IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 5.731

Volume 12, Issue 1, December 2021

Future Trends in Data Science

Yaasmin Attarwala[1] and Sakshi Baid[2]
Students, Diploma in Computer Science Engineering
SVKM's Shri Bhagubhai Mafatlal Polytechnic, Mumbai, Maharashtra, India
yaasattar163@gmail.com and sakshibaid4@gmail.com

Abstract: *With progression in technology, an enormous magnitude of information being collected from digital users by various businesses and organizations, has resulted in formation of huge data repositories commonly known by the term Big data. Data mining is a tool used for extracting hidden information from these vast databases to identify unique patterns and rules. The present paper aims to provide a detailed description of the importance of big data in today's times, its characteristics, how data mining plays an important role in big data, why it is a necessity in today's times, the process of data mining and functionalities it performs, data mining techniques such as classification, clustering etc. that help in finding the patterns to decide upon the future trends in businesses and applications of the same in various fields. The paper also discusses the important role of data mining in Business Intelligence (BI) and various industries, to identify unique patterns and obtain results from the data along with the second half of the paper focusing on further exploring the challenges that are faced in big data and tools used, the applications and upcoming trends in data science and lastly, the scope and importance of data science in the future.*

Keywords: Big data; Data mining; Business Intelligence; Data repositories; Data analysis; Data Science; Clustering, classification techniques;

I. Introduction

As we look around these days, it appears that every- thing is becoming digitalized. We go around with our GPS-enabled smart phones, continuously sharing and exchanging information and ideas on social media platforms such as Instagram, YouTube, and Twitter, and we conduct a set number of online transactions each day. The main point is that it's happening all over the world, and anything we do that involves a digital activity leaves a digital trace. This is happening not only on a human level, but we can observe large volumes of data being captured and relayed via sensor-equipped industrial machinery and plants from a bird's eye view. Big Data means massive data that is highly complex and rapidly increasing day by day due to large use of technologies. It is a term applied to datasets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Artificial intelligence (AI), mobile, social and the Internet of Things (IoT) are driving data complexity through new forms and sources of data. For example, big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media much of it generated in real time and at a very large scale. It is a challenging task to extract the relevant information from such a complex huge data using traditional methodology. Thus, this is where data mining comes into use i.e. it is a tool used to explore and analyse large amounts of data to find patterns from the big data. Data mining is thus, the analysis and scrutiny of mammoth data sets, with an aim to uncover significant pattern and rules that were previously unidentified. As every arena of human life has now become data intensive, it has stemmed in making data mining an indispensable constituent. The information and knowledge extracted can be momentously useful for the applications ranging from small business management to complex engineering design to science exploration. Speaking of businesses, companies can predict what certain kind of customers will want to buy, and when, with a great degree of accuracy these days and all this is owed to data mining that also plays a major role in BI. Companies have an enormous influx of data coming from their customer base. Every previous purchase, social media interaction, and search engine query is a clue into what a consumer may buy next. Besides the need to store this gross amount of data, businesses must be able to make sense of it. This is where business intelligence shines. Business intelligence (BI) is a collection of

DOI: 10.48175/IJARSCT-2199

applications and techniques used to transform data into actionable information.BIinvolvesenterprise-leveldataanalysisthatpinpointsareas for operational improvement and external expansion. Thus, all the above comes under the umbrella term "Data Science" which is can be explained as the science that deals with the identification, representation and extraction of meaningful information from a pool of data useful for the growth of businesses by using a mixture of algorithms and analytics.

