International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Comprehensive Study of Data Science

Chennu Ganesh Tiru Venkata Manikanta Sai, Kasaraneni Sri Harsha, Chennu Aryan Karthikeya

Abstract: Today's generation is totally dependent on technology that uses data as its fuel. The present study is all about innovations and developments in data science and gives ideas about how to efficiently use the data provided. This study will help to understand the core concepts of data science. The concept of artificial intelligence was introduced by Alan Turing in which the main principle was to create an artificial system which can run independently of human given programs and can function with the help of analyzing data to understand the requirements of the users. Data science comprises business understanding, analyzing data, ethical concerns, understanding programming languages, various fields and sources of data, skills, etc. The usage of data science has evolved over the years. In this review article, we have covered a part of data science, i.e., machine learning. Machine learning uses data science for its work. Machines learn through their experience, which helps them to do any work more efficiently. This article includes a comparative study image between human understanding and machine understanding, advantages, applications, and real-time examples of machine learning. Data science is an important game changer in the life of human beings. Since the advent of data science, we have found its benefits and how it leads to a better understanding of people, and how it cherishes individual needs. It has improved business strategies, services provided by them, forecasting, the ability to attend sustainable developments, etc. This study also focuses on a better understanding of data science which will help us to create a better world.

1. Introduction

Data science is the study of data to derive useful perceptions or intuition for various business purposes. It can also be defined as a combination of programming knowledge, mathematics and algorithms. The processes like data mining, cleansing, analyzing, visualization and generation of insights are included in data science. This process helps us to solve analytical problems that are complicated for human beings to solve. (Shah,2022) (Kips,2020)

2. Timeline of data science

•1962 - John Turkey published a book 'The Future of Data Analysis'.

•1974 - Peter Naur termed 'Data Science' in 'Concise Survey of computer methods.'

•1977 - IASC (International Association for Statistical Computing) was formed.

- John Turkey published another book about data analysis named 'Exploratory Data Analyzing'.

•1985 - Data science was used as an alternative term for statistics by C.F. Jeff Wu in the Chinese Academy of Science. •1989 - The first Knowledge Discovery and Databases (KDD) workshop was organized by Gregory Piatetsky-Shapiro. •1994 - Business Week magazine introduced the concept of database marketing.

•1999 - In 'Mining Data for Nuggets of Knowledge' Jacob Zahavi wrote about the importance and need to store and analyze data for businesses.

•2001 - For the initial cloud-based application Software-as-a-Service (SaaS) was formed.

•2002 - The Data Science Journal was published by the International Council for Science (ICS).

•2006 - For storage and processing of large data sets Apache Hadoop 0.1.0 an application has been released.
•2007- The research center for Dataology and Data Science was established at Fudan University.
•2009 - In this year the university established in 2007's center's researchers published "Introduction to Dataology and Data Science," in which they explained the various important topics of Dataology and Data Science.

- "Harnessing the Power of Digital Data for Science and Society" is published.

- Troy Sadkowsy created the data scientist group on Linkedin as a companion to his website, data scientist.com.

•2010 - Kenneth Cukier writes in the Economist special Report" Data Data Everywhere", which combines all the skills to find out the hidden mountain of data.

•2015 - Artificial Intelligence, Machine learning and deep learning make their debut in the field of data science. These technologies have been helping us to bring a new change in our lives.

•2018 - Introduction to new regulations took place this year in the field of Data Science.

•2020s - We are seeing the further evolution of Artificial Intelligence systems, machine learning and deep learning. **3. Expertise in data science**

For being involved in business or anyone related to data science one must have knowledge of various programming languages such as Java, C/C++, Python, R, SQL, etc. One must also know how to derive useful meaning from the given piece of data and one should also know how to design, model and analyze data. (Kips,2020)

3.1 Prerequisites

- Programming
- Statistics
- Machine learning
- Databases
- Modeling

4. Fields in data science

A few professionals involved in the field of data science are Machine learning Engineers, Data Architects, Data Scientists, Data analysts, Software engineers, etc. (Kips,2020)

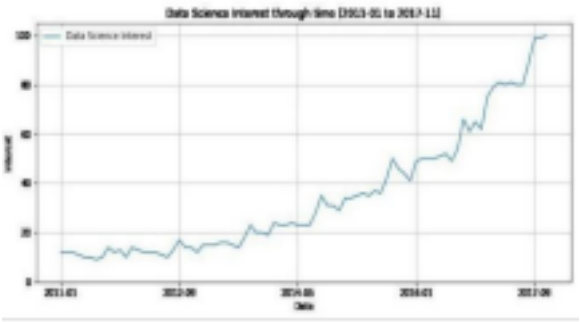

fig.1. interest in data science over the years (Griffin, 2019)

5. Types of Data Scientist

5.1 Machine Learning Experts

The advancement in technology over the years has increased the intellectual capabilities and decision-making capabilities of modern-day computers.

