Big Data Management

BIRMINGHAM CITY
University

Module Coordinator & Code:

Muhammad Afzal
CMP7203

Student Name & SID:

Muhammad Irfan
23173372

Submission Date:

13th May 2024

Contents

1 Introduction

The big data processing mainly is regarded as the set of different programming models and techniques that are utilized for accessing the large-scale data as well as extracting useful information in supporting along with providing decisions. In this era of Big Data, where the data is exponentially increasing across different fields (Mehdipour, Noori and Javadi, 2016). Majority of the data nowadays gets generated online either through social media platforms or online websites. The organizations are utilizing this data for getting relevant insights as well as in making the informed decisions for improving their profitability along with ensure the implementation of innovation (Sakr, Di Modica and Tomarchio, 2019). The Big Data analytics and processing have an impact on decision-making especially in the healthcare sector. As the healthcare managers utilize the large datasets, which are collected with the help of machine learning ML and statistical techniques, to identify the relevant patterns within data that allows them the ability to create the decision-making models (Hussain and Roy, 2019).

Similarly, in this big data management task the "diabetes_binary_health_indicator.csv" dataset is being analyzed utilizing the various data analytics techniques to find the impact of different health or general factors or indicators on the prevalence of diabetes in patients. In the data analysis at least 2 ML techniques are applied such as correlation analysis and regression analysis, along with the 2 data visualization techniques to present the findings from the analysis (Kumar and Kirthika, 2019). Furthermore, this report also provides the discussion of the Big Data properties along with the 3 processing paradigms. Moreover, the critical view that is beyond the course material with focus on the recent emerging ethical use with the use of Big Data is also provided in this report.

2 The 3 Big Data Processing Paradigms

The Big Data concept was originally linked with mainly three significant concepts regarded as the data variety, data volume and data velocity (Kumar and Kirthika, 2019). The analysis of Big Data presents the different challenges in data sampling and allowed previously for the observations and sampling of data. The three V's of the Big Data and the three dimensions of the Big Data are the defining properties in its processing (Ianni, Masciari and Sperlí, 2020). The processing of Big Data mainly includes the data extraction, data loading and data transformation ETL processes. As the data is collection from a myriad of sources that gets characterized based on the three V's of data velocity, data volume and the data variety (Ianni, Masciari and Sperlí, 2020).

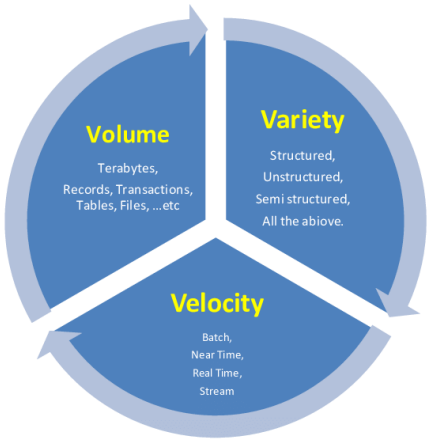


Figure 1: The 3 V's of Big Data (Mathrani and Lai, 2021).

The data volume generally refers to the amount of data, which is taken from different sources. Whereas, the data velocity represents the speed of data processing at which the Big Data is generated. However, the data variety mainly refers to different data types that are either structured, unstructured, or semi-structured data types (Mathrani and Lai, 2021). The Big Data processing paradigms includes the different techniques, which are used to handle, process, and evaluate large amounts of data as well as intricate datasets that are too large and complex for conventional data processing systems. The following are 3 main modern Big Data processing paradigms (Mathrani and Lai, 2021).

2.1 Batch Processing:

The first paradigm in Big Data is batch processing which is used to convert the data mainly in the form of batch views. The different data batches are basically created and then executed as iterative

algorithms like sorting, searching, and indexing algorithms. This batch processing is addressed through the use of a programming approach that is regarded as MapReduce (Pufahl and Weske, 2019). In batch processing, huge volumes and chunks of data get converted into the form of small batches and easily processed batches mainly analyzed with the assistance of different modes like Apache Pig, Hadoop MapReduce, and Apache Hive (Pufahl and Weske, 2019). As illustrated in figure-2 below, the Hadoop MapReduce basically works by splitting input datasets, after that it processes them independently with Map tasks, particularly in the entirely parallel manner (Pedamkar, 2019). Then it sorts the output and it is input to reduce the tasks. The job input as well as output get stored in the file systems and tasks are particularly monitored and scheduled in a framework as presented below.

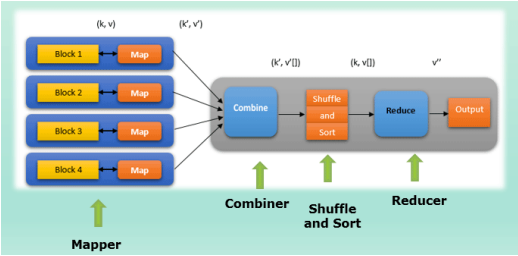


Figure 2: The batch processing with Hadoop MapReduce (Pedamkar, 2019).

2.2 Stream processing

The real-time stream processing is mainly about the continuous processing as well as analysis of the data in real-time particularly when the data gets created, ingested, and received from different sources that include the sensors, devices, applications, and platforms (Dias de Assunção, da Silva Veith and Buyya, 2018). The focus of stream processing is upon the analysis and processing of the data streams in real-time to enable real-time data analytics. This is the opposite of batch processing as in the stream processing the real-time processing of data happens (Dias de Assunção, da Silva Veith, and Buyya, 2018). For stream processing different tools like Apache Storm, Apache Spark and Apache Kafka are utilized. The following figure-3 shows the example of Stream Processing in real-time where action is taken on the data at time of it being published or generated (hazelcast, 2024).

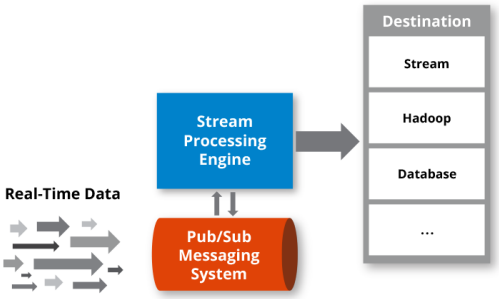


Figure 3: The real-time stream processing (hazelcast, 2024).

2.3 Interactive Processing

In graph theory, centrality measures assess the significance of nodes within a network. The Big Data processing paradigm of interactive processing is aimed toward the automatic and continuous processing of data as it is concerned with providing of users with the freedom and interaction in the evaluation and investigation of data in accordance with their requirements (teach-ICT, 2020). Through interactive processing, the user is required to provide the system with instructions to perform the processing.

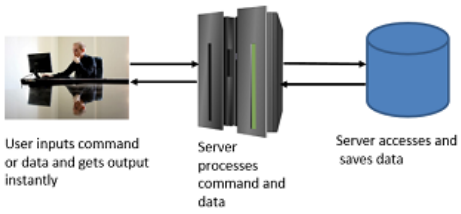


Figure 4: The interactive processing model (teach-ict, 2020).

