DEEP INSIGHT TO DATA SCIENCE LIFE CYCLE USING HEALTH CARE DATA

Indumathi.U[#1], Dhivya.V[*2], Pream Sudha.V[#3]
[#]*Department of Computer Science (PG),*
PSGR Krishnammal college for women,
Bharathiar University.
[1]indumathy444@gmail.com
[2]dhivyavelusamy704@gmail.com
[3]**preamsudha@psgrkcw.ac.in**

***Abstract*— The paper aims to provide a systematic implementation of real time data in the life cycle of data science that can help to explain the concept of data science and through which it leads to reasonable performance gains. This work is grounded on past empirical works on research, and builds on the resource-based view. By identifying the basics through which the main areas of focus should be leveraged, this paper attempts to add to literature on how big data should be examined as a source of competitive advantage. This paper leads towards an inclusive research agenda by focusing on the interplay between data science and its life cycle with a real time case study.**

***Keywords*— Data science, Datafication, Life cycle of data science.**

I. INTRODUCTION

Data science is the civil engineering of data. The term Data Science is being used increasingly. As per a set of people data science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. This might be a point of view but not the exact definition. Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics and machine learning. Data Science is a more forward looking approach and an exploratory way with a focus to analyze the past or current data and predict the future outcomes with the aim of making informed decisions. The skills of data geeks include [1] :

- Statistics (traditional analysis)
- Data munging (parsing, scraping, and formatting data)
- Visualization (graphs, tools, etc.)

Datafication is a process of considering all aspects of life and converting them into data. For example, consider Google's augmented-reality glasses which datafy the gaze. The world is being datafied, or rather every actions are [2].

Even when browsing the web, the users are unintentionally or atleast passively, being datafied through cookies which is known or not known to them. Also while walking around in a store, or even on the street, the people are being datafied in a completely unintentional way via sensors, cameras, or Google glasses [3].

To date, prominence has been on the technical aspects of big data, with limited attention paid to the organizational changes they entail and how they should be leveraged strategically. Weka is a set of machine learning algorithms that can be applied to a data set directly, or called from Java code is named after flightless New Zealand bird Weka. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualisation using which the real time data set has been processed in this paper [4].

II. LIFE CYCLE OF DATA SCIENCE

Here is a brief overview of the main phases of the Data science life cycle in fig.1.

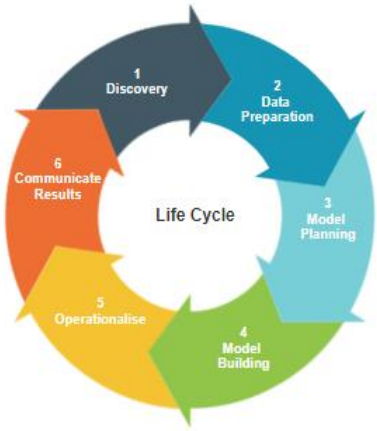

Fig.1 Life cycle of data science

A. Phase 1—Discovery

Initially before beginning the project, it is important to understand the various specifications, requirements, priorities and required budget. One must possess the ability to ask the right questions assess the required resources present in terms of people, technology, time, data etc to support the project.

B. Phase 2—Data preparation

In this phase, analytical sandbox in which analytics is performed is required for the entire duration of the project. Exploration, preprocessing and conditioning of data is required prior to modeling. Further, ETLT (extract, transform, load and transform) is performed to get data into the sandbox. Let's have a look at the fig.2 below for Statistical Analysis flow to frame the problem and formulate initial hypotheses (IH) to test.

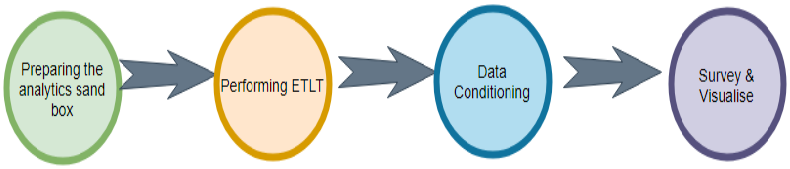

Fig.2 Data Preparation process

R can be used for data cleaning, transformation, and visualization. This will help to spot the outliers and establish a relationship between the variables. Once the data has been cleaned and prepared, it's time to do exploratory analytics on it.

