IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 12, December 2023

DOI: 10.17148/IJARCCE.2023.121215

Survey on data science: its technique, tools and Open issues

Mrs Anagha Abhijit Jawalkar

Asst.Professor, Department of Artificial Intelligence & Data Science, D.YPatil COE Akurdi Pune India

Abstract: Now a day Data science is emerging field , before data science we had statisticians. These statisticians are skilled person who are evaluate records and organizations hired them to research their standard overall performance and income. Data science is a booming field of study which has a multidimensional scope for all organizations and industries. Data Science has lots of scientific methods which includes statistical techniques, machine learning, artificialintelligence all together we all know from earlier time the mathematics can solve the once complex problems. It gives various information on emerging trends and patterns in a specific model . Data science provides with various methodsto analyzed data, and make predictions on the data. The basic objective of this paper is to explore the techniques of data science , tools which are available as an open source for data science and various tools associated with it..–

Keywords: Machine Learning, Data science, Open source, Data science Tools ,Big data analytics, Structured data; Unstructured Data

I. INTRODUCTION

The term “Data Science” has emerged because of the evolution of statistical techniques as well as different mathematical techniques, data analysis, and the field of big data.Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data.Data Science is the area of study which basically extracting insights from vast amounts of data using various scientific methods, algorithms, and processes. It helps you to discover hidden patterns from the raw data.This enables us to to translate a business problem into a research project and then translate it back into a practical solution.It is a thorough study of the large collection of the data, which involves extracting meaningful insights from raw data and processed by using the scientific method, different technologies, and algorithms.Itis a multidisciplinary field that uses tools and techniques to manipulate the data we can find something meaningful newinformation. Data science combines math and statistics, specialized programming, analytics, Artificial intelligence (AI), and machine learning(ML) with specific subject matter expertise to uncover actionable insights hidden in an organization’s data. These insights can be used to guide decision making and strategic planning.

There are various roles, tools, and processes, which enables data analysts to get actionable insights from data.Datascience project involves different stages:

- **Data collection :** This first stage begins with the data collection wherein both raw structured and unstructured data collected from all relevant sources using a variety of methods. These methods can include manual entry, web scraping, and real-time streaming data from systems and devices. Data sources can be any structured data,like different organization customer data, along with unstructured data like log files, video files, audio files pictures images

- **Data storage and data processing**: data can have different formats and structures, companies need to consider different storage systems based on the type of data that needs to be captured. Variety of Data management teams help to set standards around data storage and structure, which facilitate proper workflowsof data around analytics, machine learning and deep learning models. This stage includes cleaning data, removing duplicate transforming and combining the data using ETL (extract, transform, load)as well as method of integration technologies applied on it. This data preparation is essential for promoting data qualitybefore loading into a data warehouse.

- **Data analysis:** In this stage data scientists evolves an exploratory data analysis to examine different biases, different patterns, ranges, and distributions of values within the data. Hypothesis testing data generation alsoperformed. It also allows analysts to determine the data friendly for use predictive analytics, machine learning, and deep learning technique. According to model accuracy, organizations can rely reliant on thoseinsights are getting for business and used for decision making.

- **Communicate:** Eventually in last cycle insights are presented as reports and different types data visualizations technique to insights more clear and can be understood by someone who is really not known much about it.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 12, December 2023

DOI: 10.17148/IJARCCE.2023.121215

In data science various programming language like R and Python used these helps to get components for generating visualizations and after that data scientists can use dedicated visualization tools.

Designing an intelligent system, for that we requires knowledge of a data Scientist which will use various statistical methods & machine learning algorithms to analyze and explore data. Data Scientist is a specialized person in data science, which will analyze the data get the patterns and insights from the data and used differentmachine learning algorithms for the future prediction of events according to the requirement of model.

