Big Data in Healthcare Policy and Management

Kaushal Kishor Gupta[1], Dr Rajesh Keshavrao Deshmukh[2]

[1]*Research Scholar, Department of Computer Science and Engineering, Kalinga University, Raipur (C.G.)*

[2]*Assistant Professor, Department of Computer Science and Engineering Kalinga University, Naya Raipur, Chhattisgarh, India Year 2023*

Abstract

The integration of Big Data Analytics (BDA) into healthcare marks a significant shift toward data-driven methodologies in patient care and health management. This shift is driven by the rapid expansion in the volume, speed, and diversity of data within the healthcare industry, including electronic health records (EHRs), genomic information, and data from wearable technologies. BDA holds immense potential for the healthcare sector, offering notable enhancements in patient care outcomes, operational efficiency, and the customization of healthcare services. The digital transformation of healthcare systems through information systems, medical technologies, and portable and smart wearable devices presents numerous challenges to researchers and healthcare providers, including issues related to data storage, reducing treatment costs, and processing time (to derive valuable insights, reduce error rates, and make optimal decisions). The findings of this systematic review reveal that, despite significant efforts in the field of healthcare big data analytics, there remains a need for newer hybrid systems that leverage machine learning.

Index Terms *Big Data, Health records, Machine learning*

I Introduction

This paper presents a critical examination of the adoption and impact of Big Data Analytics (BDA) within Polish healthcare facilities, highlighting the nuanced interplay between structured and unstructured data in medical practices. Structured data, with its predefined schema, contrasts sharply with the more nebulous nature of unstructured data, or Big Data (BD), which defies traditional data processing approaches due to its sheer volume and lack of organization [1]. This dichotomy underscores the necessity for specialized technologies and methodologies to harness the potential of BD, transforming it into actionable insights that can significantly benefit healthcare organizations. The manuscript is pioneering in its approach, offering a dual perspective that merges a comprehensive review of existing literature on BD and BDA with empirical research focused on the practical application of big data analytics in Poland's medical sector [2]. This endeavour not only sheds light on the theoretical underpinnings of Big Data's role in healthcare but also provides a tangible analysis of its real-world implications, particularly in a Polish context. Healthcare's intricate ecosystem, encompassing a wide array of stakeholders and governed by stringent regulations, is at a critical juncture. The traditional doctor-patient paradigm is evolving towards a more collaborative model that emphasizes patient engagement and preventative care, a shift further catalysed by the challenges presented by the Covid-19 pandemic. The pandemic has not only underscored the importance of digital health solutions and patient data access but also highlighted the pressing need for a healthcare model that can adapt to the demographic shifts of an aging population and declining fertility rates. The paper posits that leveraging BDA extends beyond technological implementation; it requires a holistic transformation encompassing healthcare process management, service design, and business models to meet the changing needs of patients, clinicians, and healthcare administrators. Despite the growing prevalence of BDA in various sectors, the healthcare industry faces distinct challenges in integrating and utilizing Big Data to its full potential, thus calling for innovative strategies to bridge the gap between data capabilities and healthcare delivery [3]. In summary, this study not only marks the first comprehensive analysis of Big Data's

multifaceted roles within the Polish healthcare system but also serves as a call to action for healthcare providers to embrace the potential of BDA. By doing so, it aims to revolutionize patient care, enhance operational efficiencies, and foster a more proactive, patient-centric healthcare ecosystem.

II Big Data and Healthcare

Big data analytics is increasingly vital to IT. Many companies, like banks and supermarket stores, have adopted big data to boost income. However, big data analytics in healthcare presented several obstacles [4]. Because of this, healthcare is slipping behind other industries. Big data in healthcare is needed for several reasons right now. Thus, big data analytics in healthcare attempts to preserve immense potential. Big data has enabled precision medicine, which has improved public health, but it has not yet reached a wider audience. Hospital services are difficult to get for low-income and underdeveloped communities because to consumer and geographical obstacles. Drug development and use strive to enhance public health and reach more people. Drug development include target identification, hit discovery, hit-to-lead generation, lead optimization, and pre-clinical drug candidate identification.

The research found that pharmaceutical drugs that take 10–17 years and 2.6 billion dollars to develop are rejected. Figure 1, shows the primary reasons not to use big data analytics in medicine and healthcare.

Fig 1 Primary concept for evaluation

medicine development is high-risk, high-cost, long-cycle, and competitive, therefore corporations assessed patients' ability to cut research and development expenses, preventing much-needed medicine development in some developing nations. Computer-aided technology has helped solve this problem. Basic theoretical literature and survey findings show that big data has improved drug R&D. Cochrane and Galperin feel data building is essential to life science basic research. Technology experts argue that huge drug candidate data sets and various patient clinical reactions imply that current drug development has entered the big data age. Deep learning and big data modeling will help researchers develop medication safety and effectiveness evaluation methods [5]. A large amount of chemical and biological data gives computer-aided drug design an increasingly prominent role in the application, which allows for rapid drug discovery and development, reducing costs and improving efficiency. In conclusion, the big data medical platform will support R&D and medical industry optimization, which will benefit medication development and public health.