C. Phase 3—Model planning

Here, the methods and techniques to draw the relationships between variables are determined. These relationships will set the base for the algorithms which will be implemented in the next phase [6]. Exploratory Data Analytics (EDA) will be applied using various statistical formulae and visualization tools. Following are the various model planning tools:

1. **R** is a software with complete set of modeling capabilities and provides a good environment for building interpretive models.
2. **SQL Analysis services** can perform in database analytics using common data mining functions and basic predictive models.
3. **SAS/ACCESS** can be used to access data from Hadoop and is used for creating repeatable and reusable model flow diagrams.

Though there exist many tools in the market, R is the most commonly used tool. After these phases the insight into the nature of the data and the algorithms to be used are decided.

D. Phase 4—Model building

In this phase, datasets for training and testing purposes are developed. Existing tool must be considered whether the tool will be suffice for running the models or it will need a more robust environment (like fast and parallel processing). Various learning techniques like classification, association and clustering to build the model are analysed. Model building can be achieved through the following tools:

- SAS Enterpise Miner
- WEKA
- SPSC Modeler
- Matlab
- Alpine
- Miner
- Statistica

E. Phase 5—Operationalise

In this phase, final reports, briefings, code and technical documents are delivered. In addition to this sometimes a pilot project is also implemented in a real-time production environment. This will provide a clear picture of the performance and other related constraints on a small scale before full deployment [5].

F. Phase 6—Communicate results

Now it is important to evaluate that the goals planned in the first phase has been achieved or not. So, in the last phase, all the key findings are identified, communicated to the stakeholders and the results of the project are determined whether they are success or failure based on the criteria developed in Phase 1.

III. CASE STUDY

Now, a real time data set of diabetes has been considered to explain the various phases described above:

A. Step 1

First, the data based on the medical history of patients is collected as discussed in Phase 1. The dataset consists of 8 attributes. Fig. 3 refers to the sample data given below:

```
@relation patient
@attribute npreg {1,2,3,4,5,6,7,8,9,10}
@attribute glu {148,85,89,78,197,166,118,103,126,119,96,109,88,99,97,102,90,111,180,106,171}
@attribute bp {73,65,81,50,70,72,84,30,88,80,66,75,58,78,60,76,68,71,64,92,110}
@attribute skin {39,29,23,32,45,19,47,38,41,35,15,26,11,31,33,37,42,49,25,18,24}
@attribute bmi {33.6,26.6,28.1,31,30.5,25.8,45.8,43.3,39.3,29,23.2,36,24.8,27.6,24,32.9,38.2,37.1,34,39,45.4}
@attribute ped {0.627,0.351,0.167,0.248,0.158,0.587,0.551,0.183,0.704,0.263,0.487,0.546,0.267,0.512,0.966,0.665,0.503,1.39,0.271,0.235,0.721}
@attribute age {50,31,21,26,53,51,30,39,25,29,20,60,22,45,33,46,27,56,24,48,54}
@data
%
6,148,73,39,33.6,0.627,50
1,85,66,29,26.6,0.351,31
1,89,81,23,28.1,0.167,21
3,78,50,32,31,0.248,26
2,197,70,45,30.5,0.158,53
5,166,72,19,25.8,0.587,51
7,118,84,47,45.8,0.551,30
1,103,30,38,43.3,0.183,39
3,126,88,41,39.3,0.704,25
9,119,80,35,29,0.263,29
1,96,66,15,23.2,0.487,20
5,109,75,26,36,0.546,60
3,88,58,11,24.8,0.267,22
10,99,78,31,27.6,0.512,45
4,97,60,33,24,0.966,33
9,102,76,37,32.9,0.665,46
2,90,68,42,38.2,0.503,27
4,111,71,49,37.1,1.39,56
3,180,64,25,34,0.271,24
7,106,92,18,39,0.235,48
9,171,110,24,45.4,0.721,54
```

Fig. 3 Sample data set

The various attributes as mentioned in the data are given below:

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

B. Step 2

Now, once the data is collected, it needs to be cleaned and prepared for data analysis.