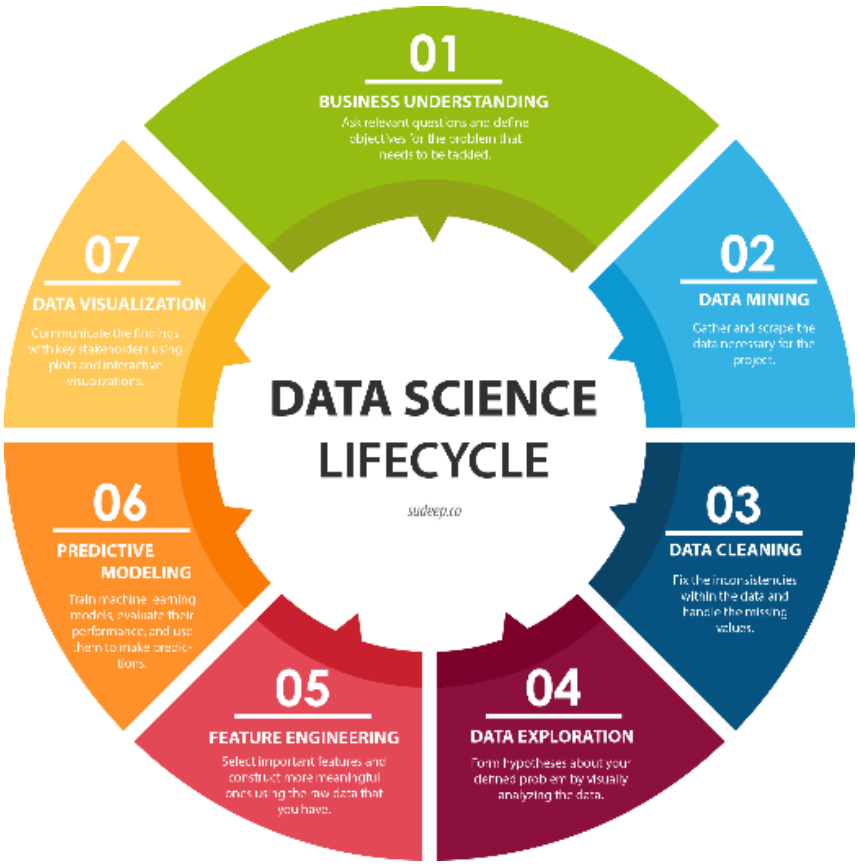

Fig 1: Data Science Life cycle

II. TECHNIQUES USED IN DATA SCIENCE

Variety of techniques are available in Data science, which totally depends on the type of data, and collectionof data, Once data are analyzed, the techniques are applied on it.

1. Machine learning: To understand data science, one needs to understand the concept of machine learning. Data science uses machine learning algorithms to solve various problems. There are Machine Learning algorithm are as Supervised Learning

1.a. Supervised Learning : It is based on the results of a previous operation that is related to the existing business operation. Based on previous patterns, Supervised Learning aids in the prediction of an outcome. Someof the Supervised Learning Algorithms are Linear Regression ,Random Forest Support Vector Machines.

1.b. Unsupervised Learning: This form of learning has no pre-existing consequence or pattern. Instead, it concentrates on examining the interactions and connections between the presently available Data points. Someof the Unsupervised Learning Algorithms are KNN (k-Nearest Neighbors),K-means Clustering, Hierarchical Clustering, Anomaly Detection.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 12, December 2023
DOI: 10.17148/IJARCCE.2023.121215

1.c. Reinforcement Learning : It is a fascinating Machine Learning technique that uses a dynamic Dataset that interacts with the real world. In simple terms, it is a mechanism by which a system learns from its mistakes and improves over time. Some of the Reinforcement Learning Algorithms are Q-Learning,State Action Reward- State-Action (SARSA), Deep Q Network.

2. Mathematical modeling: Mathematical modeling is required to make fast mathematical calculations and predictions from the available data.

3. Statistics: Basic understanding of statistics is required, such as mean, median, or standard deviation. It isneeded to extract knowledge and obtain better results from the data.

4. Computer programming: For data science, knowledge of at least one programming language is required.R, Python, Spark are some required computer programming languages for data science.

5. Databases: The depth understanding of Databases such as SQL, is essential for data science to get the data and to work with data.

III. TOOLS FOR DATA SCIENCE

The major role of Data Scientists is to perform data analysis on the basis of structured data as well as unstructured data to make prediction is the big task.so to handle this need to have a massive amount of data and for that suitable programming languages and tools are needed.so that they can clearly performed their task. In this surveypaper, We will explore some available tools for data science which is use for data analysis and perform predictions.

