International Journal of Research
(IJR)

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

Basics of Data Science

Dr. D. Hema Latha[1], Azmath Mubeen[2], Dr. D. Rama Krishna Reddy[3],

Asst. Professor, Dept. of Computer Science, University College for Women, Hyderabad, TS, India [1]
Asst. Professor, Dept. of Computer Science, University College for Women, Hyderabad, TS, India [2]
Asst. Professor, Dept. of Mathematics, UCS, Osmania University, Hyderabad, TS, India[3]

ABSTRACT

In this present era of big data Hadoop, cloud and other frameworks are available for huge data storage, now the concentration is on huge data processing. With the help of machine learning tools, techniques and algorithms, huge data processing can be done and AI can be implemented in machines and make the machines to function like human brain intelligence.

With the huge rise in data, there is a continual requirement for analyzing such a huge amount of data. Data Science concept can handle this data and develop useful machine learning models that predict the future results. Data Science is emerging multidisciplinary field with roots in mathematics, statistics, and computer science. Data science can be applied to wide range of applications such as business, finance, healthcare, transportation etc., as it can perform data extraction, data analysis, data visualization, and also manages huge amounts of data. The main objective of Data Scientists is to recognize and utilize relevant and important insights from data, so that it can be helpful for organizations in taking smarter decisions. During this process, different tools and methods to identify redundant patterns and hidden knowledge within the data can be used. Most efficient algorithms, powerful hardware and programming systems to solve the data related problems are also used.

This paper focuses on basics of Data Science and its implementation with a simple case study. Understanding and implementing data science can upgrade individual knowledge, skills and business.

Key Words – Data Science, Business Intelligence (BI), visualization techniques.

I. INTRODUCTION

Data Science is combination of machine learning principles, various tools and algorithms, the main goal of Data Science is to discover and extract hidden patterns from the raw data and process.

International Journal of Research (IJR)
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

Data Science [1] provides advance skills, techniques to analyze large amounts of data, data mining and programming. Data Science consists various phases. In order to uncover useful intelligence for the organizations, data scientists must master the total spectrum of data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.

Data Science [2] has evolved as one of the most promising and in-demand career paths for skilled professionals.

II. DATA SCIENCE LIFE CYCLE

The main phases of the Data Science Lifecycle –

Main phases of data science [3] life cycle are depicted in figure 1 and figure 2, followed by description of each phase.

