International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 5, ISSN No. 2455-2143, Pages 237-240
Published Online September 2022 in IJEAST (http://www.ijeast.com)

COMPONENTS OF DATA SCIENCE AND ITS APPLICATIONS

Sudeepa Paul, Doeyl Sah, Rupsa Das, Somhrita Goswami, Souvik Samanta, Biswadeep Roy, Debrupa Pal
Department of Computer Application
Narula Institute of Technology, Kolkata, West Bengal, India

Abstract— **It is difficult to define data science in a formal way because it is expanding rapidly and revolutionizing many industries. The most basic definition of data science, however, is the process of drawing out useful information from unstructured data. It is being applied to any human endeavor for which there is sufficient data because it is a result of scientific discovery. There have been significant and spectacular accomplishments, and even more sweeping claims have been made. Along with the advantages, there are also challenges and risks. However, a term alone cannot explain what data science specifically is. This paper examines the fundamentals, lifetime, and several applications of this field**

Keywords— **Data science tools, Components, Life Cycle, Use cases**

I. INTRODUCTION

Data science, also referred to as data-driven science, includes various statistical and computational disciplines to transform data to be used in decision-making. It uses an interdisciplinary approach to mine the vast amounts of data being gathered and produced for practical insights. Data is flooded from a variety of industries, platforms, and apps, including mobile devices, social media, online stores, surveys related to the healthcare industry, and internet searches. Data scientists must be skilled in everything from data engineering to arithmetic, statistics, complicated computing, and visualization in order to effectively sort through confusing volumes of data. Through these abilities, data scientists create statistical models that examine data and identify patterns, trends, and linkages in large data sets.

STAGES OF A DATA SCIENCE PROJECT

A. Data ingestion

The data collection phase of the lifecycle starts with gathering raw, unstructured, and structured data from all relevant sources using several techniques. These techniques can involve data entry by hand, online scraping, and real-time data streaming from machines and gadgets. Structured data, like client.

B. Data Storage and Data Processing

Companies must consider various storage systems depending on the type of data that has to be captured because data can have a variety of formats and structures. Creating standards for data storage and organization with the aid of data management teams makes it easier to implement workflows for analytics, machine learning, and deep learning models [2]. Using ETL (extract, transform, load) jobs, this stage involves cleaning, deduplicating, manipulating, and merging the data. Prior to being loaded into a data warehouse, data lake, or other repository, this data preparation is crucial for boosting data quality.

C. Data Analytics

In order to look for biases, trends, ranges, and distributions of values within the data, data scientists perform an exploratory data analysis. The generation of hypotheses for a/b testing is driven by this data analytics exploration. Additionally, it enables analysts to evaluate the data's applicability for modelling purposes in predictive analytics, machine learning, and deep learning [3]. Organizations may depend on these insights for corporate decision-making, enabling them to achieve more scalability, depending on the model's accuracy.

D. Communicate

Furthermore, findings are represented as reports and other data visualizations to help business analysts and other decision-makers better understand the insights and how they will affect the organization [4]. In addition to using specialized visualization tools, data scientists can create visualizations using components built into programming languages for data science, such as R or Python [5].

II. TOOLS FOR DATA SCIENCE

Data science efforts might be slowed down when data is trapped inside papers and images with no organized data representation. Data scientists need a variety of tools and programming languages to perform their duties as effectively as possible [6]. These tools can include, for example:

A. Data Analysis

R, Python, SAS, Jupyter, R Studio, MATLAB, Excel, and RapidMiner are a few examples.

B. Data Warehousing

SQL Informatics/Talend, Hadoop, ETL, and AWS Redshift

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 5, ISSN No. 2455-2143, Pages 237-240
Published Online September 2022 in IJEAST (http://www.ijeast.com)

C. **Data Visualization**
Cognos, Tableau, Jupyter, and R.

D. Machine learning
Mahout, Spark, and Azure ML studio

III. COMPONENTS OF DATA SCIENCE

The following list contains the key elements of data science.

A. Statistics
One of the most crucial elements of data science is statistics. In order to gather and evaluate enormous amounts of numerical data and derive useful insights from it, statistics is used [7].

B. Domain Expertise
Technical expertise is the glue that holds data science together. Expertise in a certain domain refers to specialized knowledge or abilities. Domain experts are needed in several fields of data science.

C. Data engineering
Data science, which deals with gathering, storing, retrieving, and transforming data, includes data engineering. Metadata (information about data) is also a component of data engineering.