3 The Real-Life Applications

3.1 The Healthcare Sector

The application of these paradigms in the healthcare sector, medical devices, which basically are part of the Internet of Things IoT are responsible for collecting and transmitting large amounts of patient data taken from various sources (Abouelmehdi, Hessane and Khaloufi, 2018). They utilize the batch processing mechanism in

the management of genomics datasets, healthcare electronic records, and patient data that is either taken from his medical history, interview with the GP, the wearables as well as the bio-sensors (Abouelmehdi, Hessane, and Khaloufi, 2018). This huge volume of healthcare data is stored mainly in the form of different batches or datasets and processed when one is required by the system. Like, for example, the company "Dignity Health" utilizes Big Data Analytics for preventing deadly infections by predicting the prevalence of infections using historical and real-time data from different regions (sas, 2019).

‘

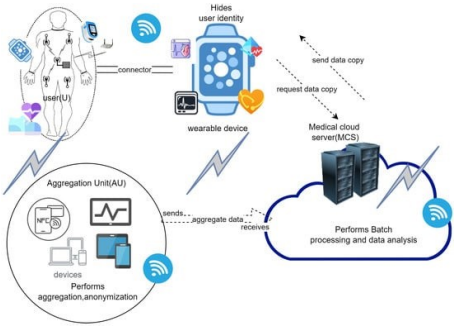


Figure 5: The real-time stream processing model implementation in Healthcare

3.2 The Cyber-security

The real-time stream processing is utilized in cybersecurity to deal with the ever-evolving and changing landscape of cyber threats (Club, 2023). The different cybersecurity software like NetScout, Sophos, Amazon CloudWatch, Microsoft Defender, and Bitdefender utilize real-time information about the latest cybersecurity threats such as viruses and malware, and update their program accordingly to provide real-time cybersecurity and threat protection (Club, 2023).

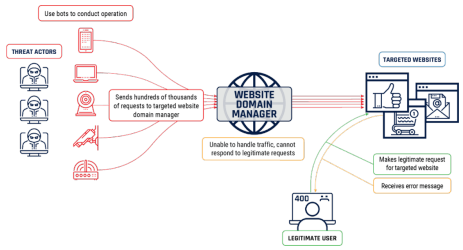


Figure 6: The real-time stream processing model implementation in Cybersecurity.

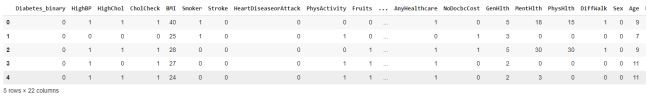
4 Big Data Processing Future in LLMs Era

In the large language model LLMs era, the future of Big Data processing is training the LLMs models on enormous datasets and assisting in sourcing of data from a myriad of sources such as books, articles, websites, and other text-based forms (Roh, Heo, and Whang, 2019). With the help of ML and AI the LLMs generate text, which is human-like content. These models can use Big Data to accelerate automate tasks like data preparation, exploration, and analysis to improve the model's decision-making and content creation (Roh, Heo, and Whang, 2019).

5 Exploratory Data Analysis

5.1 ETL Process

In exploratory data analysis, the data extraction, transformation, and loading process i.e. The ETL process is mainly regarded as a process that combines the data from multiple sources, particularly in the large and central repository known as the data warehouse (amazon, 2024). In this big data management assessment task the "diabetes_binary_health_indicator.csv" dataset extracted from the module portal was loaded in the Google Colab with the help of the Python code (amazon, 2024). Then in the transformation process, the dataset was cleaned and any missing values were removed. The following screenshot shows the "diabetes_binary_health_indicator.csv" is loaded in the Google Colab platform for performing the various data analytics tasks.



	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age
0	0	1	1	1	40	1	0	0	0	0	...	1	0	5	18	15	1	0	9
1	0	0	0	0	25	1	0	0	1	0	...	0	1	3	0	0	0	0	7
2	0	1	1	1	28	0	0	0	0	1	...	1	1	5	30	30	1	0	9
3	0	1	0	1	27	0	0	0	1	1	...	1	0	2	0	0	0	0	11
4	0	1	1	1	24	0	0	0	1	1	...	1	0	2	3	0	0	0	11

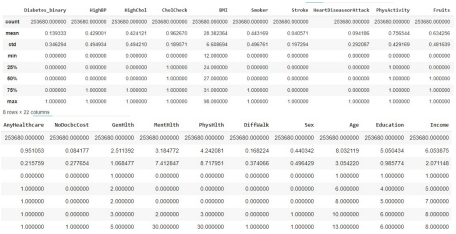
5 rows × 22 columns

Figure 7: The "diabetes_binary_health_indicator.csv" in Google Colab (Google, 2024).

5.2 Descriptive Statistics

After performing the ETL process in this exploratory data analysis of the sourced dataset, the required is to analyze the dataset using the

different data analytics techniques for finding the impact of different health or general factors or indicators on the prevalence of diabetes in patients (Kalton and Conway, 2019). The descriptive statistic mainly is the summary statistic, which quantitatively summarizes and describes the features of the dataset such as the variability, central tendency, and distribution. Such methods are effective in providing the data overview and help in identifying the relationships and patterns within the data (Bickel and Lehmann, 2019).



	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000
mean	0.139333	0.429001	0.424121	0.962670	28.382364	0.443169	0.040571	0.094186	0.756544	0.634256
std	0.346294	0.494934	0.494210	0.189571	6.608694	0.496761	0.197294	0.292087	0.429169	0.481639
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000	0.000000	1.000000	0.000000
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000	0.000000	1.000000	1.000000
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	1.000000	1.000000	1.000000

8 rows × 22 columns

AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000
0.951053	0.084177	2.511392	3.184772	4.242081	0.168224	0.440342	8.032119	5.050434	6.053875
0.215759	0.277654	1.068477	7.412847	8.717951	0.374066	0.496429	3.054220	0.985774	2.071148
0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
1.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	6.000000	4.000000	5.000000
1.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	8.000000	5.000000	7.000000
1.000000	0.000000	3.000000	2.000000	3.000000	0.000000	1.000000	10.000000	6.000000	8.000000
1.000000	1.000000	5.000000	30.000000	30.000000	1.000000	1.000000	13.000000	6.000000	8.000000

Figure 8: The Descriptive Statistics (Google, 2024).

The above illustrated descriptive statistics output in figure-8, shows that the majority of the patients assessed in this dataset did not have diabetes only 35,347 patients suffering from diabetes in the total sample of 253,680 patients making only 0.14 patients in this sample having diabetes on average (Bickel and Lehmann, 2019). The descriptive statistics output above not only provides the mean value of these variables but also the standard deviation, minimum, maximum, 25%, 50%, and 75% values as well. The pie chart in Figure-9 below shows that in this dataset approximately 86.1% were the participants that had no diabetes and 13.9% were the participants that had diabetes.

