International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 07 | July - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Data and Data Analysis for Beginners

Udit Aditya Chaturvedi

Abstract-

Big data is a new driver of the world economic and societal changes. The world's data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education. While the data complexities are increasing including data's **volume**, **variety**, **velocity** and **veracity**, the real impact hinges on our ability to uncover the **value** in the data through *Data Analytics* technologies. *Data Analytics* poses a grand challenge on the design of highly scalable algorithms and systems to integrate the data and uncover large hidden values from datasets that are diverse, complex, and of a massive scale. Potential breakthroughs include new algorithms, methodologies, systems and applications in *Data Analytics* that discover useful and hidden knowledge from the *Data* efficiently and effectively.

I am trying to make the *Data* and *Data Analysis* simple to understand since the field is already impacting the civilisation. As the technology grows, steps linked to *Data* and *Data Analysis* are taking part for example Artificial Intelligence and Machine Learning. In this paper, we will mainly focus on *Data* and *Data Analysis*.

Introduction-

This is my first paper on *Data* and *Data* Analysis which sounds quite complex and I thought of coming out with a simple version for everybody to understand. I am a youngster having recently completed my engineering studies. My interest in the subject developed when I asked my father about how Google Maps indicate about travel times, traffic congestions and their prediction about the time taken for a travel is so good. His response didn't satisfy me but I wanted to get an answer so I read about it. I found the answer quite interesting developing my keenness to know more about *Data*. I studied various papers on *Data* and following is the summary of my understanding.

What is Data-

Let us begin with understanding *Data*, is it so new OR we have started calling it *Data* after so much improvement in technology and IT. I feel, *Data* has been in use for a long time. For example, if you talk about the Doctor's profession. They detect the illness based on *Data* which is body temperature, feedback from the person suffering, heartbeat, blood pressure and other pathological details. If you talk about admissions to prestigious institutes, the same is based on *Data* like the student's marks, his project details etc. So the concept of *Data* is not new.

Let us try to understand what is *Data* and *Intelligent Data*, another term we use very frequently. *Data* is any information or event or activity or change of state. The same will be of interest to the extent of improving the general knowledge to everybody OR to an individual or a set of individuals, it can be "intelligent data" because it is of higher consequence to him or them. It will be easily understood if I give a few examples.

For example, if we are informed that Mumbai roads are normally choked during office hours, it is an information. The same information becomes intelligent for the people going to office because then they have to trace the time expected to be taken for the travel. For a person, taking someone to the hospital, he also has to know as to which route has to be taken to reach fastest. So depending on the circumstances, the same *data* can become *intelligent data*.

Second example would be…if we are informed that India is a country with highest population, it is an information. For the people, like in administration, responsible for people's accommodation, education, employment, travel, food, power etc., the information becomes *intelligent data* with lot of work to be done with respect to the infrastructure availability to sustain this growth.

Next main thing is how to get *data*…? This simply depends on the application of the *data*. *Data* is being continuously generated and one key thing is to store the *data* in data files or data servers. For example, the traffic *data* is given for commuters and it comes largely by

- Historic time duration for identified travel at a defined time.
- Satellite information regarding position tracking of the vehicle
- Continuous information of position tracking of vehicles ahead in traffic.

Similarly, *Data* are available for crime, poverty, education, population, health, travel, economy, environment, medicines and their effects OR about anything which will impact the civilisation personally or as a group.

Data Analysis-

Why *Data analysis* has gained prominence now is because of many scientific developments like internet, mobility, satellites and communication and simultaneous improvements in our knowledge to reduce dependence on man-hours but going forward based on *data*. So it has a large impact on business. And why do we find most *data analysis* to be so good or high in accuracy, simply because it is based on mathematics, largely statistics and probability.

Data Analysis is a work done by people who understand that *data*, inference from the *data* and next steps offering solution if the *data* indicates a challenge. Data-driven Decision Making(DDDM) can be defined as the process of making correct decisions based on facts, data or metrics instead of intuition or observations. So a *data* scientist must be a champion of his field.

It can be better explained by following examples.

Example1- An example of data-driven decision-making is using digital intelligence tools to look at existing demand in a market for a specific product or service before deciding to enter it.

Example2- If a weatherman senses the progress of *data* like temperature, wind pressure and humidity in the environment, he should be able to announce the weather. For example, if it is going to rain, if yes, how much will be the rain etc..