- Here, the data have been organized into a single table under different attributes making it look more structured.
- This data has a lot of inconsistencies like missing values, blank columns, abrupt values and incorrect data format which need to be cleaned.

1) In the column *npreg*, "one" is written in words, whereas it should be in the numeric form like 1.
2) In column *bp* one of the values is 0 which is impossible for as bp cannot be 0 for an alive human being.
3) As you can see the *Income* column is blank and also has no use in predicting diabetes. As it is irrelevant to have it here it must be removed from the table.

Fig. 4 below shows a look at the structured sample data.

npreg	glucose	bp	skin	bmi	ped	age	income
6	148	72	35	33.6	0.627	50	
1	85	66	29	26.6	0.351	31	
8	183	64	0	23.3	0.672	32	
one	98	66	23	28.1	0.167	21	
0	137	40	35	43.1	2.288	33	
5	116	74	0	25.6	0.201	30	
3	78	50	32	31	0.248	26	
10	115	0	0	35.3	0.134	29	
2	197	70	45	30.5	0.158	53	
8	125	96	0	0	0.232	54	
4	110	92	0	37.6	0.191	30	
10	168	74	0	38	0.537	34	
10	139	80	0	27.1	1.441	57	
1	189	60	23	30.1	0.398	59	
5	166	72	19	25.8	0.587	51	
7	100	0	0	30	0.484	32	
0	118	84	47	45.8	0.551	31	
7	107	74	0	29.6	0.254	31	
1	103	30	38	43.3	0.183	33	
1	115	70	30	34.6	0.529	32	
3	126	88	41	39.3	0.704	27	
8	99	84	0	35.4	0.388	50	
7	196	90	0	39.8	0.451	41	
9	119	80	35	29	0.263	29	
11	143	94	33	36.6	0.254	51	
10	125	70	26	31.1	0.205	41	
7	147	76	0	39.4	0.257	43	
1	97	66	15	23.2	0.487	22	
13	145	82	19	22.2	0.245	57	
5	109	75	26	36	0.546	60	

Fig.4 Structured data set with missing, irrelevant and abrupt values

So, the data will be cleaned and preprocessed by removing the outliers, filling up the null values and normalizing the data type. Finally, we get the clean data as shown below in Fig. 5 which can be used for analysis.

npreg	glucose	bp	skin	bmi	ped	age
6	148	72	35	33.6	0.627	50
1	85	66	29	26.6	0.351	31
8	183	64	28.47	23.3	0.672	32
1	98	66	23	28.1	0.167	21
0	137	40	35	43.1	2.288	33
5	116	74	28.47	25.6	0.201	30
3	78	50	32	31	0.248	26
10	115	65	28.47	35.3	0.134	29
2	197	70	45	30.5	0.158	53
8	125	96	28.47	32.49	0.232	54
4	110	92	28.47	37.6	0.191	30
10	168	74	28.47	38	0.537	34
10	139	80	28.47	27.1	1.441	57
1	189	60	23	30.1	0.398	59
5	166	72	19	25.8	0.587	51
7	100	65	28.47	30	0.484	32
0	118	84	47	45.8	0.551	31
7	107	74	28.47	29.6	0.254	31
1	103	30	38	43.3	0.183	33
1	115	70	30	34.6	0.529	32
3	126	88	41	39.3	0.704	27
8	99	84	28.47	35.4	0.388	50
7	196	90	28.47	39.8	0.451	41
9	119	80	35	29	0.263	29
11	143	94	33	36.6	0.254	51
10	125	70	26	31.1	0.205	41
7	147	76	28.47	39.4	0.257	43
1	97	66	15	23.2	0.487	22
13	145	82	19	22.2	0.245	57
5	109	75	26	36	0.546	60

Fig. 5 Cleaned data set

C. Step 3

Now some analysis has to be done as discussed earlier in Phase 3. Here it is done using WEKA software. First, the data should be loaded into the analytical sandbox and various statistical functions must be applied on it. Then, visualization techniques like histograms, line graphs, box plots etc are used to get a fair idea of the distribution of data.