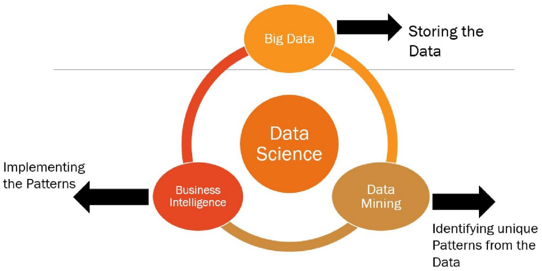


Figure 1: Terms related to Data Science and how they are interconnected to each other

II. Characteristics of BigData

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many multinational companies to process the data and business of many organizations. There are five V's of Big Data that explains the characteristics:

2.1 Volume

The name Big Data itself is related to an enormous size. Big Data is vast 'volumes' of data generated from many sources daily, such as business processes, machines, social media platforms, networks, human interactions, and many more. Facebook can generate approximately a billion messages, 4.5 billion times that the "Like" button is recorded, and more than 350 million new posts are uploaded each day.

2.2 Variety

Big Data can be structured, unstructured, and semi- structured that are being collected from different sources. Data was only being collected from databases and sheets in the past, But these days the data will comes in forms such as PDFs, Emails, audios, SM posts, photos, videos, etc.

2.2 Veracity

In many instances, the data available is inconsistent and uncertain, which might influence the accuracy of the results acquired after research and analysis. Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently.

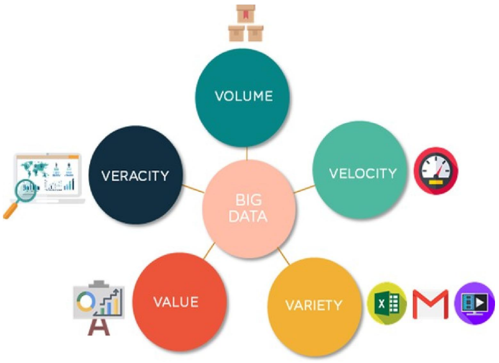


Figure 2: Characteristics of Big Data

DOI: 10.48175/IJARSCT-2199

2.3 Value

Value is an essential characteristic of big data. The primary goal of Big Data analysis is to extract meaningful information that will aid us in forecasting and making decisions.

2.4 Velocity

In the age of digitization, sometimes the speed of data creation is more important than the volume of data. For instance, every minute almost 72 hours of video files are uploaded on YouTube. The need of the hour is how to manage this data in an effective and appropriate manner. Big data velocity deals with the speed at the data flows from sources like application logs, business processes, networks, and social media sites, sensors, mobile devices, etc.

III. Role of Data Mining

Big data produces enormous data from different sources in different domains. This collection of data can be classified into the following types:

- Data Generated from Mobile
- Geographical Data
- Temporal Data
- Streaming Real Time Data
- Online Data

This variety of data is processed using machine learning and data mining techniques. As we know, Data mining is the process of extraction of useful information and patterns from huge data stored either in databases, data warehouses, or other information repositories. It is also called as knowledge discovery in database (KDD), knowledge mining from data, knowledge extraction or data /pattern analysis and is intended to be refined to remove irrelevant data, predict future out comes and help in decision making.

3.1 Process of Data Mining

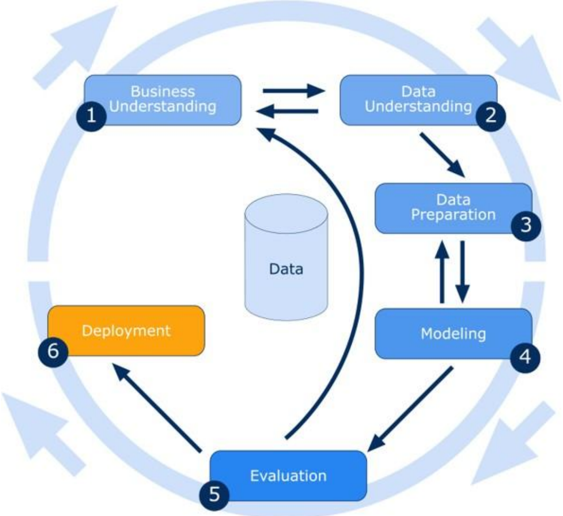


Figure 3: Process of Data Mining

1. **Business Understanding:** In this phase various activities like determining business objectives, finding current situation, determining data mining goal and producing project plan are involved.
2. **Data Understanding:** This phase is basically concerned with establishing the main characteristics of data which includes the data structures, data quality and identifying any interesting subsets of the data.
3. **Data Preparation:** The main activities performed during this phase are select data, cleaning data, data integration and data transformation for constructing the final data set into the desired form.

