International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 878-883 www.ijaem.net ISSN: 2395-5252

Eminent Tools of Data Science: Its Prosperities and Applications

Dr. G. Rajitha Devi
Asst. Prof. in Computer Science

Submitted: 05-04-2022 Revised: 16-04-2022 Accepted: 19-04-2022

ABSTRACT: Data Science has always been about data. Monitoring data have served as the confirmation that has allowed us assess, accept, and discard scientific theories. But in the last decades, scientific c data has grown fast in both size and importance. Data Science is therefore not a new science direction, but rather a new pair of glasses – a new prototype– to look at problems and questions in the existing disciplines with the new probabilities of data analytics in The intellectual algorithms or data operations in the data science make the data to be more effective in decision making and decision polices. We also focus on how data science integrate mathematical & statistical methods, logical reasoning with applications of Artificial Intelligence techniques. We also focus on various data operations tools which exist in the python, SAS and many others.

KEY WORDS: Data science, Big Data, Open source Tools, Machine Learning, Artificial Intelligence

I. INTRODUCTION

A. Artificial Intelligence and its relevance with Data Science:

Artificial Intelligence eloquent about how to make the system as intelligent like a human being. Designing intelligent system is comprehensible by incorporate the computers with learning, processing and decision making potentiality. All these abilities deal with vast knowledge which helps the system to train with intelligent behaviour. A.I speaks about innumerable approaches of learning, understanding and processing techniques which can be applied on various problems or domains.

The most popular A.I techniques are Heuristics, Support Vector Machines, Artificial Neural Networks, and Markov Decision Process. Artificial Intelligence is well known for its applications like natural language processing, data retrieval by using intelligent systems, expert systems for various domains, theorem proving & game playing, Scheduling and combinatorial problems , robotics and so on. Know question rises how the A.I is related to data science, as almost all humans' beings uses the data for their wide variety of applications in day to day life.

B.Machine Learning relevance with Data Science

To develop the Machine learning (M.L.) we need to learn the past developments of over six decades in 1950 are where Alan Turing initiates with an idea of machine computing and intelligence. M.L. is considered as subset, practical approach and application of A.I based algorithms. As the name implies machine deals with wide variety of data of various domains and design the system. This system will be able to identify the train the new set of data with the existing data samples or derive the new set of rules. Unlike algorithms to make the machine as well organized such as supervised, unsupervised and semi-supervised and reinforced algorithms.There are numerous techniques proposed by M.L like game analytics, software, voice recognition, stock trading, and internet of things (I.O.T's). The data science plays an important role by providing the data in good means to have effective M.L algorithms. Machine learning techniques are used to routinely find the recognize primary patterns inside complex data that we would otherwise brawl to determine innovations in the areas like applied computing, medical sciences, professionals & social life activities, computing paradigms, Data management systems and many more to have a better decision making. Influence the new methods of improving intellectual thinking of how to use, organize, process, load, and model or visualize the data. The emerging existing professions in the field of data science is the data scientist who draws an medium salary of $124,00 and stated this profession may be on the peak of in the coming

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 878-883 www.ijaem.net ISSN: 2395-5252

years. The tool selection to implement the data science activities like we have SAS, Python, Rapid Miner, Data Analysis Tools and Data visualization.

II. METHODOLOGY

For the deep study, a review approach was adopted. This helped us to combine the existing work and identify our strategy in comprehensive manner. We made an in-depth study of Data Science tools, revealing their benefits, challenges and applications. The prime focus of our study is to provide extensive knowledge to all users to decide which tools are better to meet their need.

Literature Survey

Data science is a multi-disciplinary field that include other scientific fields such as Statistics, Mathematics, Artificial Intelligence, Data Mining and Machine learning to extricate knowledge from structured and unstructured data.