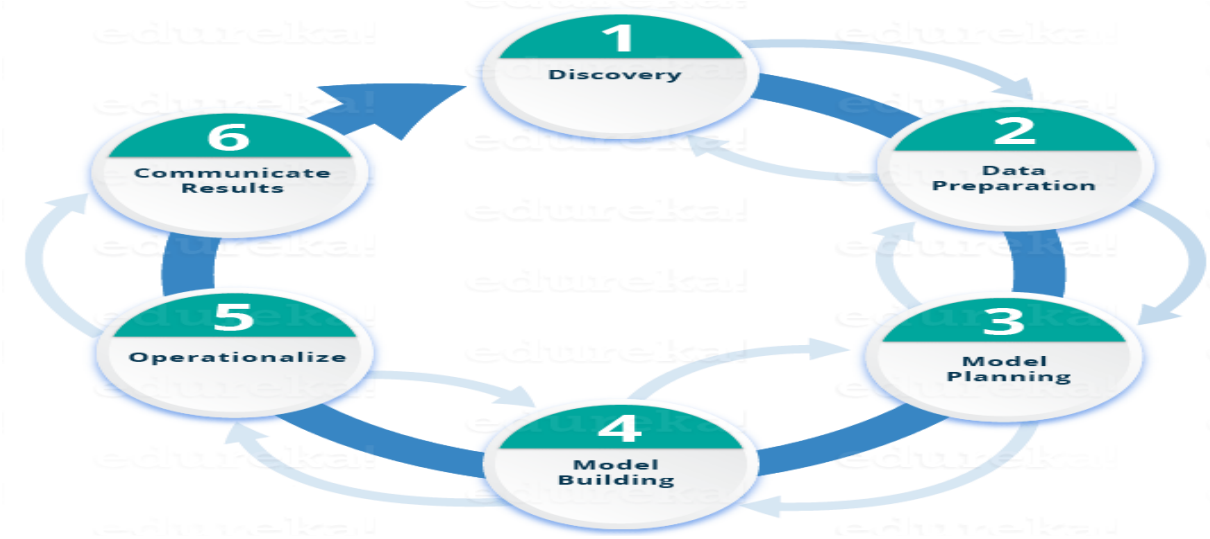

Figure 1. Phases of Data Science life cycle

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

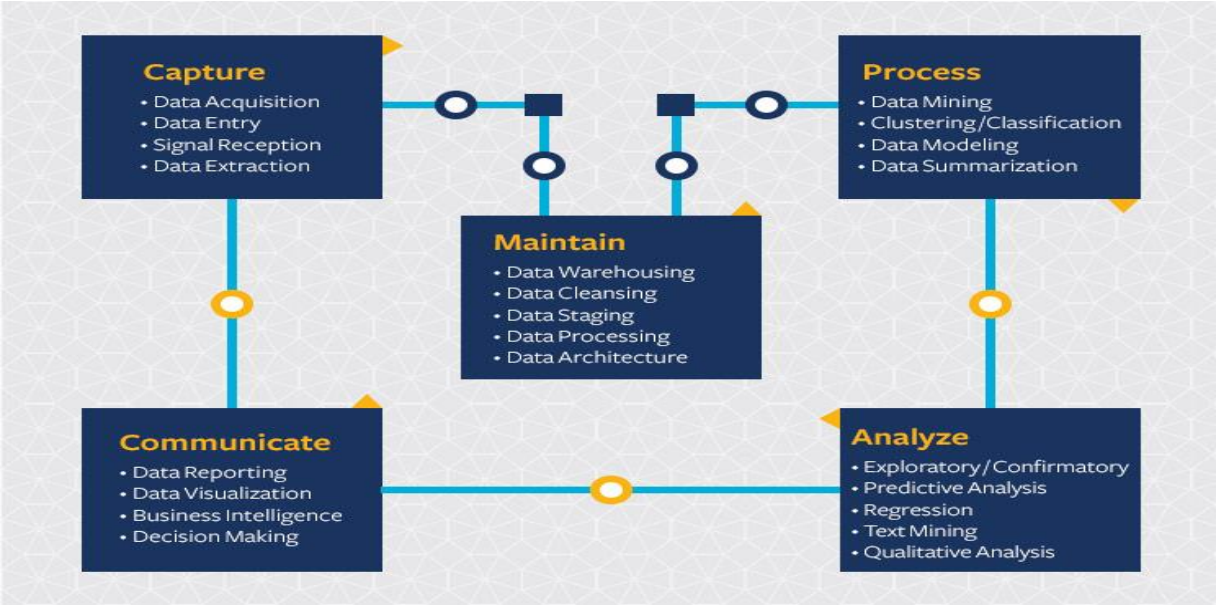

Figure 2. Phases of Data Science life cycle

2.1. Discovery Phase/Data Capture: In this first phase, Data acquisition, data entry, signal reception and data extraction processes are performed. Before the project is started various specifications, priorities and budget required should be understood and planned properly. Once has to gather the requirements with the help of questionnaire, interviews etc. The person gathering the requirements should possess the ability to ask the right questions. Here, one can assess the requirements like people, technology, time and data to support the project. In this phase, business problem framing and initial hypotheses (IH) formulation [4] is needed to test.

2.2. Data Preparation/ Maintain phase: In this phase, Data Warehousing, data cleansing, data staging, data processing and data architecture is performed. Here, analysis is performed, for the entire duration of the project. Here, analytical sandbox is maintained and Extract, Transform, Load and Transform (ETLT) are performed to get the data into the sandbox. The statistical analysis flow is shown in the below figure 3.

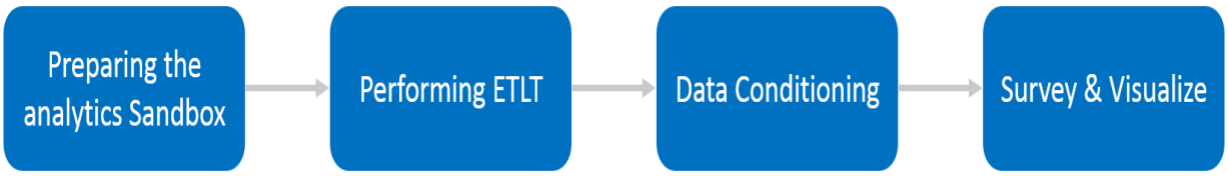

Figure 3. Statistical Analysis Flow

For data cleaning, transformation, and visualization, python is the best programming language

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

that can be used. This will help to spot the outliers and establish a relationship between the variables. Once the data is cleaned and prepared, next is to perform exploratory analytics on the cleaned and prepared data.

2.3. Model planning/Process: data mining, clustering/classification, data modeling and data summarization functions are performed in the phase. . In this model planning phase, the methods and techniques are determined, to draw the relationships between variables. These relationships will set the base for the algorithms which can be implemented in the next phase. Exploratory Data Analytics (EDA) can be applied using various statistical formulas and visualization tools.
Various model planning tools are shown in the figure 4.

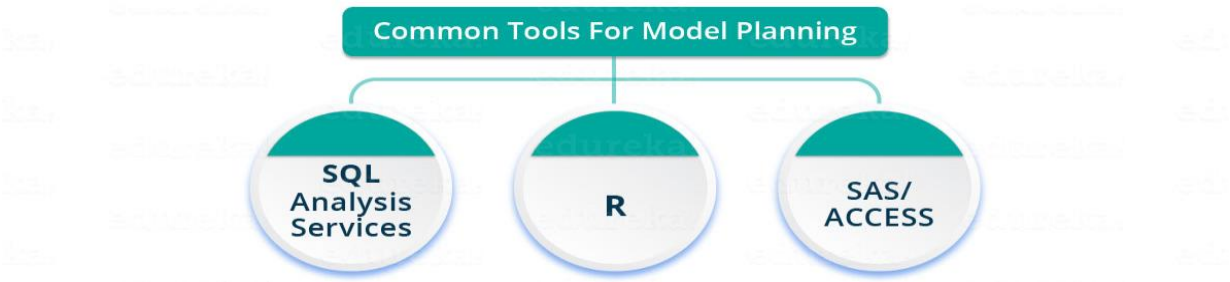

Figure 4. Model planning tools

2.3.1. SQL Analysis services can perform database analytics using common data mining functions and fundamental predictive models.

