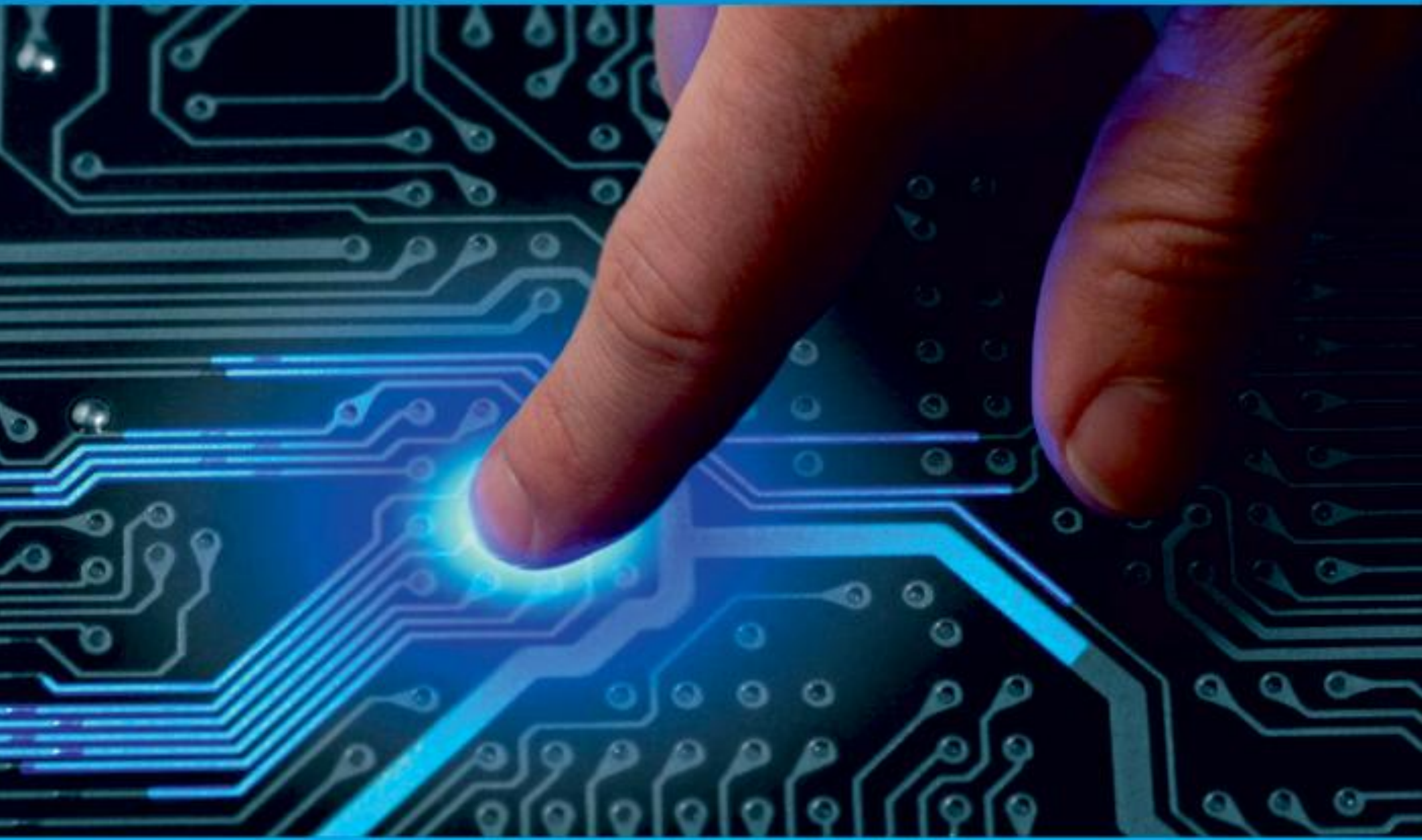
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Special Issue 2, March 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

Data Science: The emerging domain in IT

Prathamesh Yechwad[1], Ashok Bhosale[2], Dr Monika D. Rokade[3]

Student, Sharadchandra Pawar Collage of Engineering Otur Dist-Pune, India[1,2]