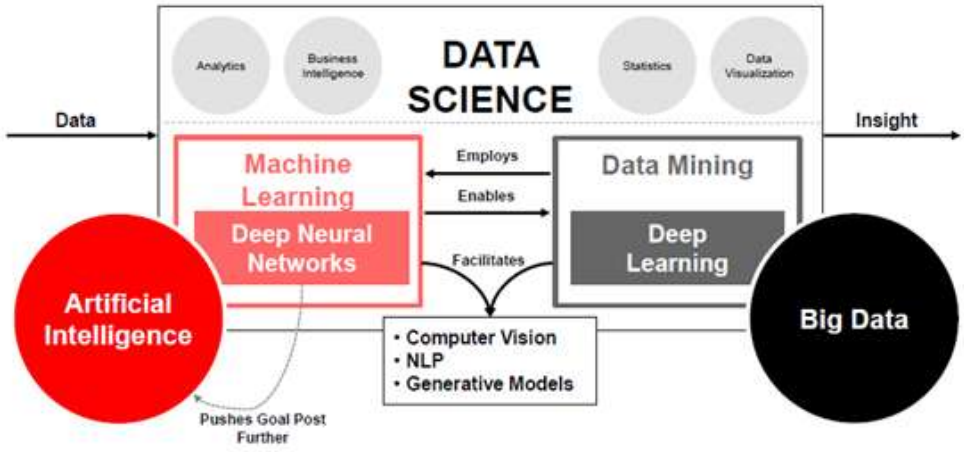

Fig. 1 Flow of Data Sciences in which involve several steps including Analysis, Business understanding, and Data visualisation to deal with Big data by using Machine learning techniques.

The term data science is introduced by John Turkey for the reformation of statistics academia, more than 50 years earlier. He pointed to the unrecognised science that is based on data science and its subject is to learn the data.

Thus using this concept Bill Cleveland, John Chambers, and Leo Breiman again worked on statistics academia and expanded its boundaries from classical-statistics to theoretical statistics. Chamber focused on data preparation and presentation rather than data modelling. Similarly, the Breiman concept based on prediction instead of inference, and finally, Cleveland suggested the name "Data Science" for fulfilling his vision.

Data science is a vast field in which is further divided into different sub-field like artificial intelligence, data mining, deep learning and machine learning that are used to extract significant data from inconsequential data and also helps to create different statistical graphs and patterns easily. The primary operations which can be performed in the data science like cleaning raw data, loading the data at the server side, process data, visualize data and acquire data by various stake holders. Detailed explanation on techniques of data requirements, data analysis, data processing, visualize or model data.

1.Python

Python is developed by Guido van Rossum in the 1980s it is an open source, interpreted and high-level languages. Specially codes for Python are written in a simple, easy format and comprehensively used for all kinds of scripts. Over the past years, it becomes popular in the field of data science. It has also many libraries for performing different tasks for different domains like Scipy, Pandas, StatsModels for Statistics. Similarly, Seaborn and Bokeh it is used for visualisation. Python is very popular among the computational scientists, and its use also increases with the passage of time. The major advantage of the python is an object-oriented language, easy to understand, readable and supporting multiple platforms. Besides this, it is slow and difficult to integrate with another language.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 878-883 www.ijaem.net ISSN: 2395-5252

2. Rapid Miner

Rapid Miner is a data science platform developed mainly for non-programmers and researchers for quick analysis of data. The user has an idea in their mind, and easily creates processes, import data into them, run them over and throw a prophency model. The tool supports importing ML models as well to web apps like flask or nodeJS, android, iOS, and more, thereby unifying the entire spectrum of the Big Data Analytics Lifecycle.

The strength of the tool is that it clarify the scattered tasks of data mining and analysis. The tool loads data from various frameworks like Hadoop, Cloud, RDBMS, NoSQL, pdf, and many more. Then it pre-processes and prepares data using standard industrial methods by grouping items by categories or spawning new child tables or join tables or interpolating missing data. Further, it trains AI models as well as optimal deep learning models such as Random Forests, XGBoots, Gradient Boost, and more, or clustering or pruning outliers, even visualizing outputs. Finally, the models are deployed on the cloud or in the production environment. The software only requires to create user interfaces for the collection of real-time data and execute it on a real model to serve a task.