2.3.2. R provides a very good and flexible environment for building interpretive models. It also consists of complete set of modeling capabilities.

2.3.3. AS/ACCESS is used to access data from Hadoop, which is used for creating iterative and reusable model flow diagrams.

Although, many tools are available in the market, R is the most commonly used and popular tool.

Once the nature of the data is understood, in the next phase algorithms are applied and a model is constructed.

2.4. Model building/Analyze: The functions performed in this phase are- exploratory/confirmatory, predictive analysis, regression, and text mining, qualitative analysis. In this phase, datasets for training and testing purpose is developed. Now, it is decided whether the existing tools will be sufficient for running the models or it will need a more robust/flexible environment such as fast and parallel processing). Various learning techniques like classification, association and clustering are analyzed to build the model. Model analysis and building tools are shown in the figure 5.

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

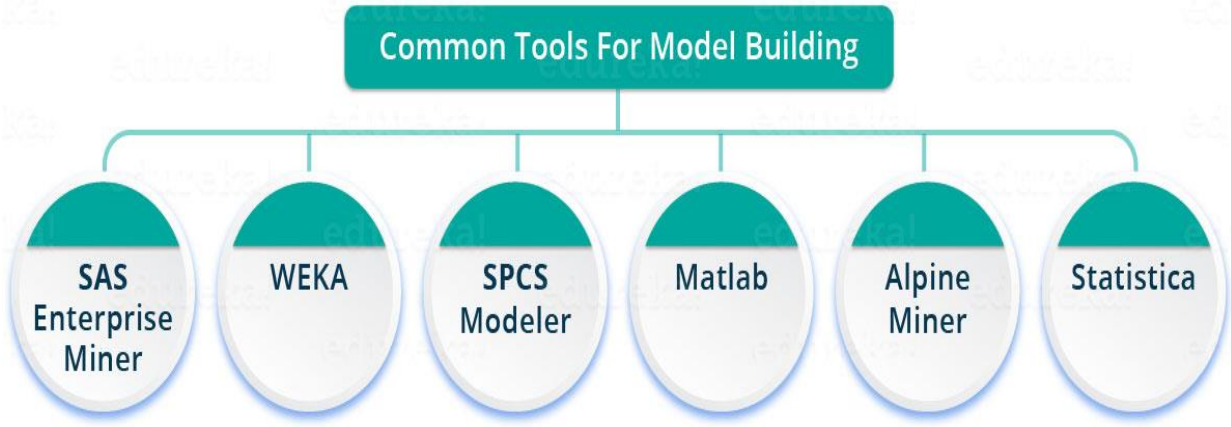

Figure 5. Model building tools

2. 4.1. SAS Enterprise Miner: SAS Enterprise Miner is used to create accurate predictive and descriptive models on large volumes of data across different sources in the organization. SAS Enterprise Miner provides numerous features and functions for the business analysts to model their data.

2. 4.2. WEKA: It is a collection of machine learning algorithms for solving real-world data mining problems. WEKA is written in Java and works on any platform. These algorithms can be applied directly on a dataset or can be imported from Java code.

2. 4.3. SPSS Modeler: It is a major tool for visual data science and machine-learning solution. This tool provides better upgraded connection between time and value and achieves desired outputs by enhancing operational tasks for data scientists.

2. 4.4. Mat lab: It is a programming and numeric computing platform that is used to analyze data, develop algorithms, and create models. This tool is used by millions of engineers and scientists.

2. 4.5. Alpine Miner: This tool is designed for customers with fast-growing data needs. This allows the addition of new variables to predictive models.

First Alpine's core product was designed and then Alpine Miner, which allows non-data scientists also, to create predictive analytics data models without using code.

2. 4.6. Statistica: Statistica is an advanced analytics software package which was originally developed by StatSoft and presently maintained by TIBCO Software Inc. This tool provides data

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

analysis, data management, statistics, data mining, machine learning, text analytics and data visualization processes.

2.5. **Operationalize:** In this phase, briefings, code, final reports and technical documents are delivered. Apart from this, a trail run/ experimental code are also implemented in a real-time operational environment. This trail run implementation provides a clear picture of the performance and other related constraints on a small scale before the deployment of the project.

2.6. Communicate results: the major functions of this phase are – data reporting, data visualization, business intelligence and decision making. So, in the last phase, the project can be evaluated and all the key findings are identified and communicated to the stakeholders and determine if the outcome/results of the project are positive or negative, that is., project is success or failure based on the criteria developed in phase 1.

III. NEED FOR DATA SCIENCE

- Most of the data which we had was structured and small in size, which could be analyzed by using simple Business Intelligence tools. In the present situation, today, most of the data is unstructured or semi-structured. Figure 6 shows more than 80% of the data currently is unstructured.