Machine Learning professionals are responsible for creating algorithms and delivering outcomes by deriving patterns from big data inputs and past trends (Aasim, 2020)

5.2 Actuarial Scientist

Actuarial Scientists use mathematical and statistical models to assess risk in the financial sector. Apart from the above-mentioned skills, knowledge about BFSI (Banking Finance Services and Insurance) is mandatory. They predict the financial prospects, such as future income, revenue, and profit/losses of an uncertain event in banks or insurance companies. A prior experience in the finance industry is a must to become an actuarial scientist. (Aasim, 2020)

5.3 Business Analysts

The business analyst helps boost a business's strength by making data-driven decisions. They use data to develop insights and suggest changes in organizations. They are liable for analyzing data using various tools such as SQL and Excel, compiling charts and graphs for data visualization, understanding business goals, and providing solutions based on past knowledge (Aasim, 2020)

5.4 Data warehousing Expert

Data warehousing is a way of managing data. Experts in this field are involved in structuring and implementing data warehouse designs. These warehouses store a large amount of information consolidated from multiple sources. (Kakkar, 2020)

5.5 Data Engineer

Data engineers build systems that synthesize data for performing tasks and make predictions based on it. They collect the raw data from the data warehouses and convert it into information that can be interpreted by analysts. Data architecture designing is one of the fundamental roles performed by them. (Aasim, 2020)

5.6 Data Mining Expert

A Data Mining expert looks into a large set of data and finds patterns in them using available software and tools. In other words, they recognize the hidden trends in the data using complex mathematical algorithms. Using these trends, they try to form assumptions about the data and learn about anomalies, and if there are any, this facilitates applying the correct statistical method for validating predictions about an event. (Kakkar, 2020)

5.7 Data Visualization Expert

A data visualization expert takes in the raw data and presents it in the form of graphs and charts for interpretation. This way of presentation serves as an excellent way of making the non-technical audience familiar with the technicalities of the data. Tableau and PoweBI are some of the most widely used data visualization tools. Python consists of various libraries which help in data visualization purposes. They help eliminate outliers and give the exact idea of how scattered our data is. (Kakkar, 2020)

5.8 Data Analyst

A data analyst is responsible for gathering the data and interpreting it to resolve a particular problem. They turn the information generated from raw data into meaningful insights and convey them to the stakeholders. They compile their findings into reports which help in understanding customer needs. Their job can be divided into five parts: collecting the data, cleaning the data, modeling the data, interpreting the data, and presenting the finding of the data. (Kakkar, 2020)

5.9 Cybersecurity Data Scientist

Cybersecurity Data Scientists help in the prevention of fraudulent activities. They develop data science models trained on past data to predict the likelihood of an intrusion or attack. This field of data science involves developing algorithms to deduce patterns from previous attacks and beforehand warning about the reliability of the system in use. (Kakkar, 2020)

6. Steps in the execution of data science

6.1 Business understanding

It's very important to focus on these 10 main concepts to understand the core of business i.e. value creation, marketing, sales, value delivery, finance, the human mind, working with yourself, working with others, understanding and analyzing systems and improving systems. (KAUFMAN,2010)

6.2 Data mining

Data mining is a process to extract useful hidden concepts from databases. (Dunham, 2002) It is also known as Knowledge Discovery in Databases (KDD). (Olaiya & Adeyemo, 2012) Data mining is a process of analyzing data to find patterns , discover new trends like you tube trends etc., and gain insight into how that data can be used. Data miners can then use those findings to make or predict an outcome . Data mining is mainly used in the field of machine learning, in the field of statistics and artificial intelligence.