Assistant Professor at Sharadchandra Pawar Collage of Engineering Otur Dist-Pune, India [3]

ABSTRACT: With the rise of advance data collection tools, the scope of inferential statistics and predictive analytics has increased to a great extent, and reliance on numbers for policy making, business solutions for complex problems, data-based decisions taken in cases of various emergencies has led to the development of field data science

KEYWORDS: Data collection, Data cleaning, machine language, predictive analysis

I. INTRODUCTION

Here we will introduce when and how this data is collected, then move on to how we process this data to make it meaningful and ready for analysis, in between we will clarify the difference between analytics and analytics, which are often misinterpreted as being the same, then you we'll walk through the sectors where data science is most commonly used and the different roles involved across the spectrum of data science. We will also introduce you to the tools used in these areas for both data collection and data analysis. We will give a brief summary of the transition of data science and finally move to the part of machine learning or modern techniques of data science [1,2].

II. DATA COLLECTION

Before we move on, we will distinguish between the terms analysis and analytics. Analysis is when we use inferential statistics to find answers to questions from the past, while analytics is when we use collected data and know techniques to predict future outcomes. So analysis is something that is done before analysis because you need to know the past to predict the future. This brings us to data collection, which has revolutionized the field of data science because more data means more accurate predictions. The traditional approach to data collection was the survey method where selected people from a group of people were selected and interviewed and the data collected was sent for processing. Then came online surveys, which became so hated by people that they would only fill them out if they were also forced to. With the invention of social media and people providing their information along with their likes and dislikes to these platforms, the scale of data collection has increased, but the technique to process such vast data has been lacking. Eventually, big data collection tools began to be used, using high-tech computers and even supercomputers to collect raw data from these platforms and process it into categories, etc., to be used for analysis [3-5].

III. DATA CLEANING

Data cleaning is the proper representation of data for analysts to work on by correcting problems with missing values, duplicate items, spelling errors, etc. Traditional data required proper data cleaning and then categorization, but with the help of big data techniques, we can now directly clean and categorize huge data. , data in zeta bytes!

1. Tools needed for data collection

A number of tools can be used for data collection and cleaning, including programming languages that are quite effective, the most widely used of which is python, a user-friendly, easy-to-understand programming language. The codes used in the language are very general in nature and therefore easy to learn. Other software used in this field is SQL, Matlab, IBM spss. Programming languages such as JAVA, R, Scala, and python are also used for big data mining, and sophisticated Hadoop software facilitates data collection [5-7].

- Data Architect – A data architect is required to create, deploy and maintain data for an organization.
- Data engineer - These are people who design software and programs to manage big data, collect it and store it in required formats.
- Data Manager – Manages stored data ready for analysis and controls data flow.

IV. WHAT IS BUSINESS INTELLIGENCE?

After the data has been collected and cleaned, analysts use various statistical techniques to establish relationships between various variables and analyze trends in business activities and ultimately find solutions to persistent problems. These numbers and data contain a huge amount of information that can be mined by the analyst, so they use other software with pre-installed functionality for easy and effective analysis and when they know about different trends in the business or how the area is increasing the number of users that can strategically invest in these areas. This is how business intelligence helps business. A more classic example for easier understanding is Consider a group of people shopping in 7 different stores of a company. Now the company wants to increase its sales to generate more profit. In order to do so, it needs to collect some data to place a digital rating where the user can rate their experience with the product. Now, a Business Intelligence analyst can provide a range of conclusions about these customers with just one additional piece of data. Recently, the role of artificial intelligence has emerged, which has proven its role in various sectors [8-9].

4.1 Customers are divided into groups of four

- Fans – These are people who shop only at this store and are satisfied with the services. So fans are people with high loyalty and also high satisfaction.
- Supporters - These are people who are loyal to the store, which means they continue to buy products from the store,

they are not satisfied with the services offered.

- Loiterers – These are people who are satisfied with the service but shop at multiple stores. They don't buy from a specific company.
- Alienated - last in the group are people who are alienated, i.e. these people neither buy much nor get much satisfaction from the product.

Now that the Business intelligence analyst gives you the numbers of these four groups, you can easily invest in the right way and increase sales. If the number of supporters is high, then the company needs to work on the services offered, store management, product price, etc.