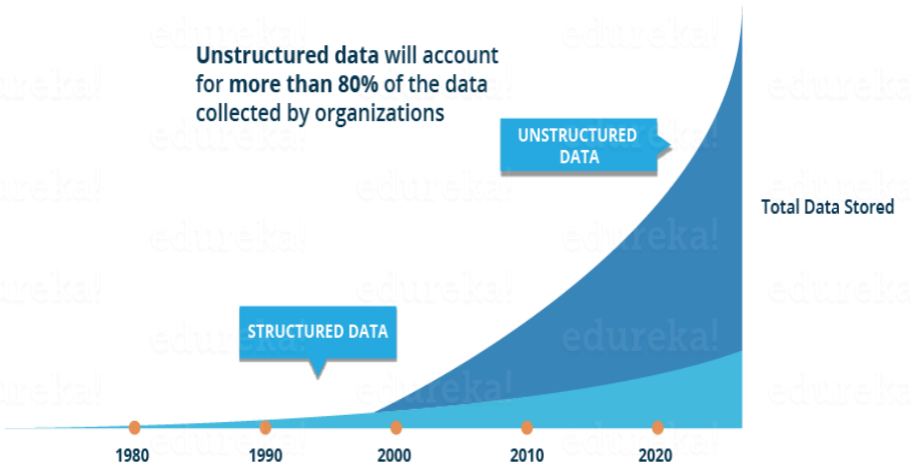

Figure 6. Graph representing structured and unstructured data

This data represented in the figure 6 is generated from different sources like financial logs, sensors text files, multimedia forms and instruments etc. Simple BI tools will not be sufficient to process the huge volume and variety of data. This is why there is a need for data science, which consists of more complex and advanced analytical tools and algorithms for processing, analyzing.

Following examples gives the need for data science.

3.1. To understand precise requirements of the customers from the existing data like the customer's past browsing history, purchase history, age and income, though this data is available earlier also, but now with the help of vast amount and variety of data, one can get training models, data sets more effectively and recommend the product to the customers with more precision.

3.2. Autonomous or self-driving or driverless cars can be possible with the help of data science, where the care can have driving intelligence. The self-driving cars collect live data from sensors, including radars, cameras, and lasers to create a map of its surroundings. Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where to take a turn – making use of advanced machine learning algorithms.

3.3. Data Science can be used in predictive analytics. Example for this is weather forecasting. Data from ships, aircraft, radars, satellites can be collected and analyzed to build models. These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities. It will help to take appropriate measures beforehand and save from damage. Figure 7 gives the info graphic display of the domains where the data science can be utilized.

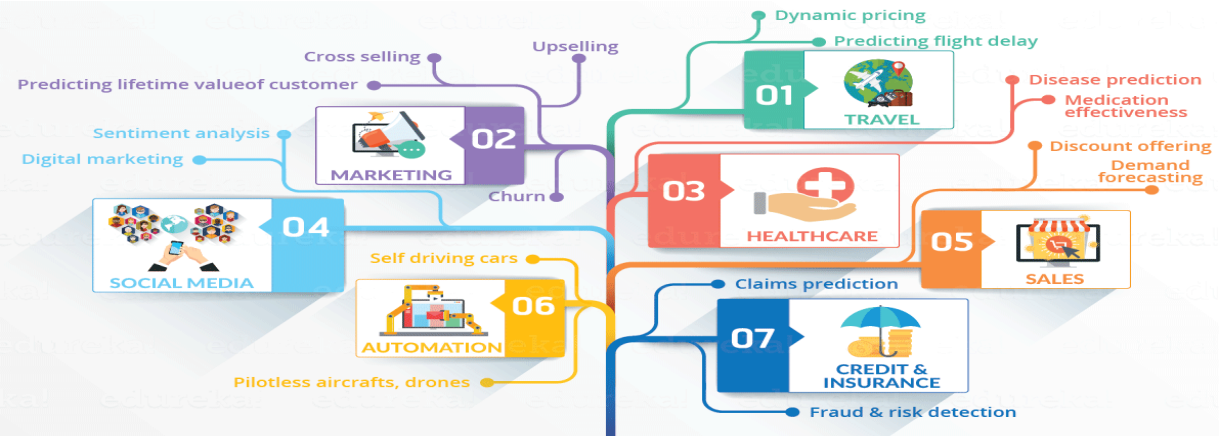

Figure 7. Applications or domains of data science

IV. ROLE OF DATA SCIENTIST

Data science [5] continues to evolve as one of the promising and in-demand career paths for skilled professionals. Data Scientist is one who practices the art of Data Science. Data Scientist collects lot of information from the scientific fields and applications either from statistics or mathematics. Data scientists should be good at analyzing huge amounts of data, must have data mining and programming skills. Data scientists must master the total spectrum of data science life cycle and possess high level of knowledge and flexibility to enhance the returns at each phase of the process.

V. FUNCTIONS OF DATA SCIENTIST

Data scientists solve complex data problems with their strong expertise in certain scientific disciplines. They work with several domains related to mathematics, statistics, computer science, etc. They utilize latest technologies and strategies, in finding solutions and reaching conclusions which are critical and important for an organization's enhancement and development. Data Scientists represents the data in a very convenient and functional format than the raw data available to them from structured as well as unstructured forms.