3. Data Analysis Tools

Data Analysis is the process of cleaning, modeling, and transforming data to discover useful information or patterns for business decision-making. The data analytics consists of various operations on the data sets or tables available in databases. The operations include data extraction, data profiling, data cleansing and data deduping, and more. There are several methods and techniques for data analysis based on business and technology. The major types of data analysis are:

- **Text Analysis:** It is the method to discover patterns in large data sets using databases or data

Mining tools. The process converts raw into useful business information.

- **Statistical Analysis:** This analysis includes past data for analysis. It is of two types:
  - **Descriptive Analysis:** shows mean and deviation for continuous data, whereas percentage and frequency for categorical data.
  - **Inferential Analysis:** In this user can find different conclusions from the same data by selecting different samples.
- **Diagnostic Analysis:** Diagnostic analysis finds the cause of the insights found in Statistical Analysis.
- **Predictive Analysis:** It predicts future insights based on previous data.
- **Prescriptive Analysis:** Data-driven companies most use this type of analysis technique. The analysis combines the insights from all previous analyses to decide on the current problem or decision.

4. Data Visualization Tools

Data visualization refers to the representation of the data in a pictorial or graphical format. Its purpose is to provide decision-makers to check analytics visually to see patterns and grasp difficult concepts. Data visualization pulls data from various disciplines, including scientific visualization, information graphics, and statistical graphics. Various approaches can achieve data visualization, a popular one being the Information Presentation, which includes statistical graphics and thematic cartography.

Data visualization tools display information in a sophisticated way such as infographics, dials and gauges, geographic maps, sparklines, heat maps, and full bar, pie, and fever charts. The visualization tool is essential in analytics, demonstrating data and making data-driven insights available to workers throughout an organization. Data visualization software plays a vital role in big data and advanced analytics projects, as well. As businesses accumulate massive troves of data during the early years of the big data trend, they need a way to quickly and easily get an overview of their data, and visualization tools prove to be a natural fit in this case. It is essential to visualize the outputs to monitor results and ensure that models are performing as intended when writing advanced predictive analysis using machine learning algorithms because it is easier to interpret visualizations of complex algorithms than to interpret numerical outputs.

5. SAS

SAS is a statistical software tool developed for advanced analytics, business intelligence, data management, a criminal investigation, predictive analysis, and data visualization.

Key features that SAS offers for visual analytics are:

- **Interactive dashboards, reports, BI and analytics:** The tool allows the user to go

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 4 Apr 2022, pp: 878-883 www.ijaem.net ISSN: 2395-5252

directly from reporting and exploration, to analysis, to sharing information through different channels with a single interface.

- **Smart Visualization:** The software compellingly presents data and results with advanced data visualization techniques and guided analysis through auto charting.
- **Location Analytics:** Combines traditional data sources with location data for analysis in a geographical context.
- **Augmented Analytics:** Reveals real stories hidden in your data within a few seconds. Automatically shows the user suggestions and identifies related measures.
- **Self-Service Analytics:** Automated forecasting, goal seeking, scenario analysis, decision trees, and more are at your fingertips, no matter what skill level the user has.
- **Text Analytics:** The tool enables us to gain insights from social media and other text data, and know whether the view is positive or negative.
- **Self-Service Data Preparation:** The user can import their data, join tables, apply essential data quality functions, and more with secure drag-and-drop capabilities.

III. DATA SCIENCE USES AND APPLICATIONS

As discussed in Data Science has conquered maximum all the organizations of the globe today. There is no such business across the globe which does not use data to improve their organizations. As such, data science has become important aid for organizations to make effective use of data. There are various organizations like banking, financial institutions, automations and engineering, conveyance, e-commerce, edification sectors, etc. that use data science.

A.Image Recognition

You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. Similarly, while using whatsapp web, you scan a barcode in your web browser using your mobile phone. In addition, Google provides you the option to search for images by uploading them. It uses image recognition and provides related search results.

B. Role of Data science in Banking or Financial institutions

Banking is one of the leading sectors which can make use of beneficiaries or customers data in effective means. These institutions can make better decisions and predict future preventions of frauds in an intelligent way. Management of customer's data involves much analytical, statistical, mathematical reasoning incorporated with A.I techniques, or algorithms, deep learning and machine learning. This also supports in maintaining customer, **predicting the plans** according to their usage & savings, investment plans and so on.