6.3 Data cleaning

It's very essential to fresh /warehouse data so that the information already present in the software works efficiently. (Bhatia, 2009) **6.4 Data Exploration**

Data exploration is a technique that is used for extracting the significant pattern in the given data.
(Kips, 2020)

6.5 Feature engineering

There are several features of a particular problem to find the solution to the given complex problem. These features play an important role to get out of this complexity. (Rawat & Khemchandani, 2017)

6.6 Projector or Data modeling

We don't always need machines to protect the upcoming situation. We ourselves can do it. All we need for this is a bit of knowledge and a bit of management to protect the upcoming future. (Kips,2020)

6.7 Data visualization

It is a better technique for understanding the perception of the data. (Kips, 2020)

7. Significance of data science

● Data science helps companies to analyze big data to derive multiple sources from it. A few applications of big data are as follows-

● It is widely used in the industries of healthcare, textiles, finance, banking, etc.

● It helps the company to analyze the needs of the customers.

● It is used in gaming, live streams and many more live sessions.

● It can be used in speech recognition and face recognition.

● It can also be used for the detection of fraud to save data.

● It is also used for logistics purposes.

(Kips,2020)

8. Importance of data science

● The crucial role of data science is the value of soaring heights with emerging technologies. Data science helps big organizations with the ability of processing and interpreting large volumes of data. This enables organizations, businesses or companies to make informed decisions about growth. It's a fact that Southwest Airlines at one point of time was able to save $100 million by influencing data. They reduced their plane's idle time and drove a change in utilizing their resources.

● Analyzing big data helps companies understand their customers and improve their experiences by learning from previously recorded data. Data science also helps in visualizing data that is understandable for business stakeholders to build their projects.

● In 2003, iTunes took 100 months to reach 100 million users while for Pokemon Go in 2016, it took days to reach the same 100 million mark. (Bansal, 2021) User outreach timelines are changing from the old models of marketing and promotions. The evolution happened due to the massive digitization of promotion platforms that run on data insights.

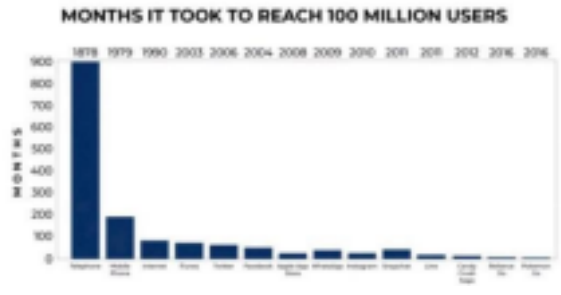

fig.2. growth of data science (Bansal, 2021)

There is also another importance of data science for industry verticals in influencing the power of data. Data science is significant for businesses as it has been disclosing wonderful solutions and intelligent decisions across many industries. Deloitte Access Economics report suggests that 76% of businesses will be pumping up their data analytics spending. (Bansal, 2021)

9. Developments in the field of data science

The emerging and developing data science as a field of study and practical application has led to the development of technologies such as machine learning, natural language processing, deep learning and computer vision.

9.1 Data-Driven Customer Experience

Companies and industries take and record our data and use it to provide us with increasingly valuable or enjoyable experiences. Our interactions with businesses are becoming more and more digital- from AI chatbots to Amazon's cashier-less convenience store. Each and every activity can be recorded, measured and analyzed and processed into a smooth and enjoyable experience. (Marr, 2021)

9.2 Deep Fakes, Generative AI and Synthetic data

Once Tom Cruise started posting some videos on Tik Tok when realistic 'deep fake' videos went viral. This new emerging technology is known as Generative AI, as it aims to generate/create something new. In this case, Tom Cruise entertained us with tales of meeting Mikhail Gorbachev which doesn't exist in reality. This technology of Generative AI is emerging quickly and getting embedded in the arts and entertainment industries.