4. **Modelling:** In Data Modelling step, we select modelling techniques, modelling parameters and assess the model created based on the business objectives.
5. **Evaluation:** The various activities performed during this phase include evaluating results and reviewing the process by verifying the steps in modelling in view of the business objectives.
6. **Deployment:** This is the execution phase which consists of plan deployment, plan monitoring and maintenance, produce and reviewing the final report and deploying the patterns for the desired outcome.

IV. Data Mining Techniques and Functionalities

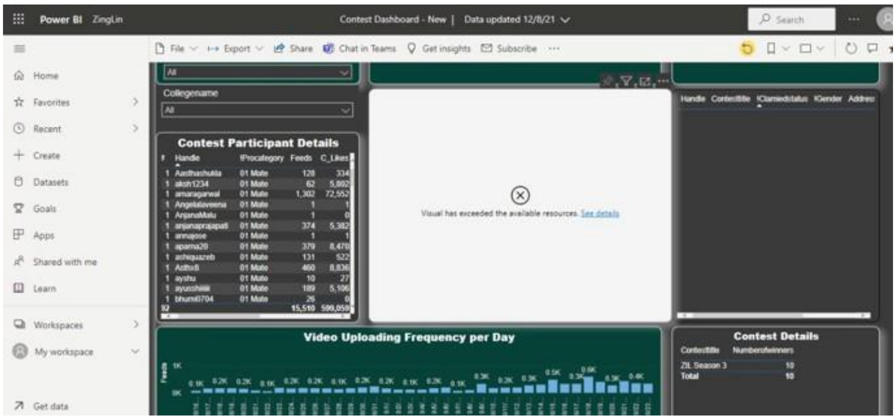
Data mining commonly involves four classes of tasks:

1. Classification, which arranges the data into predefined groups;
2. Clustering, is like classification but the groups are not predefined, so the algorithm will try to group similar items together;
3. Regression, attempting to find a function which models the data with the least error; and
4. Association Rule Learning, searching for relationships between variables.

Data mining functionalities include data characterization, data discrimination, association analysis, classification, clustering, outlier analysis, and data evolution analysis.

V. Business Intelligence

Business intelligence (BI) is a technology-driven process for analysing data and delivering actionable information that helps executives, managers and workers make informed business decisions. Data mining can be seen as the precursor to business intelligence. Upon collection, data is often raw and unstructured, making it challenging to draw conclusions. Data mining decodes these complex datasets, and delivers a cleaner version for the business intelligence team to derive insights. In other words, companies use data mining to gain an understanding of the "what" in order to answer for business intelligence to answer the "how" and "why".



All

Collegename

All

Contest Participant Details

#	Handle	Procategory	Feeds	C_Likes
1	Aasthashukla	01 Male	128	334
1	akoh1234	01 Male	62	5,892
1	amanagarwal	01 Male	1,302	72,552
1	Angelaloveena	01 Male	1	1
1	AnjanaMalu	01 Male	1	0
1	anjanaprajapati	01 Male	374	5,382
1	annajose	01 Male	1	1
1	aparna29	01 Male	379	8,470
1	ashiquazeb	01 Male	131	522
1	Adithi8	01 Male	460	8,836
1	ayshu	01 Male	10	27
1	ayusshiiii	01 Male	189	5,106
1	bhumi0704	01 Male	26	0
92			15,510	599,059

Visual has exceeded the available resources. See details

Handle	ContestTitle	Claimedstatus	Gender	Address

Contest Details

ContestTitle	Numberofwinners
ZIL Season 3	10
Total	10

Figure 4: Real Life Use case of a short video app named Zinglin using PowerBi

Transforming your raw data into business insight via the process of data mining takes place over five steps: Extract, Transform, and Load (ETL): The first stage in data mining involves extracting data from one or many sources (such as those referenced above), transforming it into a standardized format, and loading it into the data warehouse.