VI. DIFFERENCES BETWEEN DATA ANALYSIS AND DATA SCIENCE

Figure 8. Pictorial representation of Business Intelligence and Data Science

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

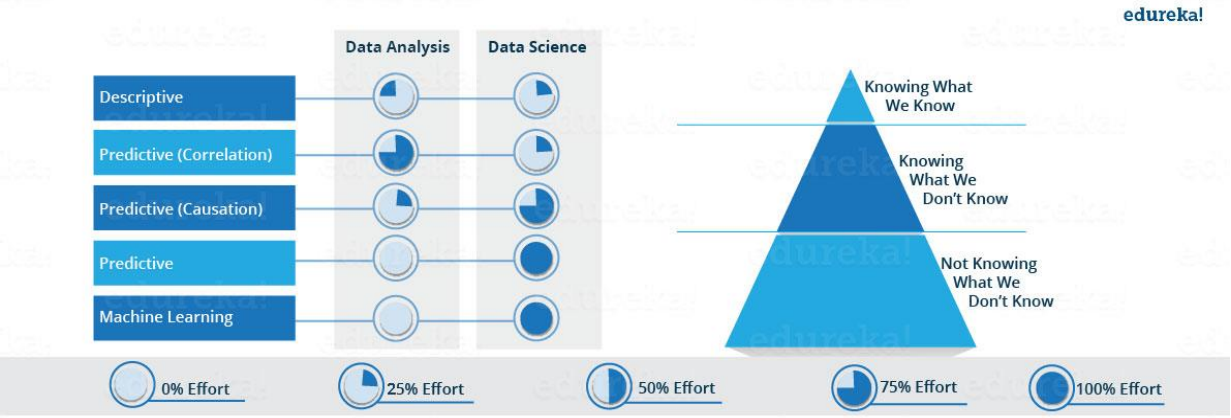

Figure 9. Business Intelligence and Data Science

Figure 9 shows that Data Analysis includes descriptive analytics and prediction also to some extent. On the other hand, Data Science is more about Predictive Causal Analytics and Machine Learning.

Figure 8 and Figure 9, clearly shows the contrasts between data analysis and data science [6], [7]. Data analyst typically focuses on what is happening by processing history of the data; however Data Scientist not just performs exploratory analysis to find insights from it, yet in addition utilizes different advanced machine learning algorithms, AI calculations to recognize the occurrence of a particular event in the future. A Data Scientist will observe the data from numerous points and angles, where some of the angles not known before.

Essentially, Data Science is utilized to perform decisions and predictions with the help of predictive causal analytics, prescriptive analytics and data science.

6.1. Predictive causal analytics – This model is used to predict the possibilities of a particular event in the future. For example, if money is provided on credit, then the probability of customers making future credit payments on time is a matter of concern, then here, a model can be built that can perform predictive analytics on the payment history of the customer to predict if the future payments will be done on time or not.

6.2. Prescriptive analytics: This can be applied for a model that has the intelligence of taking its own decisions and the ability to modify it with dynamic parameters; this can be assumed as adding a new field. Means, it predicts and also suggests a range of prescribed actions and associated outcomes. The best example for prescriptive analysis is Google's self-driving car. The data collected by the vehicles can be used to train self-driving cars and algorithms can be applied on this data to embed intelligence to it. This will enable the car to take decisions like when to take turn, which route to take, when to slow down or speed up etc.

6.3. Machine learning for making predictions — This can be used for transactional data of a finance company and need to build a model to determine the future trend. This comes under supervised learning paradigm. It is called supervised because; already data is available, based on which machines can be trained. For example, a fraud detection model can be trained using a historical record of fake/ fraud purchases.

6.4. Machine learning for pattern discovery — if the prescribed parameters are not available to make predictions, then hidden patterns is searched within the dataset in order to make meaningful predictions. This is related to unsupervised model, as it does not require any predefined labels for grouping. Clustering is the most common algorithm used for pattern discovery. For example, Telephone Company requires establishing a network by putting towers in a region. Then, clustering technique can be used to find those tower locations which will ensure that all the users receive optimum signal strength.

VII. BUSINESS INTELLIGENCE (BI) vs. DATA SCIENCE

- Business Intelligence (BI) [8] analyzes the previous data to find the details to describe business trends. For example, Here BI collects data from external and internal sources, prepare it, execute queries on it and create dashboards to answer questions like quarterly revenue analysis or business problems. BI can evaluate the impact of certain events in the near future. Table 1 shows the differences between BI and Data science [9].

Features	Business Intelligence (BI)	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro-linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

Table1 : Differences between Business Intelligence (BI) and Data Science

For smooth functioning of data science projects, all the phases of life cycle should be followed systematically. Data science [10] is an exploratory way with the focus on analyzing the past or current data and predicting the future outcomes with the goal of making informed decisions.

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

. One should have adequate data collection and analysis, requirements understanding, framing the business problem to reach the goal successfully.