C.Gaming

EA Sports, Zynga, Sony, Nintendo, Activision-Blizzard have led gaming experience to the next level using data science. Games are now designed using machine learning algorithms which improve / upgrade themselves as the player moves up to a higher level. In motion gaming also, your opponent (computer) analyzes your previous moves and accordingly shapes up its game.

D.Price Comparison Websites

At a basic level, these websites are being driven by lots and lots of data which is fetched using APIs and RSS Feeds. If you have ever used these websites, you would know, the convenience of comparing the price of a product from multiple vendors at one place. PriceGrabber, PriceRunner, Junglee, Shopzilla, DealTime are some examples of price comparison websites. Now a days, price comparison website can be found in almost every domain such as technology, hospitality, automobiles, durables, apparels etc.

E.Airline Route Planning

Airline Industry across the world is known to bear heavy losses. Except a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air fuel prices and need to offer heavy discounts to customers has further made the situation worse. It wasn't for long when airlines companies started using data science to identify the strategic areas of improvements. Now using data science, the airline companies can:

1. Predict flight delay
2. Decide which class of airplanes to buy
3. Whether to directly land at the destination, or take a halt in between (For example: A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)
4. Effectively drive customer loyalty programs

Southwest Airlines, Alaska Airlines are among the top companies who've embraced data science to bring changes in their way of working.

F.Fraud and Risk Detection

One of the first applications of data science originated from Finance discipline. Companies were fed up of bad debts and losses every year. However, they had a lot of data which

use to get collected during the initial paper work while sanctioning loans. They decided to bring in data science practices in order to rescue them out of losses. Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures and other essential.variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

G.Coming Up In Future

Though, not much has been reveled about them except the paradigms, and neither i know when they would be available for a common man's disposal. Hence, I've kept these amazing application of data science in 'Coming Up' section. We need to wait and watch how far Google can become successful in their self driving cars project. Robots, as we know, have lived for a while but aren't being used as a commodity yet due to related security issues.

IV. CONCLUSION

In this paper, we have discussed the various tools of Data Science. Open source tools and other tools are discussed with their benefits, challenges and applications. But no doubt, the open source tools such as Python, SAS, Python, Rapid Miner, Data Analysis Tools and Data visualization. Currently aforementioned companies are using open source tools, Python and R-Programming. Python and R-Programing are free for every one and could be used for data modelling, data visualisation and a lot of other applications. Data science becomes as a mandatory field which coordinates between multi disciplines like mathematics, statistical approaches, mathematical methods, logical reasoning, intelligence algorithms and machine learning practical's. All these fields correlate to access the data from various business or organizations and make use of them in effective means. These effective use of data leads to perform proper decision making to grow business further on the basis of customer chooses and satisfaction. At last we focus on how successful carriers can be built in the field of data science. The main vision of this field it used to grow all businesses.

REFERENCES

[1]. McKinsey Global Institute. The Age of Analytics: Competing in a Data-Driven World. Research Report.(under "Our Research", "Technology and Innovation"). 2016; p. 136. Available online: www.mckinsey.com/mgi (accessed on 18 June 2018).

[2]. Mahabal, A.A.; Crichton, D.; Djorgovski, S.G.; Law, E.; Hughes, J.S. From sky to earth:Data Science methodology transfer. In Proceedings of the International Astronomical Union, Sydney, Australia,17 July 2017; Brescia, M., Djorgovski, S.G., Feigelson, E., Long, G., Cavuoti, S., Eds.; Cambridge University Press: Cambridge, UK, 2017; pp. 17–26. Available online: https://arxiv.org/pdf/1701.01775.pdf (accessed on 18 June 2018).

[3]. Murtagh, F. Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics; Chapman & Hall/CRC Press: Boca Raton, FL, USA, 2017.