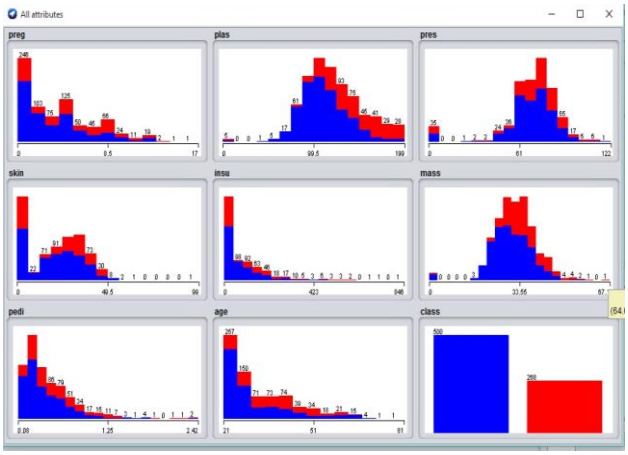

Fig. 6 Chart showing range of various attributes

D. Step 4

Now, based on insights derived from the previous steps, the best fit for this kind of problem must be recognised which is the decision tree.

- Since, the major attributes for analysis like *npreg*, *bmi*, etc., are already present supervised learning techniques are to be used to build a model here.

- Further, decision tree has been used particularly because it takes all attributes into consideration, like the ones which have a linear relationship as well as those which have a non-linear relationship. In this case, we have a linear relationship between *npreg* and *age*, whereas the nonlinear relationship between *npreg* and *ped*.
- Decision tree models are also very strong as it can be used for different combination of attributes to make various trees and then finally implement the one with the maximum efficiency.

Fig.7 shows the decision tree. Here, the most important parameter is taken as the root node which is the level of glucose. Now, the present node i.e., glucose and its value determine the next important parameters age, bmi which are taken into consideration. The process continues until the result is derived in terms of *pos* or *neg* which denotes the tendency of having diabetes is positive and the tendency of having diabetes is negative respectively.

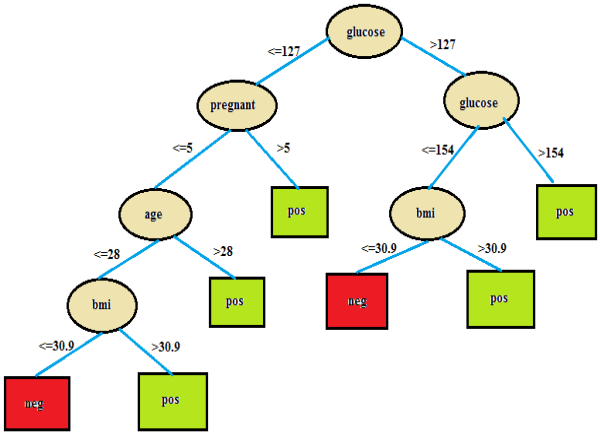

Fig. 7 Decision tree

E. Step 5

In this phase, performance constraints if any to find whether the results are appropriate must be viewed. If the results obtained are not accurate, then the project needs to be replanned and the model must be rebuilt.

F. Step 6

Once the project is executed, the output must be shared for full deployment.

IV.CONCLUSION

From the above case study, the criteria for determining the tendency of having diabetes both positive and negative are determined. Using this appropriate treatment can be given facilitating competitive performance gains. Thus following the phases of life cycle of data science the data can be processed appropriately leading to proper decision making. Once again big data proves to be an effective source of competitive advantage.

REFERENCES

[1] Shakhovska, Natalya. *Advances in Intelligent Systems and Computing*. Springer International Pu, 2017.
[2] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: big data concepts, methods, and analytics." *International journal of information management* 35.2 (2015): 137-144.5
[3] O'Neil, Cathy, and Rachel Schutt. *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.", 2013.
[4] Abbasi, Ahmed, Suprateek Sarker, and Roger HL Chiang. "Big data research in information systems: Toward an inclusive research agenda." *Journal of the Association for Information Systems* 17.2 (2016): I.
[5] Elhoseny, Hisham, et al. "A framework for big data analysis in smart cities." *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, Cham, 2018.
[6] Vines, Timothy H., et al. "The availability of research data declines rapidly with article age." *Current biology* 24.1 (2014): 94- 97.