VIII. CASE STUDY: DIABETES PREVENTION

In this case study, chance of diabetes occurrence in future, can be predicted and appropriate measures can be taken beforehand to prevent it.

Step 1:

- First, data is collected based on the medical history [11], [12] of the patient as discussed in Phase 1. Sample unstructured data is represented in table 2.

;npreg;glu;bp;skin;bmi;ped;age,income
1;6;148;72;35;33.6;0.627;50
2;1;85;66;29;26.6;0.351;31
3;1;89;80;23;28.1;0.167;21
4;3;78;50;32;31;0.248;26
5;2;197;70;45;30.5;0.158;53
6;5;166;72;19;25.8;0.587;51
7;0;118;84;47;45.8;0.551;31
8;1;103;30;38;43.3;0.183;33
9;3;126;88;41;39.3;0.704;27
10;9;119;80;35;29;0.263;29
11;1;97;66;15;23.2;0.487;22
12;5;109;75;26;36;0.546;60
13;3;88;58;11;24.8;0.267;22
14;10;122;78;31;27.6;0.512;45
15;4;97;60;33;24;0.966;33
16;9;102;76;37;32.9;0.665;46
17;2;90;68;42;38.2;0.503;27
18;4;111;72;47;37.1;1.39;56
19;3;180;64;25;34;0.271;26
20;7;106;92;18;39;0.235;48
21;9;171;110;24;45.4;0.721;54

Table 2 : Sample unstructured data

Various attributes are specified to predict the occurrence of diabetics

Attributes:

1. glucose – Plasma glucose concentration
2. bp – Blood pressure
3. npreg – Number of times pregnant
4. skin – Triceps skin fold thickness
5. bmi – Body mass index
6. ped – Diabetes pedigree function
7. age – Age
8. income – Income

Step 2:

- Once the data is collected, clean and prepare the data for data analysis.

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

- The data with inconsistencies like missing values, blank columns, abrupt values and incorrect data format should be cleaned. Sample data with inconsistencies is shown in table 3.
- The data is well organized into a single table under different attributes, which looks like structured data, which is shown in table 4.
- Let's have a look at the sample data below.

	npreg	glu	bp	skin	bmi	ped	age	income
1	6	148	72	35	33.6	0.627	50	
2	1	85	66	29	26.6	0.351	31	
3	1	89	6600	23	28.1	0.167	21	
4	3	78	50	32	31	0.248	26	
5	2	197	70	45	30.5	0.158	53	
6	5	166	72	19	25.8	0.587	51	
7	0	118	84	47	45.8	0.551	31	
8	one	103	30	38	43.3	0.183	33	
9	3	126	88	41	39.3	0.704	27	
10	9	119	80	35	29	0.263	29	
11	1	97	66	15	23.2	0.487	22	
12	5	109	75	26	36	0.546	60	
13	3	88	58	11	24.8	0.267	22	
14	10	122	78	31	27.6	0.512	45	
15	4		60	33	24	0.966	33	
16	9	102	76	37	32.9	0.665	46	
17	2	90	68	42	38.2	0.503	27	
18	4	111	72	47	37.1	1.39	56	
19	3	180	64	25	34	0.271	26	
20	7	106	92	18		0.235	48	
21	9	171	110	24	45.4	0.721	54	

Table 3: Sample data with inconsistencies.

The inconsistencies in the table 3 are:

1. In the column **npreg**, it should be numeric 1, but "one" is written in words.
2. In **bp** column one of the values is 6600 which is impossible, human **bp** cannot be that large value.
3. Income column is blank and does not make sense in predicting diabetes, also **glu** and **bmi** columns, one of the rows are empty. Therefore, it is redundant to have it here and should be removed from the table.
4. Also **glu** and **bmi** columns, one of the rows are empty.

- So, data clean and preprocess is done by removing the outliers, filling up the null values and normalizing the data type.
- Finally, the clean data as shown in table 4, which can be used for analysis.

International Journal of Research (IJR)
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

	npreg	glu	bp	skin	bmi	ped	age
1	6	148	72	35	33.6	0.627	50
2	1	85	66	29	26.6	0.351	31
3	1	89	80	23	28.1	0.167	21
4	3	78	50	32	31	0.248	26
5	2	197	70	45	30.5	0.158	53
6	5	166	72	19	25.8	0.587	51
7	0	118	84	47	45.8	0.551	31
8	1	103	30	38	43.3	0.183	33
9	3	126	88	41	39.3	0.704	27
10	9	119	80	35	29	0.263	29
11	1	97	66	15	23.2	0.487	22
12	5	109	75	26	36	0.546	60
13	3	88	58	11	24.8	0.267	22
14	10	122	78	31	27.6	0.512	45
15	4	97	60	33	24	0.966	33
16	9	102	76	37	32.9	0.665	46
17	2	90	68	42	38.2	0.503	27
18	4	111	72	47	37.1	1.39	56
19	3	180	64	25	34	0.271	26
20	7	106	92	18	39	0.235	48
21	9	171	110	24	45.4	0.721	54

Table 4: cleaned and prepared data (Structured data)

Step 3:

Here, analysis is done as discussed in Phase 3.