Store and manage: Next, businesses store and manage the data in a multidimensional database system, such as OLAP or tabular cubes.

- **Access:** After the data has been standardized, loaded, and managed into the database, business analysts, IT professionals, or data scientists gain access to the data to determine how it should be organized.
- **Analysis:** Application software analyzes and sorts the data based on incoming queries from the end user.
- **Present:** After the data has been analyzed and sorted, it is presented to the end user in an understandable format, such as a report, chart or graph.

Areas where BI and Data Mining are used vastly:

- Retail and E-commerce
- Marketing and Social Media
- Finance Telecommunications

Real-World Data Mining and Business Intelligence Cases:

- Feed visor- Uses Data to Serve Its Retail Clients Better
- Penneo- Used Business Intelligence to Understand Its Clients and Billing
- Brunner- Uses BI to Streamline Its Processes and Exceed Client Expectations
- Zinglin- Used BI to Retain and the customers and improve the traffic on the app

V. Challenges of BigData

Big Data as we know refers to a large amount of data that is difficult to analyse and store. Using various methodologies, data mining is used to examine Big Data in order to turn it into processed and usable form by removing redundancy and irrelevant data. Manually processing large amounts of data is impossible. Data mining techniques have significant advantages, despite the fact that they also provide a number of obstacles, which is a severe concern. The challenges can be broadly categorized into three types:

5.1 Data Challenges

The main challenge is to handle Big Data storage, processing and to discover hidden patterns among those data. Data challenges are identified with the characteristics of data itself (volume, variety, and velocity of data).

The main data challenges that needs to be focused are:

- Data Storage
- Velocity of Data
- Data Variety
- Computational power of data
- Data Understanding
- Quality of Data
- Data Representation

5.2 Management Challenges

Big data technologies are now used in a variety of business sectors. As a result, it is critical that they make the right decision. Finest Big Data Analytics approaches and to get the most out of your data, the changes must be managed by the company in a timely manner.

Some of the challenges are:

- Leadership
- Talent Management
- Management of Technology
- Decision taking capability
- Environment of company
- Security, Privacy, governance, and ethical perspectives

5.3 Process Challenges

Processing challenges are identified with an arrangement of how approaches: how to capture data, how to change data, how to incorporate data, how to choose the correct model for investigation, and how to present the findings. Experts can assess the model based on previous experiences, but they still lack in some areas, posing a problem in the assignment of processing large amounts of data. Choosing the best model for the data analysis is a difficult task that must be taken into

DOI: 10.48175/IJARSCT-2199

serious consideration. The following are the process challenges:

- To capture the Data from various sources
- Conversion of collected data into analyzable form
- Discovering the hidden patterns among the data
- Analyzing the scope of techniques required to filter the data.

VI. Tools for Big Data Analytics

Big Data Analytics or Data mining is useful in areas where better decision making as well as predictive analysis matter. Our main aim is to study and analyze the data in an effective and efficient manner. There are certain data software available known as big data tools which help in productive analysis of the data. Some of the tools are:

1. **Apache Hadoop:** Apache Hadoop is an open source framework, used for clustered file system and managing big data by means of MapReduce programming model. The main features of Hadoop that made it widely used among masses like Amazon are: The HDFS or Hadoop Distributed File System has the ability to hold all type of data. It can be audio, video, images, plain text and XML also. Hadoop is extensible and provides rapid data access. At the same time Hadoop uses more disk space due to its thrice data redundancy.
2. **Cloudera Distribution for Hadoop:** Cloudera Hadoop Distribution provides extensible and flexible platform. Because of this Cloudera results in better performance while managing larger volume and more of variety data. Clouder a provides more potential to process and analyze unlimited data. This makes it popular among major companies. There are automated wizards in Cloudera that help to deploy the cluster faster. This makes it easy to manage Hadoop. Cloudera is integrated in its nature therefore it provides highly secure environment for data analysis.
3. **Cassandra:** Apache Cassandra is an open source, distributed NoSQL database management system. It is used to manage large volume of data. It also uses CQL i.e. Cassandra Structure Language to interact with database. Companies like American Express, Facebook, Yahoo use Cassandra. Cassandra is highly available as it can efficiently manages data that is spread across a number of servers. As it is distributed in nature there is no single point failure. It has log structured storage and linear scalability.
4. **KNIME:** KNIME stands for Konstanz Information Miner. It is an open source tool which is widely used for research, data mining, data analytics, and Business Intelligence. It is sometimes considered as a good choice because it integrates easily with other technologies and languages. It also supports Linux, OS X, and Windows Operating Systems. Companies like Comcast, etc. Cassandra has an abundant algorithm set for efficient computations. It automates manual work productively hence reduces workload. KNIME is easy to set up and has ETL operations.
5. **Rapid Miner:** Rapid miner is a cross-platform tool which offers an integrated environment for data science, machine learning and predictive analytics. It comes under various licenses that offer small, medium and large proprietary editions as well as a free edition that allows for 1 logical processor and up to 10,000 data rows. Organizations like Hitachi, BMW, Samsung, Airbus, etc have been using RapidMiner.
6. **HPCC:** HPCC stands for High-Performance Computing Cluster. This is a complete big data solution over a highly scalable supercomputing platform. HPCC is also referred to as DAS (Data Analytics Supercomputer). This tool was developed by LexisNexis Risk Solutions. This tool is written in C++ and a data-centric programming language known as ECL(Enterprise Control Language). It is based on a Thor architecture that supports data parallelism, pipeline parallelism, and system parallelism. It is an open-source tool and is a good substitute for Hadoop and some other Big data platforms.

VII. Applications of Data Science

Predictive Causal Analytics and Machine Learning are more important in Data Science. Data Science is primarily used to make decisions and predictions using predictive causal analytics, prescriptive analytics (predictive and decision science), and machine learning. Data science efficiently solves this problem, i.e., refines or extracts uncovered/hidden

DOI: 10.48175/IJARSCT-2199

patterns/information from this large data set. Life is being made easier with application of Big Data and analytics across various fields including entertainment, civil services, and health. Many organisations gather information about which users enjoy which types of plays, movies, music, and other items. These kinds of suggestions are being developed and greater services are being delivered to the users in question. Data Scientists are in charge of such tasks. For Example, Netflix and YouTube use data mining to detect pause, play, and repeat instances in movie/music viewing habits, i.e., to find user preferences, interest, and makes decisions based on that. The recommendation algorithms are the core of Netflix product. It provides the members with personalized suggestions to reduce the amount of time and frustration to find something great content to watch. Recommendations are made by the system based on the data and demographics related to user watch list, search history.

Big Data Analytics plays a prominent role in scaling business by finding insights. The rise of the data analytics has made it possible for organisations to discover consumer shopping habits. Not only it helps in B2C services, but has significantly improved B2B services as well.

Amazon, like adtech's ”duopoly” Facebook and Google, was drawn into the advertising business by the sheer volume of consumer data at its disposal. Since its inception in 1994, the company has amassed vast amounts of data on what millions of people buy, where their purchases are delivered, and which credit cards they use. In recent years, Amazon has begun to provide more and more businesses, including marketing firms, with access to its self-service ad portal, where they can buy ad campaigns and target them to ultra-specific demographics, such as past purchasers.

Big Data in Health Sector: In a country as populated as India, it was challenging to track the number of Covid Cases reported every second amidst the pandemic. That's when an anonymous group called "millennials" launched the Covid-19 Live Dashboard which presented the number of Covid Cases reported: Positive cases, recovered cases, and the number of deaths state, city, area wise. It has also started displaying the vaccination status across the country. Another application is the Aarogya Setu app – the Self Assessment App, predicts the risk of infection based on data recorded around your vicinity.