Creating synthetic data for the training of machine learning algorithms is considered to possess huge potential. Synthetic or simply artificial faces of people who never existed can be created to train facial recognition systems. By this, the privacy or concerns involved in using real faces can be avoided. (Marr, 2021)

9.3 Smart Learning Analytics

The role of data science in the development of learning analytics changed remarkably over the few decades. Data science provided tools for designing experiments when it included only a handful of people before the Digital Age. Big data at hand have millions of topics and details about their content, data science is used to form a basis for analysis. The emergence of Smart Learning Environments has made learning analytics easier. (Giannakos, Sampson, & Kidzinski, 2016)

10. Ethical concerns of data science

Something which has advantages always comes with some disadvantages or concerns. When we talk about data science for its migration, questions related to ethics arise from the different types of technologies that can be developed from the intersection of these two. Types of ethics in data science are

•Ethics of data

•Ethics of algorithms

•Ethics of practice

The most common ethical issue in data science with emerging technologies is consent. Consent is a practice to ensure a person's self-governance is protected. Consent has an important role in a person's involvement and sharing of identifiable personal data (Salah, Canca, & Erman). Other ethical issues are privacy, the accuracy of the information, accountability, and inclusivity (Ronzhyn & Wimmer, 2019).

Today most machines are trained on the data provided to them. This means the machines are biased. For example, a machine trained on data might come to a conclusion that more percentage of successful businesses established by specific gender are better at handling certain jobs (gender bias). Detecting bias is not easy as many systems act as a black box. Whatever happens inside the black box remains in it forever and it cannot be explained by the developer also. Human beings are capable of finding negative uses of every technology. If the data collected by

systems or data scientists get into the wrong hands it may cause a lot of damage. In the hands of terrorists, it can become a tool for spreading terror, in the hands of cyber criminals it can be easy for hacking or damaging, etc. (Kips Content Development Team, 2020)

The speed at which big organizations access huge volumes of data or big data through their users prompts questions on ethics to arise. Collecting a lot of information can affect people's behaviour, raising questions on the violation of privacy,legal rights,etc. big organizations should survey user's reactions to the usage of their data by policy-making, research, etc. ·(Nair, 2020)

11. Methods used for analyzing data

11.1 Cluster Analyzing

It is a process of grouping similar data into clusters based on observation. It is an unsupervised data classification procedure. **11.1.1 Hierarchical Cluster Analyzing**

In this technique, a tree diagram (Phylogenetic tree) of the data cluster is made and has been observed with the help of a dendrogram. Algorithms like agglomerative (grouping observations) or divisive (dividing the data sets) are applied. Measurement of various variables plays an important role in this technique. These variables are compared between subjects and the Euclidean distance is adjusted.

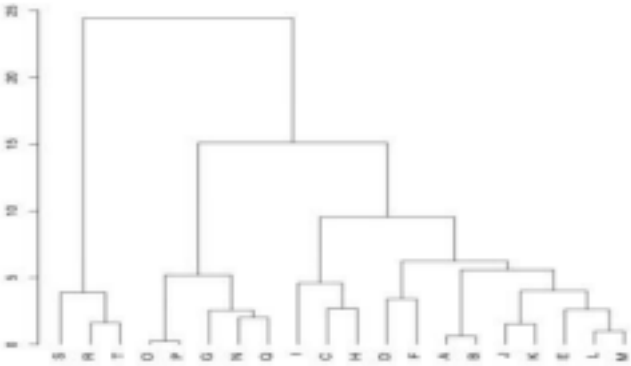

fig.3. hierarchical cluster analysis (Desagulier, 2019)

11.1.2 Centroid-Based Clustering

Each cluster is referred to as a single central entity which is not necessarily a part of the data set in this method. **11.1.3 Distribution-Based Clustering**

It creates groups of data points based on probability distribution. Some important and complex properties like correlation and dependence can be captured by this method.

11.1.4 Density-Based Clustering

Considering a group as a contiguous region of high point density, distinctive clusters in data are identified in a data space by separating other clusters belonging from sparse regions. Data points in sparse regions are known as outliers or noise.

11.1.4.1 DBSCAN Clustering

It stands for Density Based Spatial Clustering of Applications with Noise. In 1996 Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu proposed it. The data points are divided into densely grouped and sparsely grouped data points.

11.2 K-Nearest Neighbour

Also termed KNN, it is a supervised data classification procedure. It is based on the estimation that a data point has a likelihood that it would be a member of any group nearest to it. (Sinharay,2010).

11.3 Cohort Analyzing

The procedure of classifying data into groups is called cohorts based on common ground and these groups are defined by a fixed period. Determining the objective of the analysis, carving the metrics which associate with the objective, determining the necessity of the cohorts, analyzing raw data and recording the data are the major steps involved.

fig.4. cohort analyzing (Yuan, 2019)

11.3.1 Time-Based Cohort Analyzing - The classification of cohorts is based on the time period I.e. when a person signs up or starts using the product or services.