Data analysis process-

As the *data* availability continues to grow both in amount and complexity, so too does the need for an effective and efficient process by which to harness the value of that *data*. The *data analysis* process typically moves through several iterative phases. Let's take a closer look at each:

- **Identify** the questions you'd like to answer. What problem is the company trying to solve? What do you need to measure, and how will you measure it? *Data* is necessary and is served as inputs for analysis, which is specified based upon the requirements.

- **Collect** the raw *data* sets you'll need to help you answer the identified question. *Data* collection might come from internal sources, like a company's client relationship management (CRM) software, or from secondary sources, like government records or social media application programming interfaces (APIs). The *data* may also be collected from sensors in the environment like traffic cameras, satellites, recording devices etc..

- **Clean** the *data* to prepare it for analysis. This often involves purging duplicate and anomalous *data*, reconciling inconsistencies, standardizing *data* structure and format, and dealing with white spaces and other syntax errors. For example; with financial information, the totals for particular variables may be compared against separately published numbers that are believed to be reliable.

- **Analyze** the *data*. By manipulating the *data* using various data analysis techniques and tools, you can begin to find trends, correlations, outliers, and variations that tell a story. During this stage, you might use *data* mining to discover patterns within databases or data visualization software to help transform data into an easy-to-understand graphical format.

- **Interpret** the results of your analysis to see how well the *data* answered your original question. What recommendations can you make based on the data? What are the limitations to your conclusions?

Types of Data Analysis-

Descriptive Analysis- This type of analysis helps describe or summarize quantitative data by presenting statistics. It gives an overview of *data* and are usually shown in tables, graphs, summary statistics etc.. For example, descriptive statistical analysis could show the distribution of sales across a group of employees and the average sales figure per employee. It answers the question- '**What Happened'.**

Diagnostic Analysis- Let's say if the descriptive analysis shows an unusual influx of patients, drilling down the *data* further may show that many of these patients shared symptoms of a particular virus. It answers the question- **'Why did it happen'**.

Predictive Analysis- this type of analysis uses *data* to form predictions about the future. It utilizes historical and current facts to reach future predictions. For example, you may notice that a given product had its best sales during the months of July and August each year, leading you to predict a similar high point in the upcoming year. It answers the question- **'What might Happen in the Future'**.

Prescriptive Analysis- This type of analysis takes insights from all types of analysis mentioned above to form recommendations. It answers the question- **'What should we Do about it'**.

Inference Analysis- This type of analysis uses a small sample to draw conclusions about a large population. It uses statistical models and probability theory to estimate population parameters and test population hypotheses based on sample *data*. For example, you select a random group of 11th graders in your state and collect data of their SAT scores. You can use inferential analysis to estimate and make hypothesis about whole population of 11th graders of that state based on sample *data*.

Applications of Data Analytics in Buisness-

As mentioned earlier, *data* are used by all of us in person or in Groups. Some examples are being given here from the field of business and the significance of *data* in business…

Education Industry-

Education systems have huge records of *data* about students, faculty, courses etc. *data analysis* can help provide better decisions in various sectors which will result into effective operation and improve performance overall.
Few Areas where *Data Analysis* has helped in bringing a change are:

- Customized and Dynamic Learning Programs: For effective learning, the learning programs are customized on the basis of student's history. This in return also improves results of students.
- Reframing course Material: A course is designed keeping in mind the only objective is the student should learn and gain knowledge. With the help of student's *data* one can closely observe what and how effectively a student is learning. Hence, *data analysis* can help in modifying the course material for better learning.
- Career Prediction: Proper analysis of student's *data* reveals the true interest, areas of weakness and strengths of the student. Therefore, one can help in predicting appropriate career choices for a student.

Health Care Industry-

The healthcare sector is bound to keep records of patients, doctors as well data about different kind of diseases. *Data analytics* help in discovering new drug development methods as well. This will automatically generate volumes of *data* and proper and effective analysis of this *data* has helped us in following ways:

- Effective analysis of *data* about diseases helps us to efficiently diagnose any disease resulting into early treatment.
- Proper analysis of previously available *data* can help in identifying epidemics as well as controlling them in early stages, saving more and more lives.

Media and Entertainment Industry-

Some of the ways where *Data analysis* helps are:

- Predicting the interests of audiences.
- Optimized or on-demand scheduling of media streams in digital media distribution platforms.
- Getting insights from customer reviews.
- Effective targeting of the advertisements.