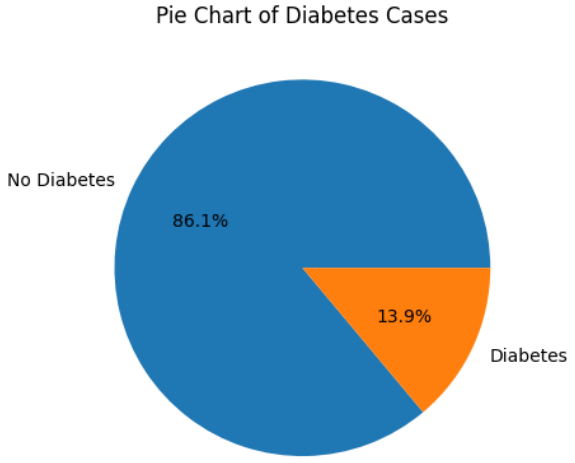


Figure 9: The Pie Chart of Diabetes Cases

The histogram depicted in Figure-10 provides a visual representation of the Body Mass Index (BMI) distribution across the study population, derived from the dataset analyzed. This chart is essential for understanding the weight status categories that may affect health outcomes, especially as they relate to diabetes prevalence observed in the dataset. The majority of the data clusters around a BMI of 25 to 30, indicative of an overweight population, with a pronounced right skew extending towards higher BMI values. This skew suggests a significant portion of the population is categorized as obese, which is consistent with global health concerns regarding increasing obesity rates. The right-skewed distribution shown here highlights the presence of extreme values or outliers in higher BMI ranges, up to 100, which are less frequent but notable for their potential health impacts. The sharp peak and distribution around the modal value near 30 further signify that the median BMI lies in the overweight category, reinforcing concerns about obesity-related health risks in this cohort.

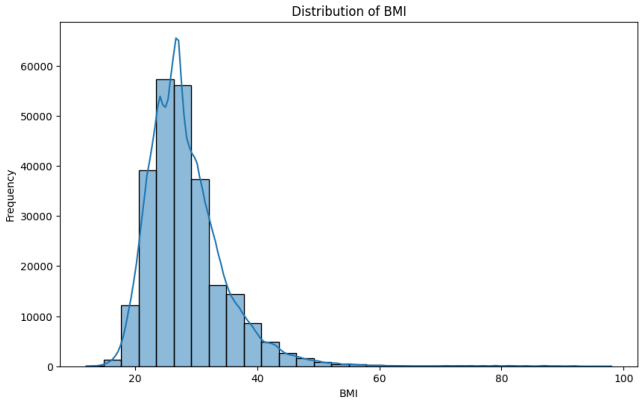


Figure 10: Distribution of BMI

5.3 Scatter Plots

The scatter graph or scatter plot is a statistical data visualization technique that utilizes dots to represent values for 2 different numeric variables (BYJU'S, 2022). Each dot gets positioned either on the vertical or horizontal axis, indicating the values of individual data points.

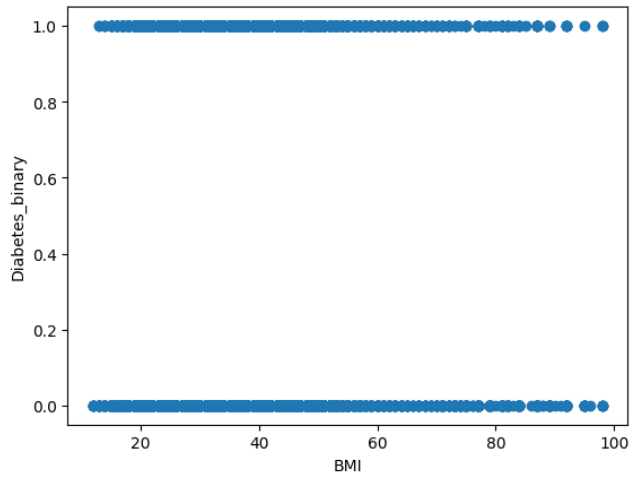


Figure 11: The scatter plot of Diabetes_binary and BMI variables (Google, 2024).

The figure-11 above presents the scatter plot of the two variables of Diabetes_binary and BMI in this dataset. This scatter plot is utilized for observing the association among these two variables. The scatter plot output shows that there is somewhat a positive association between the increase in BMI and the prevalence of diabetes among patients. The findings of Gray et al. (2015), research shows the fact that the rise BMI is associated with a significant rise in hypertension, dyslipidemia, and diabetes mellitus prevalence. Another study also showed that the diabetes prevalence was positively associated with the increase in BMI among participants as obese individuals with BMI greater than 25 kg/m2, had a 30.4% chance of diabetes whereas lower BMI individuals had a 5% chance of diabetes (Medhi et al., 2021).

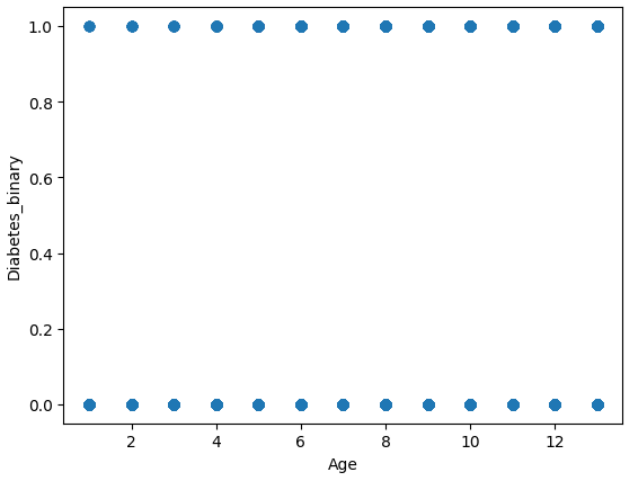


Figure 12: The scatter plot of Diabetes_binary and Age variables (Google, 2024)

The figure-12 above presents the scatter plot of the two variables of Diabetes_binary and Age in this dataset. This scatter plot is utilized for observing the association between these two variables wit its output showing that the prevalence of diabetes among participants increased with age. The research also shows that with age the chances of an individual having Type 2 diabetes greatly increase (Gray et al. 2015). The figure-13 below presents the scatter plot of the two variables of Diabetes_binary and General Health in this dataset. This output shows that there is a negative association with the prevalence of diabetes among patients with their General Health. This is also backed by literature evidence as research shows that patients who have good healthcare and diet are less likely to suffer from Type 2 diabetes as compared to patients who have bad healthcare and diet (Medhi et al., 2021).