11.3.2 Segment-Based Cohort Analyzing - The classification of cohorts is based on the type of cohorts I.e. the type of product or services opted by people.

11.3.3 Size-Based Cohort Analyzing - The classification of cohorts is based on the number (size) of people involved in the usage of a product or service.

11.4 Descriptive Analysis - It forms patterns of data points that describe the data and analyze it.

11.5 Predictive Analysis - Predictive analysis is used to extract a variety of statistical techniques from data mining, predictive modeling and machine learning to predict the future. Predictive analysis can be both qualitative and quantitative. For example, weather forecasting can be about predicting the upcoming weather conditions. Predictive analysis asks questions on the basis of past data having faced different conditions. (Waller & Fawcett, 2013)

11.6 Diagnostic Analysis - It is an advanced form of data analytics to find out the answer to a question by asking 'why did it happen?'

11.7 Regression Analyzing - It helps to estimate the relationship between two or more variables and also helps us to understand the impact of the variables by classifying the variables into independent and dependent variables.

11.7.1 Logistic Regression Analyzing - It is generally used for classification problems. The result is always displayed as 0 (False) or 1(True).

11.7.2 Linear Regression Analyzing - It is generally used for regression problems. The result is displayed as continuous values.

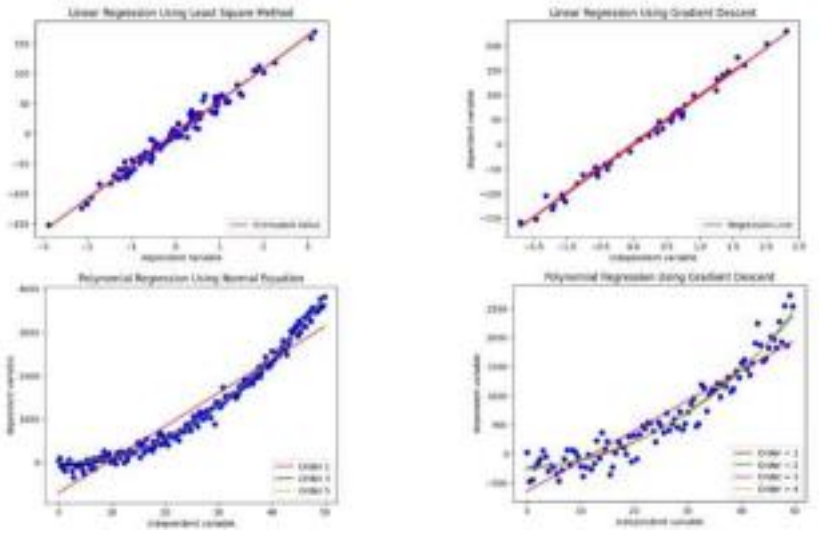

fig.5. Linear regression analysis (Brital, 2021)

11.8 Factor Analyzing

It is a technique in which the variables are converted into fewer numbers of factors which makes analyzing easier. It is a subpart of the General Linear Model (GLM).

11.9 Text Analyzing

It is a process of drawing meaningful insights from the keywords extracted.

11.10 Time Series Analyzing

In this type of analysis, the data points represent the periods of time and these points are then analyzed. This also helps us to represent the growth of an organization over the years. It also helps to track the progress of economic conditions, an asset or any system. (Hayes, 2022)

11.11 Decision Tree

It is a process of making a flowchart that carries out an idea and branches out the consequences.

11.12 Conjoint Analyzing - It helps to analyze data obtained in a survey and is generally used by different industries in advanced market research analysis. There are two types of Conjoint analysis: Choice-based Conjoint (CBC) Analysis and Adaptive Conjoint Analysis (ACA).

11.13 Monte Carlo Method - John von Neumann and Stanislaw Ulam during World War II to improve decision-making under uncertain conditions, they invented this method. It was named after Monaco town. It is a predictive model and analyzes different probabilities to give output.

11.14 Neural Network - also known as Artificial Neural Network (ANN) focuses on mimicking the structure of the human brain and enables the artificial system to learn, recognize and make human-like decisions. Like the human brain, they are developed to function by processing data that is learned by doing and hence it does not require programming. Similar to the human brain's neurons it uses nodes for processing data.