If the number of roaming users was too high, the company would issue loyalty cards encouraging people to buy more, and these are the ways data science helps businesses grow and now plays a key role in the corporate world.

4.2 WHY USE BUSINESS INTELLIGENCE TOOLS?

First, uncovering information that was previously limited to the abilities of top investigative professionals is now something anyone can do with these tools. What's more, not only that, these devices give you the bits of knowledge you need to accomplish things like development, batch-critical solutions, gather all your information in one place, predict future outcomes, and so much more [7 ,9, 10].

Top Business Intelligence devices that will help you make the right choice.

- SAP Business Intelligence

Business Intelligence offers several advanced investigative measures, including continuous BI insightful investigation, AI, and disposition and investigation. Specifically, the Business Intelligence phase offers details and examination, information representation and investigation applications, office incorporation, and portable investigation. SAP is a vigorous programming designed for all jobs (IT, end users and executives) and offers a huge number of functions in a single phase.

MicroStrategy is a business intelligence tool that offers incredible (and fast) dashboarding and information exploration to help explore pattern, spot new opportunities, improve profitability, and the sky's the limit. You can usually access it from your workspace or using a portable device.

- Datapine

Datapine is an all-over BI phase that in any case supports the confusing procedure of investigating information for non-specialist clients. Due to the far-reaching approach of self-administrative research, datapine's answer enables information experts and business clients alike to effectively coordinate different information sources, conduct controlled information investigations, create intelligent business dashboards, and create significant business pieces of knowledge.

- Yellow-finned BI

Yellowfin BI is an end-to-end business insight and trial tool that combines representation, artificial intelligence and coordinated efforts. Likewise, you can effectively transfer huge amounts of information with natural too open dashboards virtually anywhere.

- QlikSense

QlikSense is the result of Qlik, an organization that is additionally known for another business intelligence tool called QlikView. QlikSense's user interface is advanced for the touch screen, making it a well-known bi device. The main difference with QlikView is the Storytelling element. Clients add their experience to the information, and by using images and features, making the right investigation and selection decisions has become significantly easier and better.

- Microsoft Power BI

Microsoft Power BI is a suite of electronic tools for business trials that exceeds expectations in information representation. It allows clients to gradually discern patterns and has fresh out of the box new connectors to improve your crusade game. Additionally, this product allows clients to coordinate their applications and deliver reports and constant dashboards.

V. MACHINE LEARNING (ML)

Machine learning is the part of science that makes a computer work without being programmed. Machine learning is a part of data science that deals with artificial intelligence. In a way, it helps make computers learn from data. Machine learning is not the same, it is different from the past. Both artificial intelligence (AI) and machine learning are often used interchangeably. Artificial intelligence is also a part of science that makes a computer act according to human tasks. Most machine learning algorithms are used in the development of artificial intelligence. These algorithms are used in various applications like email, filtering, etc. In our daily life, we go through many activities that are powered by machine learning, like recommendations on Netflix, YouTube, etc., and search engines like Google. Even voice assistants such as Google Assistant and Siri [10-13].

What is machine learning? ML is a part of artificial intelligence that allows computers to learn on their own without being explicitly programmed. The ML algorithm uses statistical data to calculate the output. The knowledge needed to create an effective machine learning system is

- The entire model
- Scalability
- Advanced algorithms along with the basics
- Ability to prepare data

VI. METHODS OF MACHINE LEARNING

These methods control different ways to train ML algorithms. We need to look at the kind of data it receives to know the pros and cons of each method. There are two types of data i.e. tagged and untagged data.

- Unlabeled data has no parameters to process. It is processed in a machine-readable pattern. There are three methods in machine learning

1. Supervised learning

Supervised learning is the simplest form of machine learning. It is considered a paradigm of machine learning. It is essentially a task-driven process. First, it needs an example with a labeled dataset to work with. Then load the algorithm with one sample data set at a time and see if the prediction made by the system is correct or not. Follow this procedure, meanwhile the system will start looking for the relationship between examples and labels. After processing the data, the algorithm gets an idea of how the system works and infers the relationship between input and output. Common applications are

- Popularity of advertisements – the selection of desired advertisements is based on this algorithm. Ads that are

present on Google while browsing are caused by this algorithm.