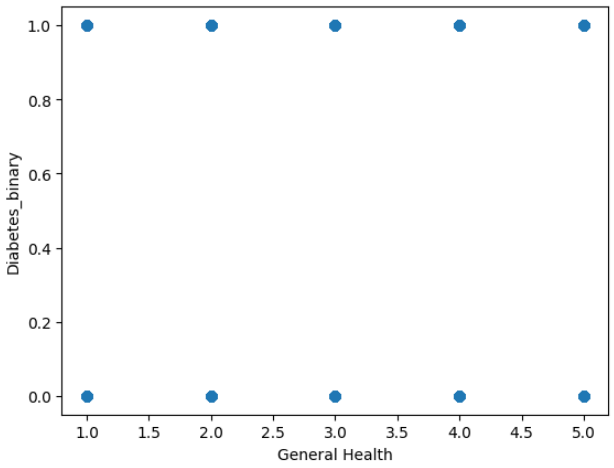


Figure 13: The scatter plot of Diabetes_binary and General Health variables (Google, 2024).

5.4 Heat Map

Another advanced data visualization technique utilized in this assessment is the heat map, which is the data graphical representation in which values are depicted with the help of colors. This graphical representation helps in understanding the association among the different variables in the dataset in a visually appealing manner. The following figure-16 presents the heat map of the correlation analysis performed in this study.

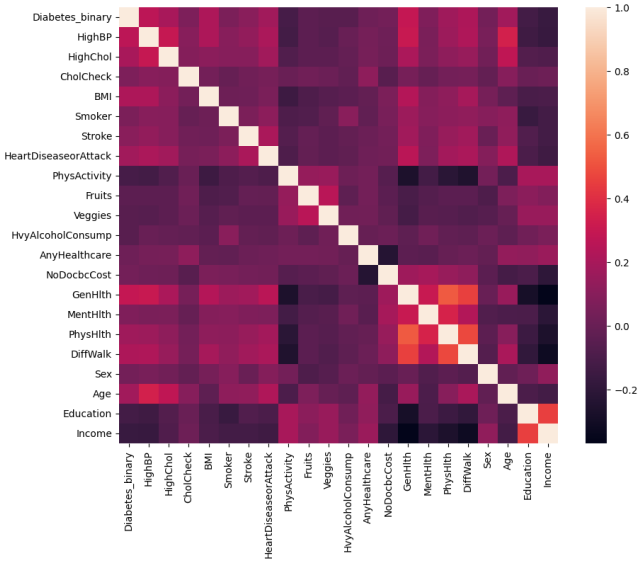


Figure 14: The heat map of correlation (Google, 2024).

Basically, the darker shade in this heat map represents little to no association between the dependent variable of "Diabetes_binary" and the independent variables of this study. Whereas, the lighter color represents some association between these variables in the dataset. From this heat map, it can be easily interpreted that the dependent variable of "Diabetes_binary" has somewhat of an association with the independent variables of HighBP, HighChol, BMI, HeartDiseaseorAttack, GenHlth, PhysHlth, DiffWalk, and Age respectively (Medhi et al., 2021). The UK National Institute of Health NIH study also revealed that the key causes of Type 2 Diabetes are obesity, overweight or high BMI and physical inactivity (NIDDKD, 2016).

6 Machine Learning ML Models

6.1 Correlation

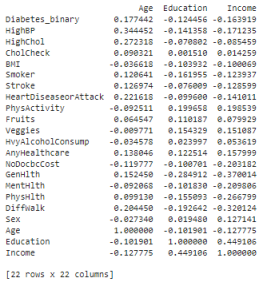
The first statistical analysis machine learning ML model being utilized in this assessment is the correlation analysis, which is also known as the bi-variate analysis and it is primarily concerned mainly with finding out of whether the association among the variables in the dataset exists or not (Sherpa, 2023). It also helps in determining the action and magnitude of the relationship present among the variables in a dataset. The figure-15 below presents the output of the correlation analysis of the "diabetes_binary_health_indicator.csv" dataset that was conducted with the help of the relevant Python code in the Google Colab platform.

	Diabetes_binary	HighBP	HighChol	CholCheck	\
Diabetes_binary	1.000000	0.263129	0.200276	0.064761	
HighBP	0.263129	1.000000	0.298199	0.098508	
HighChol	0.200276	0.298199	1.000000	0.085642	
CholCheck	0.064761	0.098508	0.085642	1.000000	
BMI	0.216843	0.213748	0.106722	0.034495	
Smoker	0.060789	0.096991	0.091299	-0.009929	
Stroke	0.105816	0.129575	0.092620	0.024158	
HeartDiseaseorAttack	0.177282	0.209361	0.180765	0.044206	
PhysActivity	-0.118133	-0.125267	-0.078046	0.004190	
Fruits	-0.040779	-0.040555	-0.040859	0.023849	
Veggies	-0.056584	-0.061266	-0.039874	0.006121	
HvyAlcoholConsump	-0.057056	-0.003972	-0.011543	-0.023730	
AnyHealthcare	0.016255	0.038425	0.042230	0.117626	
NoDocbcCost	0.031433	0.017358	0.013310	-0.058255	
GenHlth	0.293569	0.300530	0.208426	0.046589	
MentHlth	0.069315	0.056456	0.062069	-0.008366	
PhysHlth	0.171337	0.161212	0.121751	0.031775	
DiffWalk	0.218344	0.223618	0.144672	0.040585	
Sex	0.031430	0.052207	0.031205	-0.022115	
Age	0.177442	0.344452	0.272318	0.090321	
Education	-0.124456	-0.141358	-0.070802	0.001510	
Income	-0.163919	-0.171235	-0.085459	0.014259	

	BMI	Smoker	Stroke	HeartDiseaseorAttack
Diabetes_binary	0.216843	0.060789	0.105816	0.177282
HighBP	0.213748	0.096991	0.129575	0.209361
HighChol	0.106722	0.091299	0.092620	0.180765
CholCheck	0.034495	-0.009929	0.024158	0.044206
BMI	1.000000	0.013804	0.020153	0.052904
Smoker	0.013804	1.000000	0.061173	0.114441
Stroke	0.020153	0.061173	1.000000	0.203002
HeartDiseaseorAttack	0.052904	0.114441	0.203002	1.000000
PhysActivity	-0.147294	-0.087401	-0.069151	-0.087299
Fruits	-0.087518	-0.077666	-0.013389	-0.019790
Veggies	-0.062275	-0.030678	-0.041124	-0.039167
HvyAlcoholConsump	-0.048736	0.101619	-0.016950	-0.028991
AnyHealthcare	-0.018471	-0.023251	0.008776	0.018734
NoDocbcCost	0.058206	0.048946	0.034804	0.031000
GenHlth	0.239185	0.163143	0.177942	0.258383
MentHlth	0.085310	0.092196	0.070172	0.064621
PhysHlth	0.121141	0.116460	0.148944	0.181698
DiffWalk	0.197078	0.122463	0.176567	0.212709
Sex	0.042950	0.093662	0.002978	0.086096
Age	-0.036618	0.120641	0.126974	0.221618
Education	-0.103932	-0.161955	-0.076009	-0.099600
Income	-0.100069	-0.123937	-0.128599	-0.141011