The neural network is a sub-part of machine learning which is one of the methods used for data modeling and statistical analysis. Neural networks are generally used for classification or forecasting and recognize patterns that are useful for making decisions.

It is also used for -

• Handwriting recognition

• Language Translation

• Voice recognition

• Automated driving vehicles

It contains numerous layers like the input layer, hidden layers and output layer. These layers consist of several interconnected notes which have an inbuilt 'activation function'. This function depends on weights and thresholds.

11.14.1 Methods used by Neural Network for Learning

• Supervised learning - The system is provided with many inputs and is trained to provide the correct output. • Unsupervised learning - The system is provided with inputs and is expected to analyze them to provide the output.

• Reinforcement learning - The system learns based on the feedback given by the users. Negative feedback helps them to understand the wrong output whereas positive feedback helps them to recognize the correct output.

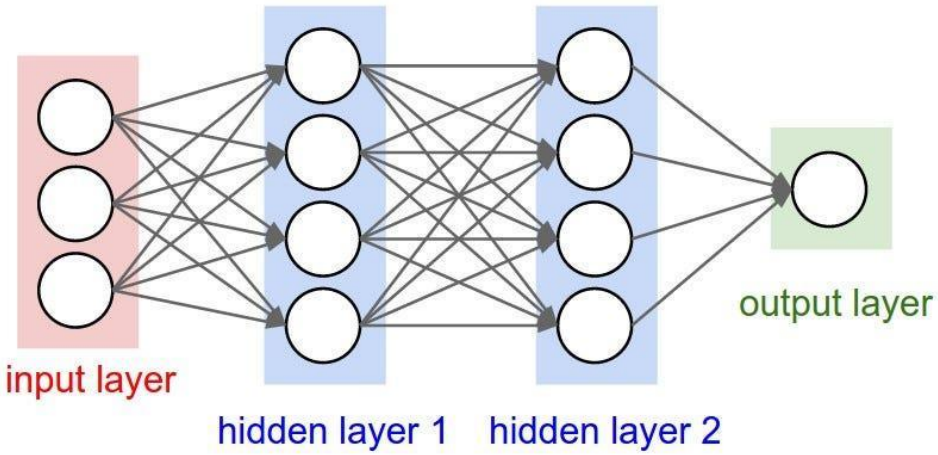

Fig 6. Neural Network (Dabbura, 2018)

12. Machine learning and data science

Machine learning - Machine learning refers to the capability of the artificially created system to learn from experience, for example by processing data the process by which a natural language translation program improves itself when used more and more. This is one of the subfields of artificial intelligence. (Kips,2020)

Machine learning is a sub-part of data science. After analyzing data the process of utilizing these data and mankind the system learns the insights from the data provided is called machine learning. It helps to make artificial devices 'smarter'. (Tavasoli, 2022)

12.1 Timeline of machine learning

The term machine learning was coined in 1959 by Arthur Samuel, an IBM employee and pioneer in the field of computer gaming and artificial intelligence. Also, the synonym self-teaching computers were used in this time period

It was first introduced in 1950 by Alan tuning in his journal 'Computing Machinery and Intelligence' so that the learning machine could learn and become smart and artificially intelligent.

Growth in machine learning:-

According to the estimation, the market size of machine learning in the year 2021 is 21 billion us dollars and by 2025 it will be 58.2 us dollars and by 2030 it will be 305 us dollars.

12.2 Features of Machine Learning

• It is a field that tries to mimic Human capabilities to learn new things and adapt to the surroundings.
• Machine learning is a method of analyzing data that automatically generate a model to analyze the given data. • Machine learning allows its user to feed the computer the important algorithms and data to analyze the input data.
• Machine learning is, however, a small part of data science, without data it is almost impossible for the models made by machine learning to execute its work. (Harrington, 2012) (Kips, 2020)

12.3 Approaches in machine learning

● Machine learning approaches are traditionally divided into three broad categories, which correspond to learning paradigms, depending on the nature of the "signal" or "feedback" available to the learning system.

● Supervised learning, unsupervised learning and Reinforcement learning are several approaches to machine learning also in the coming years.