[4]. Hayashi, C. What is Data Science? Fundamental concepts and a heuristic example. In Data Science,Classification, and Related Methods; Hayashi, C., Yajima, K., Bock, H.H., Ohsumi, N., Tanaka, Y., Baba, Y., Eds.;Springer: Heidelberg, Germany, 1998; pp. 40–51

[5]. Ohsumi, N. Fromdata analysis to data science. In Data Analysis, Classification, and RelatedMethods; Kiers, H.A.L.,Rasson, J.-P., Groenen, P.J.F., Schader,M., Eds.; Springer: Heidelberg, Germany, 2000; pp. 329–334.

[6]. Escoufier, Y.; Fichet, B.; Lebart, L.; Hayashi, C.; Ohsumi, N.; Baba, Y. (Ed.) Data Scienceand Its Applications;Academic Press: Tokyo, Japan, 1995.

[7]. Cao, L. Data science: A comprehensive overview. ACM Comput. Surv. **2017**, 50, 43:1–43:42. [CrossRef]

[8]. Ueno, M. As the oldest journal of Data Science. Behaviormetrika **2017**, 44, 1–2. CrossRef]

[9]. Englmeier, K.; Murtagh, F. Data Scientist—Manager of the discovery lifecycle. In Proceedings of the 6th International Conference on Data Science, Technology and Applications—Volume 1: DATA, Madrid, Spain,26–28 July 2017; pp. 133–140.

[10]. Coombs, C.H. A Theory of Data; Wiley: Hoboken, NJ, USA, 1964.

[11]. Japec, L.; Kreuter, F.; Berg, M.; Biemer, P.; Decker, P.; Lampe, C.; Lane, J.; O'Neil, C.; Usher, A. AAPOR Report on Big Data; Technical Report; American Association for Public Opinion Research (AAPOR):Oakbrook Terrace, IL, USA, 2015; 50p. Available online: http://www.aapor.org/Education-

Resources/Reports/Big-Data.aspx (accessed on 18 June 2018).
[12]. Abbany, Z. A Public Transport Model Built on Open Data, News Article. Available online: http://www.dw. com/en/a-public-transport-model-built-on-open-data/a-41546053 (accessed on 27 November 2017).
[13]. Darabi, A. The UK's Next Census Will Be Its Last—Here's Why, News Report. Vailable online: https://apolitical.co/solution_article/uks-next-census-will-last-heres (accessed on 5 December 2017).
[14]. Murtagh, F.; Orlov, M.; Mirkin, B. Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research. J. Classif. **2018**, 35, 5–28. [CrossRef]
[15]. Hand, D. Statistical challenges of administrative and transaction data. J. R. Stat. Soc. Ser.A **2018**, 181, 1–24.[CrossRef]
[16]. Anderson, C. The End of Theory: The Data Deluge Makes The Scientific Method Obsolete, Wired Magazine.Available online: http://www.wired.com/science/discoveries/ magazine/16-07/pb-theory (accessed on 16 July 2008).
[17]. Murtagh, F. Origins of modern data analysis linked to the beginnings and early development of computer science and information engineering. Electron. J. Hist. Probab. Stat. **2008**, 4, 26.
[18]. Englmeier, K.; Murtagh, F. What can we expect from data scientists? J. Theor. Appl. Electron. Commer. Res.**2017**, 12, i–iv.
[19]. Murtagh, F.; Farid, M. Contextualizing Geometric Data Analysis and Related Data Analytics: A Virtual Microscope for Big Data Analytics. J. Interdiscip. Methodol. Issues Sci. Spec. Issue Digit. Contex. **2017**, 3, 1–19 .
[20]. Allin, P.; Hand, D.J. New statistics for old?—Measuring the wellbeing of the UK. J. R. Stat. Soc. A **2017**, 180,3–43. [CrossRef]
[21]. Wessel M. You Don't Need Big Data—You Need the Right Data. Harvard Business Review, 3 November 2016. Available online: https://hbr.org/2016/11/you-dont-need- big-data-you-need-the-right-data (accessed on 18 June 2018).
[22]. Jobs Rated Report 2017: Ranking 200 Jobs. Available online: https://www.careercast.com/jobs-rated/2017-jobs-rated-report (accesssed on 18 June 2018).