	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost
Diabetes_binary	-0.118133	-0.040779	...	0.016255	0.031433
HighBP	-0.125267	-0.040555	...	0.038425	0.017358
HighChol	-0.078046	-0.040859	...	0.042230	0.013310
CholCheck	0.004190	0.023849	...	0.117626	-0.058255
BMI	-0.147294	-0.087518	...	-0.018471	0.058206
Smoker	-0.087401	-0.077666	...	-0.023251	0.048946
Stroke	-0.069151	-0.013389	...	0.008776	0.034804
HeartDiseaseorAttack	-0.087299	-0.019790	...	0.018734	0.031000
PhysActivity	1.000000	0.142756	...	0.035505	-0.061638
Fruits	0.142756	1.000000	...	0.031544	-0.044243
Veggies	0.153150	0.254342	...	0.029584	-0.032232
HvyAlcoholConsump	0.012392	-0.035288	...	-0.010488	0.004684
AnyHealthcare	0.035505	0.031544	...	1.000000	-0.232532
NoDocbcCost	-0.061638	-0.044243	...	-0.232532	1.000000
GenHlth	-0.266186	-0.103854	...	-0.040817	0.166397
MentHlth	-0.125587	-0.068217	...	-0.052707	0.192107
PhysHlth	-0.219230	-0.044633	...	-0.008276	0.148998
DiffWalk	-0.253174	-0.048352	...	0.007074	0.118447
Sex	0.032482	-0.091175	...	-0.019405	-0.044931
Age	-0.092511	0.064547	...	0.138046	-0.119777
Education	0.199658	0.110187	...	0.122514	-0.100701
Income	0.198539	0.079929	...	0.157999	-0.203182

	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	\
Diabetes_binary	0.293569	0.069315	0.171337	0.218344	0.031430	
HighBP	0.300530	0.056456	0.161212	0.223618	0.052207	
HighChol	0.208426	0.062069	0.121751	0.144672	0.031205	
CholCheck	0.046589	-0.008366	0.031775	0.040585	-0.022115	
BMI	0.239185	0.085310	0.121141	0.197078	0.042950	
Smoker	0.163143	0.092196	0.116460	0.122463	0.093662	
Stroke	0.177942	0.070172	0.148944	0.176567	0.002978	
HeartDiseaseorAttack	0.258383	0.064621	0.181698	0.212709	0.086096	
PhysActivity	-0.266186	-0.125587	-0.219230	-0.253174	0.032482	
Fruits	-0.103854	-0.068217	-0.044633	-0.048352	-0.091175	
Veggies	-0.123066	-0.058884	-0.064290	-0.080506	-0.064765	
HvyAlcoholConsump	-0.036724	0.024716	-0.026415	-0.037668	0.005740	
AnyHealthcare	-0.040817	-0.052707	-0.008276	0.007074	-0.019405	
NoDocbcCost	0.166397	0.192107	0.148998	0.118447	-0.044931	
GenHlth	1.000000	0.301674	0.524364	0.456920	-0.006091	
MentHlth	0.301674	1.000000	0.353619	0.233688	-0.080705	
PhysHlth	0.524364	0.353619	1.000000	0.478417	-0.043137	
DiffWalk	0.456920	0.233688	0.478417	1.000000	-0.070299	
Sex	-0.006091	-0.080705	-0.043137	-0.070299	1.000000	
Age	0.152450	-0.092068	0.099130	0.204450	-0.027340	
Education	-0.284912	-0.101830	-0.155093	-0.192642	0.019480	
Income	-0.370014	-0.209806	-0.266799	-0.320124	0.127141	



	Age	Education	Income
Diabetes_binary	0.177442	-0.124456	-0.163919
HighBP	0.344452	-0.141358	-0.171235
HighChol	0.272318	-0.070802	-0.085459
CholCheck	0.090321	0.001510	0.014259
BMI	-0.036618	-0.103932	-0.100069
Smoker	0.120641	-0.161955	-0.123937
Stroke	0.126974	-0.076009	-0.128599
HeartDiseaseorAttack	0.221618	-0.099600	-0.141011
PhysActivity	-0.092511	0.199658	0.198539
Fruits	0.064547	0.110187	0.079929
Veggies	-0.009771	0.154329	0.151087
HvyAlcoholConsump	-0.034578	0.023997	0.053619
AnyHealthcare	0.138046	0.122514	0.157999
NoDocbcCost	-0.119777	-0.100701	-0.203182
GenHlth	0.152450	-0.284912	-0.370014
MentHlth	-0.092068	-0.101830	-0.209806
PhysHlth	0.099130	-0.155093	-0.266799
DiffWalk	0.204450	-0.192642	-0.320124
Sex	-0.027340	0.019480	0.127141
Age	1.000000	-0.101901	-0.127775
Education	-0.101901	1.000000	0.449106
Income	-0.127775	0.449106	1.000000

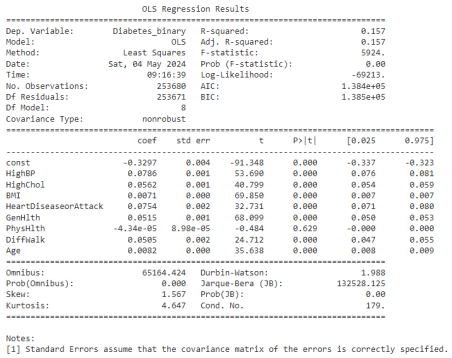
[22 rows x 22 columns]

Figure 15: The correlation analysis (Google, 2024).

The results of this correlation analysis suggest that the main dependent variable of this study, which is the “Diabetes_binary” variable, shows that it has a positive association with the independent variables of HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, Age, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, AnyHealthcare and NoDocbcCost respectively (Sherpa, 2023). Whereas, the main dependent variable of “Diabetes_binary” has negative correlation with the independent variables of PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, Education, and Income respectively.

6.2 Linear Regression Analysis

The next machine learning ML statistical model utilized in this Big Data analytics assessment was the Linear regression analysis. The linear regression analysis is basically the statistical analysis method that is utilized for predicting the variable value based on another variable’s value. The predicted variable in the linear regression analysis model is the dependent variable (IBM, 2024). The variable you are using to predict the other variable’s value is called the independent variable, which in this case is the “Diabetes_binary” variable. Whereas, the key independent variables or predictor variables according to the findings from the correlation analysis are HighBP, HighChol, BMI, HeartDiseaseorAttack, GenHlth, PhysHlth, DiffWalk, and Age respectively having a correlation of 0.15 or higher with the dependent variable (WHO, 2024). The figure-16 below presents the complete linear regression analysis output that was generated using the relevant Python code in the Google Colab platform. The complete code of generating this respective output in the Google Colab environment is presented in the appendix section of this report. Considering the R-squared value of this regression analysis it is clear that almost 0.157 or 15.7% variability observed mainly in the dependent or target variable gets explained by this regression model (CFI, 2023).