12.4 Advantages of Machine Learning

Everything comes with advantages and disadvantages at the same time but we will see it in a positive aspect, a few advantages of machine learning are:

• Identification of the new trend

• It learns continuously from its experiences

• It easily manages the data

• Automation of everything

• Provides a wide range of adaptability

• Machine learning also requires the same qualities and expiries to understand it just like it was in the case of data science. • It learns from its experience and never makes the same mistake again in the future which is a main advantage of machine learning.
Machine learning is still in its working stage and it will further acquire a very important place in the whole world. A situation will come when the whole world would be dependent upon machines for some of the other execution of work.

12.5 Limitation of machine learning:-

● Although machine learning has been transformative in some fields,machine-learning programs often fail to deliver the expected results. Reasons for this are numerous: lack of (suitable) data, lack of access to the data, data bias, privacy problems, badly chosen tasks and algorithms, wrong tools and people, lack of resources, and evaluation problems.

● Example:-In 2018, a self-driving car from Uber failed to detect a pedestrian, who was killed after a collision. Attempts to use machine learning in healthcare with the IBM Watson system failed to deliver even after years and billions of dollars were invested

12.6 Machine learning and conventional human learning

12.6.1 Machine becoming intelligent

This idea was first introduced by Alan Turing who asked a very simple and revolutionary question 'Can machines think?'.This was the first leap toward Artificial Intelligence. (Turing,1950)

12.6.1.1 The Imitation Game

This game was introduced by Alan Turing to better understand the concept of machine thinking. In this game, there are three players: a human, an intelligent machine and an interrogator. The interrogator has to be inside a different room and has to decide who is the human and which one is the machine by asking questions. They are labeled as different variables to the interrogator. So that the interrogator can't identify them on the basis of the tone of their voice the answer must be typed. In order to confuse the interrogator they may or may not lie. (Turing,1950)

12.6.2 Memory

Computers have long-term memory and can store any information if asked by the user whereas humans store most of the information for a short term. Computers also have selective memory i.e. it can store the required data and delete the data which is not required but humans can't have selective data.

12.6.3 Prediction and forecasting

Most of the predictions made by machines are by recognising patterns in the data and by analyzing the data gained from various experiences whereas humans predict most of the things subconsciously.

12.6.4 Accuracy

Due to logical reasoning, analyzing and unbiased processing the output provided by machines are more accurate.

12.6.5 Power consumption

The power consumed by the machines is very high compared to that of the human brain.

12.6.6 Emotion

Today's AI-based machines lack the aspect of emotions but scientists are working on it. When it comes to emotions, all human decisions and opinions are highly based on it and it is the main aspect to differentiate between humans and machines.

12.6.7 Interpretation

All the operations performed by today's AI are performed without interpreting the outcomes. They are unable to interpret the reason for processing data.

12.6.8 Creativity

AI is not creative. They can only hallucinate human creative work.

12.6.8.1 Examples of limitations of creativity of AI

•AI can compose new songs by using the pre-recorded human voice, pre-made genre and by combining pre-made music and lyrics but it cannot create a new music genre.

•AI can generate new paintings by analyzing the paintings provided to it as data but it cannot create new ideas. Example - Rainy street in the night painting

Fig 7. Painted by human (Lagenberg, 2021)

Fig 8. Painted by AI

12.7 Applications and Real life uses of Machine learning

Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day. We are using machine learning in our daily life even without knowing it such as Google Maps, Google Assistant, Alexa, etc. Below are some most trending real world applications and uses of Machine Learning:-

● Image recognition - generally use the technology of machine learning for recognising images in various social media apps like WhatsApp, Facebook, Twitter, Google, etc.

● Vision systems - Machine learning uses its technology for predicting and analyzing the given data it can be used in Content based image retrieval, forensic labs and in many more ways.

● Medical diagnosis - Machine learning uses its system to save the life of people by providing them treatment based on 24 *7 monitoring. It helps people to get well as soon as possible by providing them with a better way to treat a

particular disease.

● Self-automated machines - Machine learning is adjusted into various devices, transport systems, technologies, etc to enhance its working.

● Natural Language processing-Machine learning is also present in the development of natural language (the ability to speak human-like language) which includes various devices like Alexa, Siri, etc.

● Commercial Establishments - establishing big businesses requires a lot of hard work and dedication. This work can be made easier by using machine learning to advance the quality.

● Decision making - Machine learning plays a very important role in making better decisions for example diagnosing patients, weather forecasting, etc.