Transportation Industry-

Data Analysis has helped the transportation industry to find traffic patterns, to plan routes efficiently and to identify accident prone areas. In order to identify traffic patterns during certain hours of the day, data analysis has helped to manage congestion and traffic to avoid wastage of time. This helps to ensure safe travel. Travel by aggregators which can be good for travel by ladies alone by tracking the vehicle movements or by location tracking of criminals by tracing their mobiles.

Banking sector-

The banking sector in today's world is dynamic in nature as its *data* changes every second. Also, *data* is added and modified with every passing millisecond. Therefore, *data analysis* and predictive modelling helps in securing this *data* from any illegal activities like money laundering, misuse of credit/debit cards. It also helps in predicting change in future in trends and activities. We have many examples of successful use of *data* and have facilities like Net Banking and Bill Payments, Digital Payments used for a large number of transactions

Manufacturing Industry-

If it is with respect to equipment service, for example, for a motor, the *data* can be with respect to temperature of the windings and the stator, for a breaker, it can also include number of short circuit clearances. And for this measurement, the manufacturer gives the instruments for continuous monitoring. Having such *data* helps in predicting failures and improves service for the equipment.

Service Industry-

There are products being designed now to simplify the prediction of failure. Take an example of a motor where failure is anticipated by tracking temperature increase in windings and stator and the vibrations. These measurements come from sensors mounted on the motor.

Data Risk and Data Security-

There is a big concern on safety of *data*, *data* tracing, *data* misuse and *data* security. There have been frauds in money mainly due to two things. Improper/fake tool used or by mistake of the individual. Big issue is of data security and personal *Data* are a point of concern. The information obtained using data analytics can also be misused against a group of people of certain country or community. Some of the analytics tools developed by companies are more like a black box model. Nobody understands the logic, the system uses to learn from *data*. These hidden complexities and biases could hamper the ability of the system to make correct decisions.

Another negative (?) point is that it gives out what an individual was doing or where he was, doing what…? These, sometimes, can be tricky and give a concern to an individual that he is being tracked continuously.

Next Steps to Data Analysis-

Along with *Data*, there has been remarkable research and development of technology like Artificial Intelligence and Machine Learning. In this research paper, we are not covering Artificial Intelligence and Machine Learning in detail but are covering the same for brief introduction.

Today we have AI tools like ChatGPT which capture information from multiple publications on the net and come out with a good answer, sometimes better than for one human being. Worldwide, with the development of artificial intelligence, many jobs are expected to be redundant for example for Doctors, Lawyers etc. and repetitive work in the factories because we have many Digital Factories now coming up with enhanced production and higher testing during production.

But I feel these are still early days despite a good development. There have been many wrong outcomes from ChatGPT. I read about a case in US where ChatGPT provided fake information regarding earlier cases. It was quoted by the lawyer without verification and he was reprimanded by the judge. The field, I would say, is very exciting but we need to develop such tools carefully and be cautious about being thrown off guard by the outputs.

Conclusion-

I hope I have captured the details in simple English for beginners. This review paper aims to increase the level of awareness of intellectual and technical issues surrounding the analysis of massive data. Recent years have seen rapid growth in computing systems to serve as backbone of the modern internet based information ecosystem. It is important to acknowledge that the goals of massive *data analysis* go beyond computational and representational issues that have been the province of classical search engines and database processing to tacking the challenges of statistical inference, where the goal is to turn *data* into knowledge and to support effective decision making.

References-

1. https://www.rfwireless-world.com/Terminology/Advantages-and-Disadvantages-of-Data-Analytics.html
2. https://chartio.com/learn/data-analytics/types-of-data-analysis/
3. https://www.simplilearn.com/tutorials/data-analytics-tutorial/what-is-data-analytics
4. https://www.scribbr.com/statistics/inferential-statistics
5. Big Data Analytics: A Literature Review Paper- Nada Elgendy and Ahmed Elragal
6. Big Data Analytics: Applications, Challenges & Future Directions- Tanya Garg and Surbhi Khulla
7. Big Data Analytics- Sachchidanand Singh and Nirmala Singh
8. Different Types of Data Analysis; Data Analysis Methods and Techniques in Research Projects- Hamed Taherdoost Research Club, Research and Development Department.