analytics,” Communications in Computer and Information Science, vol. 885, pp. 237–251, 2018, doi: 10.1007/978-3-319-98998-3_19.
[5]. https://www.geeksforgeeks.org/r-programming-language-introduction [August, 2021]
[6]. https://www.guru99.com/python-tutorials.html
[7]. https://www.javatpoint.com/advantages-of-python-that-made-it-so-popular-and-its-major-applications
[8]. H.-P. Halvorsen, “Technology blog - https://en.wikipedia.org/wiki/Python(programminglanguage), 00 20
[9]. https://intellipaat.com/blog/advantages-and-disadvantages-of-python/
[10]. https://www.w3resource.com/projects/python/python-projects-numbers.php