● Traffic Prediction - If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions. It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways: Real Time location of the vehicle from the Google Map app and sensors. And Average time has taken on past days at the same time.

● Product Recommendations- Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendations to the user. Whenever we search for some product on Amazon, then we start getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning. (Kips,2020)

12.8 Vision for Machine Learning and Data Science

Machine Learning (ML) is so versatile and powerful that it's one of the most exciting technologies of our time. Amazon uses it, Netflix uses it, Facebook uses it, and many more companies use it. With all other hyped technologies, there are a lot of misconceptions about machine learning too. In this article, we will discuss the future of machine learning and its value throughout industries. Machine learning works on the principles of computer algorithms that learn in a reflex manner through trials and experiences. ML is an application of Artificial Intelligence that permits program applications to anticipate results with utmost precision. It makes a distinction to create computer programs and to assist computers to memorize without human intercession. Machine learning plays an important role in the field of enterprises as it enables entrepreneurs to understand customers' behavior and business functioning behavior. Nowadays leading companies like Google, Amazon, Facebook, Tesla, and many more are efficiently utilizing these technologies, as a result, machine learning has become a core operational part of functioning. The skills that data scientists use to perform their work will change, with coding and AI becoming more and more essential. In parallel, they also need to be much more business-minded. In the past, data scientists focused instead more on statistics and modeling and less on coding. This shift is due in part to the increasing complexity of data. Data sets are growing larger and more disparate, making it more difficult to glean insights from them. Meanwhile, the tools that data scientists use to analyze data have become more sophisticated. As datasets have gotten larger and more complex, the need for data scientists to have strong coding skills has therefore increased. The same is true for machine learning. **(Moioli, 2022)**

13. Data warehouse

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process". **(Inmon, 1995)**

References

1. Kips Content Development Team. (2020). KIPS Artificial Intelligence Class-9. Kips Learning Private Limited.

2. Griffin, Paige (2019, May 9) https://medium.com/quick-code/top-data-science-trends-to-watch-for-in-2019- cafc8034db4c

3. Omair Aasim , Benney Mathenews (January , 2023). Retrieved from https://www.projectpro.io/article/10-different-types of-data-scientists/179

4. Vandana Kakkar , (December , 2022) Retrieved from https://www.jigsawacademy.com/blogs/data-science/types-of-data scientists/

5. Alan Turing, "Turing." Computing machinery and intelligence. Mind 59.236 (1950): 433-60. 6. Rawat, T., & Khemchandani, V. (2017). Feature engineering (FE) tools and techniques for better classification performance. International Journal of Innovations in Engineering and Technology, 8(2), 169-179.

7. Hayes, A. (2022, June 12). Retrieved from https://www.investopedia.com

8. Press, G. (2013, May 28). Retrieved from https://www.forbes.com

9. Sharma, A. (2020, September 8). Retrieved from https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering works/

10. Tavasoli, S. (2022, September 13). Retrieved from https://www.simplilearn.com

11. S. Sinharay (2010). Retrieved from International Encyclopedia of Education

12. Amal Joby (19 July 2021). Retrieved from https://learn.g2.com/k-nearest-neighbor

13. Shreya Joshi (2022, August 5). Retrieved from https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/ 14. Retrieved from https://byjus.com/maths/cluster-analysis/

15. Retrieved from https://www.dominodatalab.com/data-science-dictionary/density-based-clustering 16. (2019, September 19) Retrieved from https://en.m.wikipedia.org/wiki/DBSCAN

17. Retrieved from https://www.wallstreetmojo.com/cohort-analysis/

18. Retrieved from https://byjus.com/mlp-dssl/

19. Retrieved from https://www.alchemer.com/resources/blog/regression analysis/#:~:text=Regression%20analysis%20is%20a%20powerful,variables%20on%20a%20dependent%20variable

20. Retrieved from https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

21. Retrieved from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/

22. Retrieved from https://monkeylearn.com/text-analysis/

23. (2021, December 6)Retrieved from https://asana.com/resources/decision-tree-analysis

24. Hayes, Adam. (2022, June 12). Retrieved from https://www.investopedia.com/terms/t/timeseries.asp

25. Retrieved from https://sawtoothsoftware.com/conjoint-analysis

26. Bansal, S. (2021, January 6). Why is data science important for you and businesses? Retrieved from https://www.analytixlabs.com