OLS Regression Results

Dep. Variable:	Diabetes_binary	R-squared:	0.157
Model:	OLS	Adj. R-squared:	0.157
Method:	Least Squares	F-statistic:	5924.
Date:	Sat, 04 May 2024	Prob (F-statistic):	0.00
Time:	09:16:39	Log-Likelihood:	-69213.
No. Observations:	253680	AIC:	1.384e+05
Df Residuals:	253671	BIC:	1.385e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.3297	0.004	-91.348	0.000	-0.337	-0.323
HighBP	0.0786	0.001	53.690	0.000	0.076	0.081
HighChol	0.0562	0.001	40.799	0.000	0.054	0.059
BMI	0.0071	0.000	69.850	0.000	0.007	0.007
HeartDiseaseorAttack	0.0754	0.002	32.731	0.000	0.071	0.080
GenHlth	0.0515	0.001	68.099	0.000	0.050	0.053
PhysHlth	-4.34e-05	8.98e-05	-0.484	0.629	-0.000	0.000
DiffWalk	0.0505	0.002	24.712	0.000	0.047	0.055
Age	0.0082	0.000	35.638	0.000	0.008	0.009

Omnibus:	65164.424	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	132528.125
Skew:	1.567	Prob(JB):	0.00
Kurtosis:	4.647	Cond. No.	179.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 16: The linear regression analysis output (Google, 2024).

Furthermore, the coefficient values suggest that the independent variables of HighBP, HighChol, HeartDiseaseorAttack, GenHlth, and DiffWalk are the most statistically significant in predicting the prevalence of Diabetes among patients according to this analysis. This means that the change in these independent variables of HighBP, HighChol, HeartDiseaseorAttack, GenHlth, and DiffWalk definitely causes the change in the dependent variable of Diabetes_binary respectively (CDC, 2022).

6.3 K-Means Cluster Analysis

The k-means clustering analysis basically is the algorithm, that groups similar objects, particularly into the groups that are called clusters. The cluster analysis endpoint is the set of clusters, in which each cluster generally is distinct mainly from the other cluster, along the objects in every cluster are generally similar to one another (Sherpa, 2023). In this analysis, the python code as provided in the appendix section of this report was used in Google Colab, for finding the K-Means clustering to assess the intrinsic groups that are within the dataset along with drawing inferences among them. The following figure-17 below presents the output of the K-means clustering analysis among the variables of “Diabetes_binary” and “BP” in this dataset. The output clearly indicates that the

majority of the data in the case of these key variables is clustered around two points i.e. 1 and 0, indicating that the patients with diabetes have the chances of high blood pressure BP respectively.

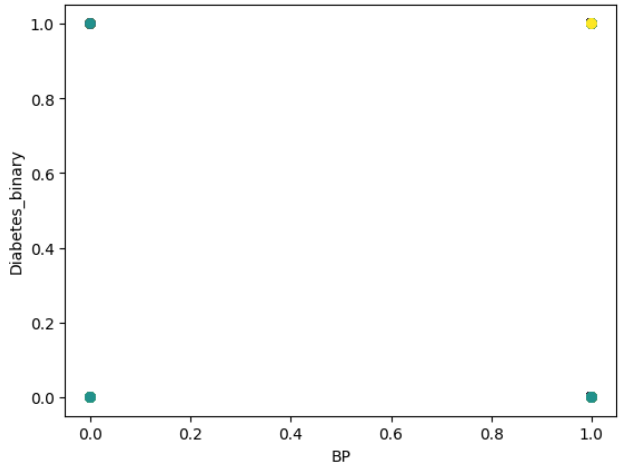


Figure 17: The K-means cluster analysis output (Google, 2024).

As research shows that diabetes is capable of causing damage mainly by scarring the kidneys which in turn leads to salt as well as water retention, which as a result raises blood pressure (CDC, 2022). Furthermore, with the course of time diabetes damages a patient's small blood vessels, which causes the blood vessel walls to get stiffen as well as not function properly. Such changes are capable of contributing to the higher blood pressure issue among diabetes patients (Bickel and Lehmann, 2019).

6.4 Decision Tree Analysis

The decision tree analysis mainly is a process to calculate the decision tree equation and the graphic representation mainly of the different alternative solutions, which are available for resolving the given problem for determining the highly effective future course of action. The code for calculating the decision-tree and is graphical output is presented in the appendix section of this report. The following figure-18 shows there is 95% probability that this model is accurate in its prediction of whether a patient has diabetes or not.

```
[0 0 0 ... 0 0 0]
Predicted      0     1
Actual
0          19123    79
1           1035  2236
0.9504294041738975
```

Figure 18: The decision tree analysis output (Google, 2024).

The figure-19 below presents the graphical representation of this decision tree but it is too large.

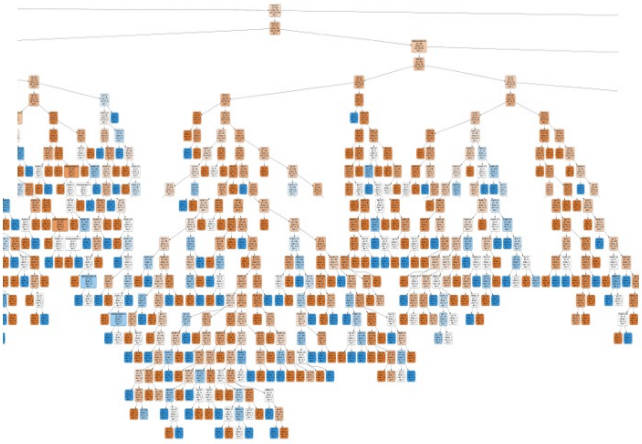


Figure 19: The decision tree graphical representation (Google, 2024).

7 Graph Analysis

In recent years, the field of graph analysis has witnessed significant advancements, facilitated by the evolution of graph database technologies such as Neo4j. Graph databases, characterized by their ability to efficiently model and query relationships between data points, are increasingly pivotal in areas ranging from social network analysis to bioinformatics and financial networks. Neo4j offers a robust platform for developing graph-based solutions, leveraging its Cypher query language and scalable architecture.