D. Visualization
The goal of data visualization is to portray information visually so that viewers can quickly grasp its relevance. Accessing the vast amount of data presented in visuals is made simple by data visualization.

E. Advanced computing
Advanced computing is the data science's primary function. Designing, writing, debugging, and maintaining the source code of computer programs are all part of advanced computing.

F. Mathematics
The essential component of data science is mathematics. The study of quantity, structure, space, and changes is a component of mathematics. A data scientist needs to have a solidunderstanding of mathematics [8].

G. Machine Learning
Machine learning is backbone of data science. Machine learning is all about to provide training to a machine so that it can act as a human brain. In data science, we use various machine learning algorithms to solve the problems.

IV. MACHINE LEARNING IN DATA SCIENCE

In order to become a data scientist, one should also be familiar with machine learning and its methods. This is because many machine learning algorithms are employed in data science. Several machine learning algorithms are used in data.

A. Regression algorithm
The most widely used supervised learning-based machine learning algorithm is linear regression. Regression is a technique used in this algorithm to model target values based on independent variables. It depicts the shape of the linear equation, which establishes a connection between a collection of inputs and a prognostic outcome. The main applications of this technique are forecasting and prediction [9]. It is known as linear regression because it demonstrates the linear relationship between the input and output variables as described in Figure1.

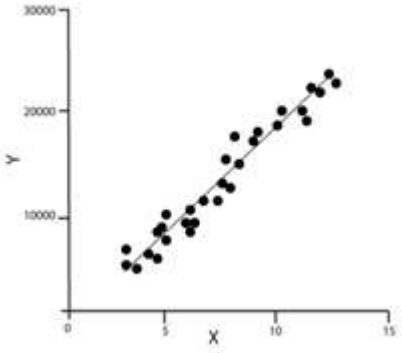

Figure 1: Plotting through regression

Equation1 can describe the relationship between x and y variables

$$Y=mx+c \quad (1)$$

where y is dependent variable, x is independent variable, m isslope and c is intercept.

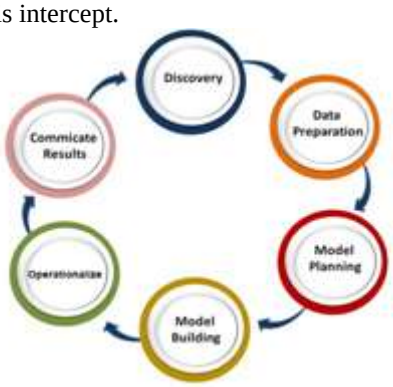

Figure 2: Phases of data science life cycle

B. Decision Tree
Another machine learning method that is a part of the supervised learning algorithm family is the decision tree algorithm. One of the most well-liked machine learning algorithms is this one. Bothclassification and regression issues can be solved using it.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 5, ISSN No. 2455-2143, Pages 237-240
Published Online September 2022 in IJEAST (http://www.ijeast.com)

C. K-means clustering

One of the most widely used machine learning algorithms, , K- means clustering is a form of unsupervised learning technique. The clustering issue is resolved. k-means clustering approach canbe used to tackle situations when we have a data set of items with specific attributes and values that is needed for classification into groups [10].

V. DATA SCIENCE LIFECYCLE

Different phases of life cycle are described in Figure 2 [11].

A. Discovery

The preliminary step is asking the correct questions, and is a necessary part of the discovery process. Project's budget, priorities, and fundamental requirements are decided before beginning any data science project. The project's prerequisites, including the quantity of workers, available technology, available time, available data, and the project's end aim, must all be established during this phase before the business challenge can be framed at the first hypothesis level.

B. Data Preparation

Data munging is another name for data preprocessing. Tasks during this phase are data cleansing, data reduction, data integration, and data transformation. After completing the all these steps, this data is readily utilized for further operations before beginning any data science project. The project's prerequisites, including the quantity of workers, available technology, available time, available data, and the project's end aim, must all be established during this phase before the business challenge can be framed at the first hypothesis level.

C. Model Planning

The numerous approaches and techniques are identified in this stage in order to establish the relationship between the input variables. By utilizing various statistical formulas and visualization tools, exploratory data analytics (EDA) is applied to explore the relationships between variables and determine the type of information data that can be provided.

D. Model Building

The process of creating models begins in this stage. Datasets are constructed for testing and training, to develop the model, a variety of techniques are used, including association, classification, and clustering.