- Facial recognition – Facebook uses this algorithm to recognize your face. If you have a system that takes pictures, then it works based on supervised learning.

2. **Unsupervised learning**
Unsupervised learning is the opposite of supervised learning. Uses untagged data. It is a data-driven process. Because it does not have labeled data, the system processes hidden structures. These hidden structures make unsupervised learning versatile. It can adapt to any data by changing the structure. A system where you can see unsupervised learning are

- Recommendation System - The recommendations provided by Netflix, YouTube and many other sites are based on
unsupervised learning.

3. Reinforcement learning
Reinforcement learning learns from mistakes in the same way that people learn data from the mistakes they routinely make. It uses a trial and error method. It does not use both tagged and untagged data. It's behavior-driven learning, making a lot of mistakes at first and learning from them. Over time, the system corrects the error and makes fewer errors than before [10-13].

VI. CONCLUSION

Data science education is well into its formative stages of development; it is evolving into a self-supporting discipline and producing professionals with distinct and complementary skills relative to professionals in the computer, information, and statistical sciences. However, regardless of its potential eventual disciplinary status, the evidence points to robust growth of data science education that will indelibly shape the undergraduate students of the future. In fact, fueled by growing student interest and industry demand, data science education will likely become a staple of the undergraduate experience. There will be an increase in the number of students majoring, minoring, earning certificates, or just taking courses in data science as the value of data skills becomes even more widely recognized. The adoption of a general education requirement in data science for all undergraduates will endow future generations of students with the basic understanding of data science that they need to become responsible citizens. Continuing education programs such as data science boot camps, career accelerators, summer schools, and incubators will provide another stream of talent. This constitutes the emerging watershed of data science education that feeds multiple streams of generalists and specialists in society; citizens are empowered by their basic skills to examine, interpret, and draw value from data.

REFERENCES

1. Nadikattu, Rahul Reddy, Research on Data Science, Data Analytics and Big Data (17 April 2020). INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND - Volume 9, Issue 5, May 2020 Pages: 99-105.. Available at SSRN: https://ssrn.com/abstract=3622844 or http://dx.doi.org/10.2139 /ssrn.3622844
2. K. Terao, Machine learning synthetic data, scanning probe data and reciprocal spatial data on quantum materials. (2019).
3. V. Setlur and M. Tory, Exploring synergies between visual analytic flow and linguistic pragmatics. AAAI Spring Symposium. (2017).
4. L.A. Enneking, Using Data Collection Activities in the Middle School Mathematics Classroom. (2008).
5. P.W.Group and G. Garrett, The Data Engineering Project (Educating for the Future PhUSE Working Group). (2019).
6. Nadikattu, Rahul Reddy, Data Warehouse Architecture - Leading the Next Generation Data Science (11 September 2019). Rahul Reddy Nadikattu "Data Warehouse Architecture - Leading the Next Generation of Data Science" International Journal of Computer Trends and Technology 67.9 (2019):78-80.. Available at SSRN: https://ssrn.com/abstract=3622840 or http: / /dx.doi.org/10.2139/ssrn.3622840
7. J.M .Fernández and A. Valencia, XML databases, are they ready for bioinformatics? Spanish Bioinformatics Conference. (2004).
8. M.Cox, S.F. Austin and A.B. Gresham, The Role of Customer Service in Small Business Strategic Planning. (1997).
9. P. Khatri, Emergence of AI through Machine Learning and Data Science. The Journal of Innovations, 14. (2019).
10. D.C.Desai, C.Dhanasekaran, A.Narayanapur and V.Joshi, (Social Media and Multimedia Data Analytics through Machine Learning. (2017).

11. MODELING A NEW WORK PROCEDURE BASED ON EMOTIONAL ANALYSIS OF FLOOR PLANS USING MACHINE LEARNING ALGORITHMS AND SEMIOTICS. (2020).
12. B.Geluvaraj, P.M .Satwik and T.A. Kumar, The Future of Cyber Security: Major Roles of Artificial Intelligence, Machine Learning and Deep Learning in Cyberspace. (2019).
13. S.K. Vishwakarma, Machine Learning

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details