- First, the data is loaded into the analytical sandbox [13] and various statistical functions are applied on it. For example, R has functions like *describe* which gives the number of missing values and unique values. Summary function can be used which gives statistical information like mean, median, range, min and max values.
- Then, visualization techniques [14] like histograms, box plots and line graphs are used to get a fair idea of the distribution of data. Figure 9, shows the graphs which represents distribution of data related to attributes like npreg, glu, bp, skin, ped, bmi and age.

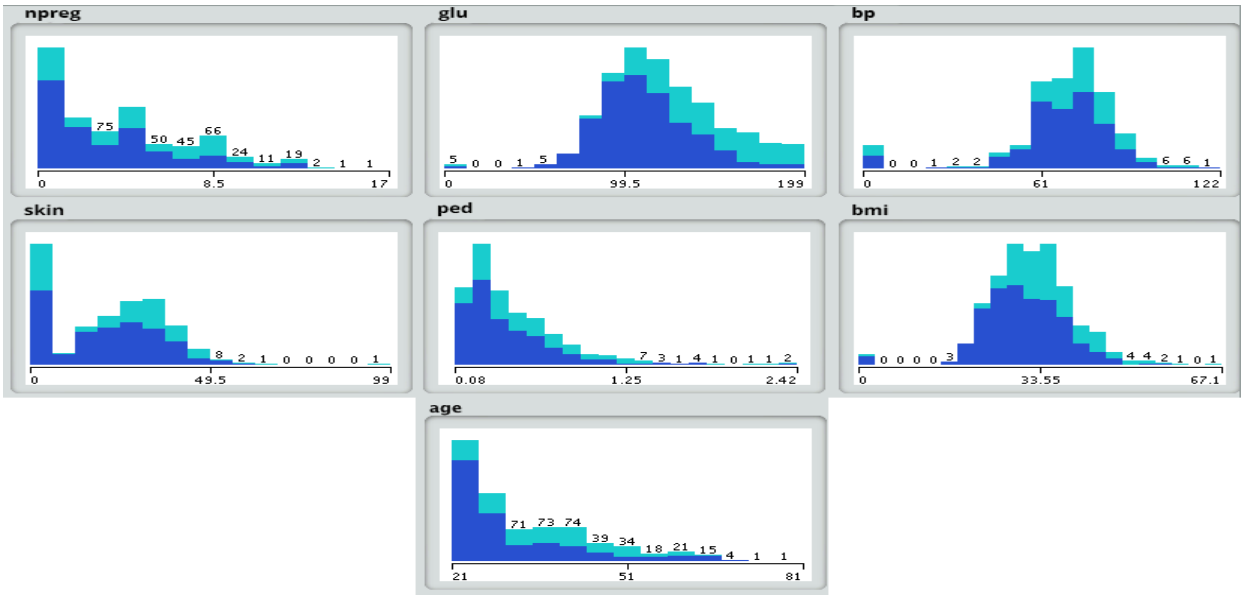

Figure 9: graphs representing distribution of data related to attributes.

Step 4:

From the previous details, decision tree is taken as the best fit, for clear representation:

- Based on the major attributes already existing in this case study for analysis, like ***npreg, bmi,*** etc., supervised learning technique is used here to build a model.
- In particular decision tree is used because it takes all attributes into consideration and clearly represents the attributes with linear relationship as well attributes with non-linear relationship. In this case study, there is linear relationship between ***npreg*** and ***age***, and nonlinear relationship between ***npreg*** and ***ped.***
- As decision tree models [15] are also very robust and flexible one can use the different combination of attributes to make various trees and then finally implement the one with the maximum efficiency.

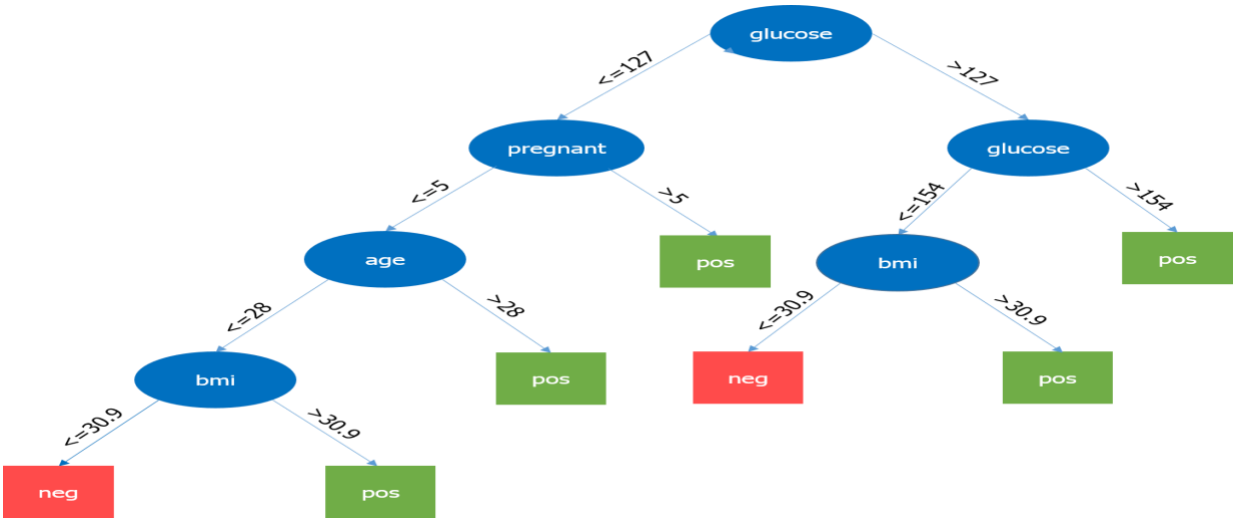

Figure 10: Decision graph representing distribution of data related to attributes.