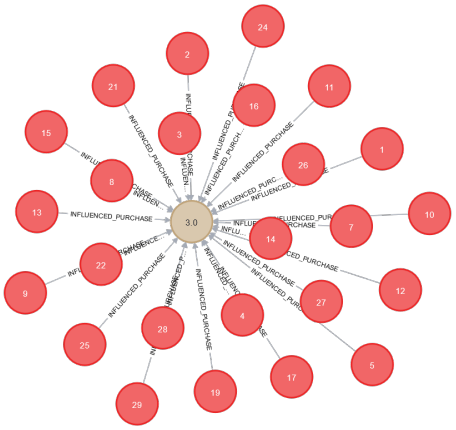


Figure 20: Influenced purchase via Ads

The figure 20 depicts the results of the relationship between “INFLUENCED_PURCHASE” Where the node represents the customer in this context customers are players of the game. The directed edges show the influence one entity had on another’s purchasing. This analysis plays a vital role in understanding consumer behavior and its purchasing decisions. The node in the center with the value “3.0”, appears to be a most influential node in the network which shows its significance and impact on other consumer decisions. This analysis helps businesses to optimize their marketing strategies and enhance their customer experience and engagement.

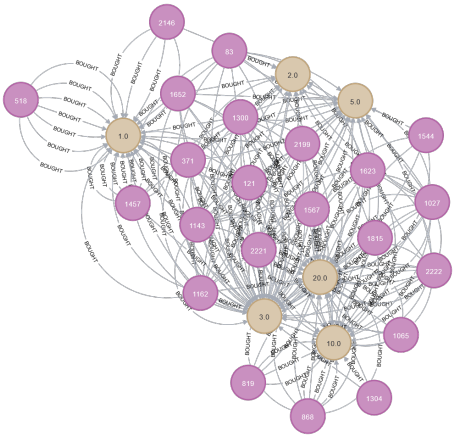


Figure 21: Items Bought Via ads by consumer

The figure 21 illustrates the network analysis, which consists of the “BOUGHT” created using Neo4j queries. It helps to visualize the purchases which went through the ads. Each node here depicts the player. The directed edges symbolize the transaction of buying product. Nodes such as 1.0,2.0.30 are highlighted with in the network, these entities are central buyers which influencing or interlinked with multiple other nodes. However, such purchasing patterns help businesses to identify key consumers of product also show the direct and indirect influences to the buying of products.

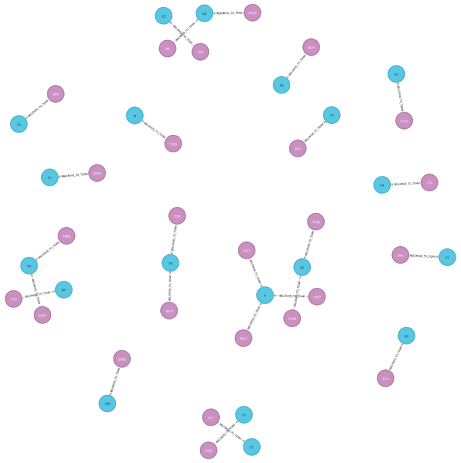


Figure 22: Team player relationship

The figure 22 visualizes the relationship between entities based on a “BELONGS_TO_TEAM” With in a network. Each node signifies an individual player, and the edges denote the relationship and affiliation with the specific teams. The nodes are color-coded, suggesting the classification of the entities into different groups. This structure helps in understanding how the entities a grouped in larger organizational context. Such insights are valuable for analyzing organizational structures, enhancing team dynamics, and optimizing collaboration across different segments of the business.

8 Role of Ethics in Big Data

There is a huge role of ethics in Big Data processing and management, because the main ethical dilemmas that is faced by Big Data Analytics generally revolve around 3 critical aspects of data management i.e. data security, data privacy and bias. The data privacy concerns mainly stem from the collection of huge amounts of personal information to be analyzed by governments or organizations (IABAC®, 2023). Furthermore, with rise in the use of Big Data, the potential for cyber-attacks, data theft and invasion of the individual’s privacy does escalate. The company’s do acquire the individual’s personal data either from various online sources like social media sites or blogs for different purposes, which

raises concerns about data privacy, data ownership, personal consent, and information control (Howe III and Elenberg, 2020). Furthermore, data security issues are also prevalent with Big Data as data gets stored on cloud platforms that constantly face the issue of cyberattacks and data breaches.

Ethically it is the responsibility of the government and organizations to protect the personal and private information of customers and show high accountability, safety and responsibility in management of the customer's personal data that they collect on the regular basis. Furthermore, the bias in data analytics must be removed along with any sort of discrimination so that any sort of unfairness or discrimination can be removed in areas such as law enforcement, hiring, lending and marketing

9 Conclusion & Recommendations

The main aim and purpose of this Big Data processing assessment report was to not only provide a discussion of the Big Data properties along with the 3 processing paradigms But also statistically analyze the "diabetes_binary_health_indicator.csv" dataset through the utilization of various data analytics techniques to find the impact of different health or general factors or indicators on the prevalence of diabetes in patients. In this data analysis task, at least 2 ML techniques were applied i.e. correlation analysis and regression analysis as well as the 2 data visualization techniques of scatter plot and heat map to present the findings from the analysis in a visually appealing manner. The three key big data processing paradigms discussed in this report were batch processing, real-time stream processing, and iterative processing. Then the discussion on the real-life application of these processing paradigms in the healthcare sector and the cybersecurity sector was provided along with the discussion upon role of big data processing in the LLMs era.

Furthermore, after performing the exploratory data analysis the correlation and linear regression analysis techniques were utilized for assessing the association between the main dependent variable of "Diabetes_binary" in this study and the key independent variables according to the findings from the correlation analysis of HighBP, HighChol, BMI, HeartDiseaseorAttack, GenHlth, PhysHlth, DiffWalk and Age respectively. Considering the R-squared value of the regression analysis it is clear that almost 0.157 or 15.7% variability observed mainly in the dependent or target variable gets explained by this regression model. Furthermore, the coefficient values suggest that the independent variables of HighBP, HighChol, HeartDiseaseorAttack, GenHlth, and DiffWalk are the most statistically significant in predicting the prevalence of Diabetes among patients according to this analysis. This means that the change in these independent variables of HighBP, HighChol, HeartDiseaseorAttack, GenHlth, and DiffWalk definitely causes the change in the dependent variable of Diabetes_binary respectively.

10 References

1. Abouelmehdi, K., Hessane, A.B. and Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. Journal of Big Data, [online] 5(1). doi:https://-doi.org/10.1186/s40537-017-0110-7.

2. amazon (2024). What is ETL? - Extract Transform Load Explained - AWS. [online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/what-is/etl/#: :text=Extract%2C%20trans-form%2C%20and%20load%20

3. García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. Springer.

4. Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2), 111-117.

5. Benjelloun, S., Aissi, M.E.M.E., Loukili, Y., Lakhrissi, Y., Ali, S.E.B., Chougrad, H. and Boushaki, A.E. (2020). Big Data Processing: Batch-based processing and stream-based processing. [online] IEEE Xplore. doi:https://-doi.org/10.1109/ICDS50568.2020.9268684

6. Bickel, P.J. and Lehmann, E.L. (2019). Descriptive Statistics for Nonparametric Models I. Introduction. The Annals of Statistics, 3(5), pp.1038–1044. doi:https://-doi.org/10.1214/aos/1176343239.