1. **Data Analysis :** R, Spark, Python and SAS
2. **Data Warehousing :** Hadoop, SQL, Hive
3. **Data Visualization** : R, Tableau, Raw
4. **Machine Learning** : Spark, Azure ML studio, Mahout

IV. APPLICATIONS OF DATA SCIENCE

Every industry benefits from the experience of Data Science companies, but the most common areas where DataScience techniques are employed are the following:

1. Banking and Finance: The banking industry can benefit from Data Science in many aspects. Fraud Detection is awell-known application in this field that assists banks in reducing non-performing assets.

2. Healthcare: Health concerns are being monitored and prevented using Wearable Data. The Data acquired from thebody can be used in the medical field to prevent future calamities.

3. Marketing: Marketing offers a lot of potential, such as a more effective price strategy. Pricing based on Data Sciencecan help companies like Uber and E-Commerce businesses enhance their profits.

4. Government Policies: Based on Data gathered through surveys and other official sources, the government can useData Science to better build policies that cater to the interests and wishes of the people.

5. Internet Search : Google search uses Data science technology to search for a specific result within a fraction of asecond

6. Recommendation Systems : To create a recommendation system. We can use this recommendation like suggestion to particular thing with the help of different applications on Facebook , YouTube, everything is done with the help of Data Science**.**

7. Image & Speech Recognition : Now a days Speech recognizes systems like Siri, Google Assistant, and Alexa run beavailable with the help of the Data science technique. Moreover, Facebook recognizes your friend when you upload a photo with them, with the help of Data Science.

8. Gaming world : EA Sports, Sony, Nintendo are using Data science technology. This enhances your gaming experience. Games are also developed with the help of Machine Learning techniques, and they can update automatically by themselves when you go to higher levels**.**

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed journal & Refereed journal ⁜ Vol. 12, Issue 12, December 2023
DOI: 10.17148/IJARCCE.2023.121215

V. CHALLENGES OF DATA SCIENCE TECHNOLOGY

- A high variety of information & data is required for accurate analysis
- Not adequate data science talent pool available
- Management does not provide financial support for a data science team
- Unavailability of/difficult access to data
- Business decision-makers do not effectively use data Science results
- Explaining data science to others is difficult
- Privacy issues
- Lack of significant domain expert
- If an organization is very small, it can't have a Data Science team

VI. CONCLUSIONS

Finally in this survey paper we can conclude that there are a number of techniques and tools available for performing data analysis offcourse by data scientists and for Data analysis we need to learn this techniques andtools thoroughly. Begins with the Data collection storage and processing and Data Analysis with the help of Python Libraries and making machine learning models and make prediction are the steps of data science projects and to get the proper visualization is also important. Many tools of data science tools can perform Complex dataOperation in one framework so that it is convenient to implement the functionalities of data science without knowing any language experience.

REFERENCES

[1]. Russell, Stuart J., and Peter Norvig. Artificial intelligence: a modern approach. Malaysia; Pearson EducationLimited,, 2016.
[2]. Nicolae, Bogdan, et al. "Park, Yoonho. Leveraging Adaptive I/O to Optimize Collective Data Shuffling Patterns for Big Data Analytics. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. PP (99) pp: 1-13." (2020).
[3]. Islam, Mohaiminul. "Data Analysis: Types, Process, Methods, Techniques and Tools." International Journalon Data Science and Technology 6.1 (2020): 10.
[4]. Dhar, Vasant. "Data science and prediction." Communications of the ACM 56.12 (2013): 64-73.
[5]. Bejjam, Suvarnamukhi & Seshashayee, M.. (2018). Big Data Concepts and Techniques in Data Processing. International Journal of Computer Sciences and Engineering. 6. 712-714.
[6]. Van Der Aalst, Wil. "Data science in action." Process mining. Springer, Berlin, Heidelberg, 2016. 3-23.
[7]. Ethem Alpaydin (2004). Introduction to Machine Learning, MIT Press, ISBN 978-0-262-01243-0.
[8]. Stuart Russell & Peter Norvig, (2009). Artificial Intelligence- A Modern Approach. Pearson, ISBN 9789332543515.
[9]. C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies:A survey on big data, Information Sciences, 275 (2014), pp.314-347.
[10]. K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.