From the above decision graph [16], [17], root node is glucose, as it is the main important parameter. Now, the current node and its value determine the next important parameter to be taken. This process goes on until the results in terms of ***pos*** or ***neg*** are extracted. ***Pos*** means the possibility of having diabetes is positive and ***neg*** means the possibility of having diabetes is negative.

Step 5:

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 10
October 2021

In this step, a small pilot project can be constructed to check if the results are appropriate. The performance constraints are also checked if any. If the results are not accurate, then the model re-plans and reconstruction is needed.

Step 6:

Once the project is executed successfully, the output is shared for deployment.

CONCLUSION

This paper mainly focuses on basics of Data Science and its implementation with a simple case study. Understanding and implementing data science can upgrade the knowledge, skills and business. In this chapter the authors explain about the fundamentals of data science with the insight of its life cycle and need for data science. And also discusses about the role and functions of data scientist. Differences between business intelligence and data science are discussed. Analysis and design is described in detail with the help of a case study: Diabetes Prevention.

REFERENCES

[1].What is data science?: a complete data science tutorial for beginners [Blog]. Retrieved 8. 10. 2019 from https://data-flair.training/blogs/what-is-datascience/. 3. Dhar, V. (2013).

[2]. Brodie, M.L. (2015). Understanding Data Science: An Emerging Discipline for Data-Intensive Discovery, in Shannon Cutt (ed.), Getting Data Right: Tackling the Challenges of Big Data Volume and Variety, O'Reilly Media, Sebastopol, CA, USA, June 2015.

[3]. Brodie, M.L. (2018a). What is Data Science? to appear in (Braschler, et. al. 2018).

[4]. Data science and prediction. Communications of the ACM, 56(12), 64– 73. 4. Foote, K. D. (2016). A brief history of data science. Retrieved 17. 10. 2019 from https://www.dataversity.net/brief-history-data-science/#. 5. Grossmann, W. and Rinderle-Ma, S. (2015).

[5]. Data science for business: what you need to know about data mining and data-analytic thinking (1 st ed.). Sebastopol: O'Reilly. 8. Provost, F. and Fawcett, T. (2013).

[6]. Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51–59. 9. Van der Aalst, W. (2016).

[7]. Data science in action. In W. van der Aalst, Process mining (pp. 3–23). Berlin; Heidelberg: Springer. 10. What is data science? [Blog]. (2019).

[8]. Fundamentals of business intelligence. Berlin; Heidelberg: Springer. 6. Merritt-Holmes, M. (2016).

[9]. Differences between data science and business intelligence. Retrieved 15. 10. 2019 from https://www.itproportal.com/2016/08/18/10- differences-between-data-science-and-business-intelligence/. 7. Provost, F. and Fawcett, T. (2013).

[10]. O. Kwon, N. Lee, and B. Shin. (2014) “International Journal of Information Management Data quality management, data usage experience and acquisition intention of big data analytics,” Int. J. Inf. Manage., vol. 34, no. 3, pp. 387–394

[11]. “Bytes to Bucks_ The Valuation of Data - HealthCare Appraisers.”.

[12]. K. Abouelmehdi, A. Beni-hssane, H. Khaloufi, and E. Nationale. (2017) “Science Direct Science Direct Big data security and privacy in healthcare: A Review Big data security and privacy in healthcare: A Review,” Procedia Comput. Sci., vol. 113, pp. 73–80.

[13]. V. Palanisamy and R. Thirunavukarasu. (2017) “Implications of big data analytics in developing healthcare frameworks – A review,” J. KingSaud Univ. - Comput. Inf. Sci.

[14]. Shixia Liu, Xiting Wang, Mengchen Liu and Jun Zhu, "Towards better analysis of machine learning models: A visual analytics perspective", Visual Informatics, vol. 1, pp. 48-56, 2017.

[15]. D. Ren, S. Amershi, B. Lee, J. Suh and J.D. Williams, "Squares: Supporting interactive performance analysis for multiclass classifiers", IEEE TVCG, vol. 23, no. 1, pp. 61-70, 2017.

[16]. B. Alsallakh, A. Hanbury, H. Hauser, S. Miksch and A. Rauber, "Visual methods for analyzing probabilistic classification data", IEEE TVCG, vol. 20, no. 12, pp. 1703-1712, 2014.

[17]. J Wang, S Fang, H Li, J Goni, AJ Saykin and L Shen, "Multigraph Visualization for Feature Classification of Brain Network Data", EuroVis Workshop on Visual Analytics (EuroVA), pp. 61-65, 2016.