7. BYJU'S (2022). Scatter Plot - Definition, Examples and Correlation. [online] BYJUS. Available at: https://byjus.com/maths/scatter-plot/.

8. CDC (2022). Diabetes risk factors. [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/diabetes/basics/risk-factors.html.

9. CFI (2023). R-Squared. [online] Corporate Finance Institute. Available at: https://corporatefinanceinsti-tute.com/resources/data-science/r-squared/#: :text=The%20most%20com-mon%20interpretation%20of..

10. Club, T.C. (2023). 25 Best Cybersecurity Software of 2023. [online] The CTO Club. Available at: https://thectoclub.com/tools/best-cybersecurity-software/.

11. Dias de Assunção, M., da Silva Veith, A. and Buyya, R. (2018). Distributed data stream processing and edge computing: A survey on resource elasticity and future directions. Journal of Network and Computer Applications, 103(1), pp.1–17. doi:https://-doi.org/10.1016/j.jnca.2017.12.001.

12. Google (2024). Google Colab. [online] colab.research.google.com. Available at: https://colab.research.google.com/-drive/15QfajXAUx-fakK7HMZ1xwUQIgVrieYxVP?usp=shar-ing

    [Accessed 13 May 2024].

13. Gray, N., Picone, G., Sloan, F. and Yashkin, A. (2015). Relation between BMI and Diabetes Mellitus and Its Complications among US Older Adults. Southern Medical Journal, [online] 108(1), pp.29–36. doi:https://-doi.org/10.14423/smj.0000000000000214.

14. hazelcast (2024). Real-Time Stream Processing. [online] Hazelcast. Available at: https://hazelcast.com/glossary/real-time-stream-processing/.

15. Howe III, E.G. and Elenberg, F. (2020). Ethical Challenges Posed by Big Data. Innovations in Clinical Neuroscience, [online] 17(10-12), pp.24–30. Available at: https://www.ncbi.nlm.nih.gov/pmc/arti-cles/PMC7819582/.

16. Hussain, A. and Roy, A. (2019). The emerging era of Big Data Analytics. Big Data Analytics, 1(1). doi:https://-doi.org/10.1186/s41044-016-0004-2.

17. IABAC® (2023). The Ethical Implications of Big Data Analytics. [online] IABAC®. Available at: https://iabac.org/blog/the-ethical-implications-of-big-data-

analytics: :text=The%20ethical%20dilemmas%20of%20Big%20Data%20Analytics%20revolve%20around%20three.

18. Ianni, M., Masciari, E. and Sperlí, G. (2020). A survey of Big Data dimensions vs Social Networks analysis. Journal of Intelligent Information Systems, 57(1), pp.73–100. doi:https://doi.org/10.1007/s10844-020-00629-2.

19. IBM (2024). What Is Linear Regression? | IBM. [online] www.ibm.com. Available at: https://www.ibm.com/topics/linear-regression: :text=IBM-.

20. Kalton, G. and Conway, F. (2019). Descriptive Statistics. Applied Statistics, 12(3), p.195. doi:https://doi.org/10.2307/2985799.

21. Kumar, S.S. and Kirthika, Ms.V. (2019). Big Data Analytics Architecture and Challenges, Issues of Big Data Analytics. International Journal of Trend in Scientific Research and Development, Volume-1(Issue-6), pp.669–673. doi:https://doi.org/10.31142/ijtsrd4673.

22. Mathrani, S. and Lai, X. (2021). Big Data Analytic Framework for Organizational Leverage. Applied Sciences, 11(5), p.2340. doi:https://doi.org/10.3390/app11052340.

23. Medhi, G.K., Dutta, G., Borah, P., Lyngdoh, M. and Sarma, A. (2021). Prevalence of Diabetes and Its Relationship With Body Mass Index Among Elderly People in a Rural Area of Northeastern State of India. Cureus, 1(1). doi:https://doi.org/10.7759/cureus.12747.

24. NIDDKD (2016). Symptoms Causes of Diabetes | NIDDK. [online] National Institute of Diabetes and Digestive and Kidney Diseases. Available at: https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes: :text=Overweight%2C%20obesity%2C%20and%20physical%20inactivity.

25. Pedamkar, P. (2019). How MapReduce Work? | Working And Stages Of MapReduce. [online] EDUCBA. Available at: https://www.educba.com/how-mapreduce-work/.

26. Pufahl, L. and Weske, M. (2019). Batch activity: enhancing business process modeling and enactment with batch processing. Computing, 101(12), pp.1909–1933. doi:https://doi.org/10.1007/s00607-019-00717-4.

27. Roh, Y., Heo, G. and Whang, S.E. (2019). A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. IEEE Transactions on Knowledge and Data Engineering, 33(4), pp.1–1.

28. Sakr, S., Di Modica, G. and Tomarchio, O. (2019). Editorial for Special Issue of Journal of Big Data Research on 'Geo-distributed Big Data Processing and Management'. Big Data Research, 16(1), p.59. doi:https://doi.org/10.1016/j.bdr.2019.05.001.

29. sas (2019). Big data in health care. [online] Sas.com. Available at: https://www.sas.com/en_us/insights/articles/big-data/big-data-in-healthcare.html.

30. Sherpa, S.T. (2023). Correlation - Correlation Co-efficient, Types and Formulas. [online] BYJUS. Available at: https://byjus.com/maths/correlation/.

31. teach-ict (2020). Computer Science learning for school students. [online] Teach-ict.com. Available at: https://www.teach-ict.com/-glossary/I/interactive%20processing.htm#: :text=Interactive%20processing%20means%20that%20the.

32. WHO (2024). Diabetes. [online] www.who.int. Available at: https://www.who.int/health-topics/diabetes?gad$_source = 1gclid = Cj0KCQjwudexBhDKARIsAI - GWYWXkB4DdURPaNMePdvXRO6NqERqLq Bxwe3kaAuB6EALw_wcBtab = tab_1$

[Accessed 4 May 2024].

33. Mehdipour, F., Noori, H. and Javadi, B. (2016). Energy-Efficient Big Data Analytics in Datacenters. Advances in

Computers, 100(1), pp.59–101. doi:https://-doi.org/10.1016/bs.adcom.2015.10.002.

34. Robinson, I., Webber, J. and Eifrem, E., 2013. Graph Databases. O'Reilly Media, Inc

35. Jordan, G., 2014. Practical Neo4j. Apress.doi:https://doi.org/10.1016/bs.ad-com.2015.10.002.

36. Bullmore, E. and Sporns, O., 2009. The application of graph theoretical analysis to complex networks in the brain. Neuroscience, 10(3), pp.186-198.

11 Appendix

GitHub repository for the code.

https://github.com/MuhammadIrfan5/bigaatamanagementneo