E. Operationalize

The project's final reports are delivered at this phase, along with technical documents, code, and briefings. Before the actual deployment, this phase gives a thorough overview of the performance of the entire project and other components.

F. Communicate Results

In this phase, checking is performed to verify whether the objectives set during the first phase are achieved or not. The business team is conveyed about the findings and the outcome.

VI. USE CASES

A. Image recognition and speech recognition

Images and audio are presently recognised using data science. When friends are suggested to be tagged on a photograph in Facebook, the picture recognition technique used in this automatic tag suggestion is a component of data science. Speech recognition algorithms make this feasible when you use phrases like "Ok Google," "Siri," "Cortana," etc. and these devices answer in accordance with voice commands [12][13].

B. Gaming World

Machine learning algorithms are being used more and more in the game industry. Data science is frequently used by companies like EA Sports, Sony, and Nintendo to improve user experience.

C. Internet Search

A variety of search engines are utilized like Google, Yahoo, Bing, Ask, etc., when it is needed to find something online. A search result is received in just a few milliseconds as data science technologies are used by all of these search engines to improve the search experience.

D. Recommendation Systems

Most businesses, like Amazon, Netflix, Google Play, and others, use data science technologies to improve the user experience by providing tailored recommendations. For instance, data science technology is responsible for the suggestions for related products customers receive when they conduct a search on Amazon.

E. Healthcare

Data science has many advantages in the healthcare industry. Tumor identification, medication development, medical image analysis, virtual medical robots, and other applications of data science are being applied [14][15].

VII. CONCLUSION

It can be concluded that the future belongs to data scientists. Making appropriate business judgments will be possible as more data becomes available. Even though this article only provided an outline of data science, this field is one that has a lot of potential and will only develop further if there are online networking platforms. Data science is progressing positively in the field of education. Since a few years ago, it has developed and is now creating workers with exceptional and complementary talents in the fields of statistics, information, and computing. Data science is becoming more and more in

demand in industry, and researchers are becoming more and more interested in it.

VIII. REFERENCE

[1] Aalst, Wil van der. "Data science in action." In Process mining, pp. 3-23. Springer, Berlin, Heidelberg, 2016.

[2] Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.", 2013.

[3] Soriano-Valdez, David, Ingris Pelaez-Ballestas, Amaranta Manrique de Lara, and Alfonso Gastelum-Strozzi. "The basics of data, big data, and machine learning in clinical practice." Clinical Rheumatology 40, no. 1, pp 11-23, 2021

[4] Waller, Matthew A., and Stanley E. Fawcett. "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management." Journal of Business Logistics 34.2, pp 77-84, 2013

[5] Provost, Foster, and Tom Fawcett. "Data science and its relationship to big data and data-driven decision making." Big data 1.1, pp 51-59, 2013

[6] Barlas, Panagiotis, Cathal Heavey, Ivor Lanning,. "A survey of open-source data science tools." International Journal of Intelligent Computing and Cybernetics, 2015

[7] Cees De Laat, Demchenko, Yuri and Peter Membrey. "Defining architecture components of the Big Data Ecosystem." International conference on collaboration technologies and systems (CTS). IEEE, 2014.

[8] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349.6245, pp 255-260, 2015

[9] Joshi, Jigyasu, and Shweta Saxena. "Regression analysis in data science." Journal of Analysis and Computation 14.6, 2020

[10] Dierckens, Karl E., et al. "A data science and engineering solution for fast k-means clustering of big data.", IEEE Trustcom/BigDataSE/ICESS. IEEE, 2017.

[11] Wing, Jeannette M. "The data life cycle.", 2019.

[12] Anupama C. Raman Raj, Pethuru, and. The Internet of Things: Enabling technologies, platforms, and use cases. Auerbach Publications, 2017.

[13] Hong, Tianzhen, Zhe Wang, Xuan Luo, and Wanni Zhang. "State-of-the-art on research and applications of machine learning in the building life cycle." Energy and Buildings 212, 2020

[14] Aron Henriksson,, Dalianis, Hercules, Maria Kvist, Rebecka Weegar and Sumithra Velupillai "HEALTH BANK-A Workbench for Data Science Applications in Healthcare." CAiSE Industry Track 1381,pp 1-18, 2015

[15] Catherine Chronaki , Robert Stegwee and Schulz, Stefan, "Standards in healthcare data." Fundamentals of Clinical Data Science, pp 19-36, 2016