According to a Forbes Article, Uber Charges more if they think you are willing to pay. What they act upon is called Dynamic Pricing. Dynamic Pricing refers to how they act to determine price. It operates on the basis of Algorithmic Pricing, which determines what price to deliver based on various variables such as your location, time of day, traffic patterns, and even your Uber user history. This information is gathered, and the algorithm predicts the highest price you are likely to pay. This ”willingness to pay” algorithm can predict how likely you are to agree to the current ride price. The results of the determining are incorporated into demand predictions by micro-segments, determining what price to set the ride at each time.

VIII. Emerging Trends and Future of Data Science

Rapid digitalization is one of the things the Covid-19 epidemic has encouraged. Data looks significantly different now than it did before the pandemic, thanks to a greater emphasis on internet services and e-commerce, and properly using it has never been more crucial than now.

The following latest data science trends indicate that a fresh perspective is emerging. Data is no more a science reserved for a small number of experts, but rather a valuable chance for every individual in a company to learn and refine their skills, adapt to new data science techniques, and rethink data collecting and analysis.

- **Rise of Data Fabric:** With the complexity of data and its potential value, the need for a unified foundation on which to build and store each business's composable data and analytics has grown. Using data fabric as the central architecture allows for the effective cohesion of hardware and software, allowing access across a variety of internal and external locations without violating data privacy laws. Existing data lakes, hubs, and warehouses can be combined with new software tools and approaches to revolutionise data governance for individual businesses. As a result, less integration and maintenance are required, allowing businesses to provide more effective updates to customer experience more quickly.
- **Big Data on the Cloud:** Businesses are increasingly turning towards cloud services for data storage, processing, and distribution. One of the major data management trends in 2021 is the use of public and private

cloud services for big data and data analytics. With the rapidly accelerating developments in cloud technology, new data science trends have attracted many businesses to rethink their data storage. Providers such as Amazon, Microsoft, and Google are now the prime way for businesses to store their data and offer built-in analytics that help streamline the data management process. Another caveat to cloud-based data is homomorphic encryption, which means that computational calculations and analysis can be performed on data without decryption; therefore keeping your data even more secure and removing the need for the holder of the decryption key to be in the same location as the data. Using Cloud-based solutions minimizes the risk of bugs and errors, as the services are used widely and matured over the years of maintenance.

- **Use of Augmented Analytics:** Going hand-in-hand with cloud-based data is the trend of augmented and user-friendly analytics. While it was previously necessary for trained specialists to interpret and evaluate data, employees at any level are now able to do so thanks to integrated data technology. The rise of Internet of Things (IoT) devices ensures that every employee is in possession of a smart device that is capable of processing data of some kind. Employees from different departments are able to share and compare data and come up with solutions and ideas that will benefit everyone via predictive analytics and trend forecasting. This analytics trend marks a move towards universal access to analytics in conjunction with a sharper focus on the specific needs of different business departments, business types, and individual employees. Instead of relying on the opinions and experience of a select handful of specialists and a set of predetermined general questions, businesses can benefit from the varied, experiential view points of all of their employees.

Other honourable mentions include Explosion in Deep fake video and audio, More applications created with python, Increased demand for end-to-end AI solutions, Increased in Consumer Data Protection and AI automation.

IX. Conclusion

This paper discussed several essential terminologies re- lated to data science, its applications as well as gave a brief view on the emerging trends that we are to see in the years to come. Data science is going to play a very important role in the future as it is already evident from today's times itself and hence this paper also covers the challenges that are faced as data is being collected on every nook and corner you go, being it a small or big business, leading to rapid digitization and hence putting our privacy and security at a greater risk than ever before. Hence, future researchers should focus on finding better solutions to these problems using the data mining algorithms, techniques and tools as well as keeping into account to ensure the growth of businesses in an efficient manner.

Acknowledgments

The authors would like to express our humble gratitude and earnestly acknowledge the sincere efforts and valuable time given by our guide Mrs. Geetha S. Their valuable guidance and feedback has helped us in completing this paper. We would also like to thank our peers and everyone who helped and motivated us to work on the paper.

References

[1]. D. Goyal, R. Goyal, G.Rekha, S. Malik and A. K. Tyagi, ”Emerging Trends and Challenges in Data Science and Big Data Analytics,” 2020 International Conference on Emerging Trends in Information Tech- nology and Engineering (ic-ETITE), 2020, pp. 1-8, doi: 10.1109/ic- ETITE47903.2020.316.

[2]. N. Lopes and B. Ribeiro, ”Novel Trends in Scaling Up Machine Learning Algorithms,” 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 632-636, doi: 10.1109/ICMLA.2017.00-90.

[3]. Z. A. Al-Sai, R. Abdullah and M. h. husin, ”Big Data Impacts and Challenges: A Review,” 2019 IEEE Jordan International Joint Confer- ence on Electrical Engineering and Information Technology (JEEIT), 2019, pp. 150-155, doi:10.1109/JEEIT.2019.8717484.

[4]. A. Ashabi, S. B. Sahibuddin and M. S. Haghighi, ”Big Data: Cur- rent Challenges and Future Scope,” 2020 IEEE 10th Symposium on Computer Applications and Industrial Electronics (ISCAIE), 2020, pp. 131-134, doi:10.1109/ISCAIE47305.2020.9108826.

[5]. C. Komalavalli and C. Laroiya, ”Challenges in Big Data Analytics Techniques: A Survey,” 2019 9th International Conference on Cloud Computing, Data Science and Engineering (Confluence), 2019, pp. 223-228, doi:10.1109/CONFLUENCE.2019.8776932.

[6]. Mittal, Sonu and Shuja, Mirza and Zaman, Majid. (2016). A Review ofDataMiningLiterature.IJCSIS.14.437-442.

[7]. Prasdika, Prasdika and Sugiantoro, Bambang. (2018). A Review Paper on Big Data and Data Mining Concepts and Techniques. IJID (International Journal on Informatics for Development). 7. 33. 10.14421/ijid.2018.0 7107.

[8]. Anshu, Review Paper on Data Mining Techniques and Applications (MARCH 31, 2019). International Journal of Innovative Research in Computer Science and Technology(IJIRCST),Volume-7,Issue-2, March 2019, Available at SSRN: https://ssrn.com/abstract=3529347

[9]. Li, Yihao, and Theresa Beaubouef. ”Data Mining: Concepts, Back- ground and Methods of Integrating Uncertainty in Data Mining.” CCSC:SCStudentE-Journal3(2010):2-7.

[10]. Mehrotra, Ankit, and Reeti Agarwal. ”A Review of Use of Data Mining during COVID-19 Pandemic.” Turkish Journal of Computer and Mathematics Education(TURCOMAT)12.6(2021):4547-4552.

[11]. Bharati, M. and Ramageri, Bharati. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering.

[12]. Rouhani, Saeed and Asgari, Sara and Mirhosseini, Vahid. (2012). Review Study: Business Intelligence Concepts and Approaches.

[13]. Ranjan, Jayanthi. (2009). Business intelligence: Concepts, compo- nents, techniques and benefits. Journal of Theoretical and Applied Information Technology. 9.60-70.

[14]. Alessandro Massaro , Valeria Vitti, Angelo Galiano , Alessandro Morelli (2019). Business Intelligence Improved by Data Mining Algorithms and Big Data Systems: An Overview of Different Tools Applied in Industrial Research. Computer Science and Information Technology, 7(1),1-21.DOI:10.13189/csit.2019 .070101.

DOI: 10.48175/IJARSCT-2199