III Stages of Big data Analysis

The implementation of Big Data Analytics (BDA) in any sector, including healthcare, typically follows a structured process divided into several key stages. Each stage is critical in transforming vast amounts of raw data into actionable insights.

Data Collection

The first stage involves gathering data from various sources. In healthcare, this could include electronic health records (EHRs), medical imaging, genomic sequences, wearable device data, and more. The goal is to collect a comprehensive set of data that can be analyzed to glean insights.

Data Preparation

Once data is collected, it needs to be cleaned and organized. This stage involves removing errors, inconsistencies, or irrelevant data, and integrating data from different sources. The preparation stage is crucial for ensuring the quality and reliability of the data analysis.

Data Storage

After preparation, the data must be stored in a manner that facilitates easy access and analysis. This often involves using databases or data lakes that can handle large volumes of structured and unstructured data [6]. Effective data storage solutions are scalable and secure to accommodate the growing size and sensitivity of data.

Data Analysis

This stage involves applying statistical models, machine learning algorithms, and data mining techniques to the prepared and stored data to identify patterns, trends, and correlations. The specific methods used can vary widely depending on the objectives of the analysis, ranging from simple descriptive statistics to complex predictive or prescriptive analytics.

Data Visualization and Interpretation

The insights gained from data analysis are often complex and not immediately intuitive. Data visualization tools are used to create graphs, charts, and dashboards that make the results more accessible and understandable. This stage is crucial for communicating findings to stakeholders who may not have a technical background.

Actionable Insights and Decision Making

The ultimate goal of Big Data Analytics is to inform decision-making. The insights derived from the data need to be actionable, meaning they can guide strategic decisions, policy formulation, and operational improvements. In healthcare, this could mean identifying risk factors for disease, optimizing treatment plans, improving patient outcomes, or enhancing operational efficiency [7].

Monitoring and Feedback

After implementing changes based on insights from BDA, it's important to monitor the outcomes and collect feedback. This stage involves assessing the effectiveness of the decisions made and identifying areas for further improvement. Monitoring and feedback ensure that the analytics process is iterative, with each cycle providing opportunities for refinement and greater insights.

IV Experimental Outcomes for proposed model

We used the public Pima Indians Diabetes Database for our project. Various diabetes diagnostic measures are in this dataset. The data came from KAGGLE. All known cases are of individuals over 21. Our suggested model comprises additional stages (Fig. 4).

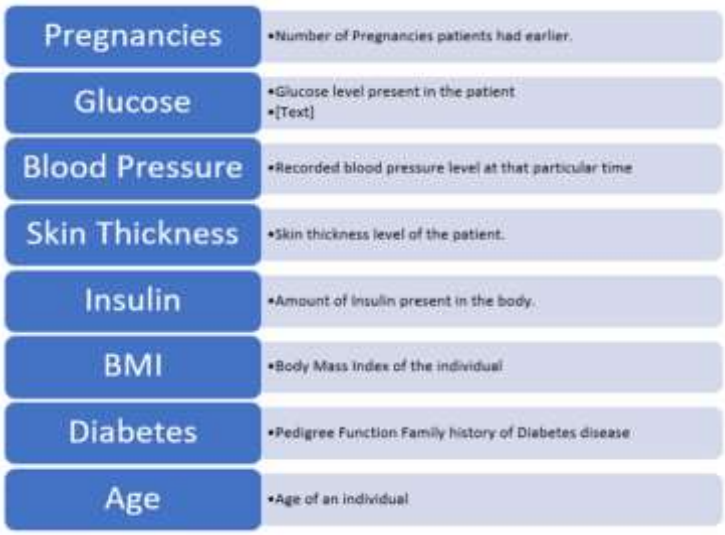

Figure 3: Instances Variable

The figure 3 have provided appears to be a visual representation of different medical parameters that are commonly used in the prediction of diabetes, particularly type 2 diabetes. Here's a breakdown of each parameter and its relevance to diabetes analysis and prediction:

- **Pregnancies:** The number of times a patient has been pregnant. This is relevant because gestational diabetes is a type of diabetes that develops during pregnancy, and women who have had gestational diabetes or have given birth to a baby weighing more than 9 pounds are at higher risk for developing type 2 diabetes later in life [8].
- **Glucose:** The glucose level present in the patient's blood. This is a direct measure of blood sugar levels and is the primary diagnostic marker for diabetes. High levels of glucose could indicate an inability to process sugar effectively, which is a hallmark of diabetes.
- **Blood Pressure:** The level of blood pressure at that particular time. While not a direct indicator of diabetes, high blood pressure is common in people with diabetes and can complicate the management of the condition, leading to increased risk of cardiovascular disease.
- **Skin Thickness:** The skinfold thickness (usually measured at the triceps). This can be an indirect measure of body fat, which can correlate with insulin resistance—a condition in which the body's cells do not respond normally to insulin.
- **Insulin:** The amount of insulin present in the body. In a patient with diabetes, the pancreas either produces little or no insulin (type 1 diabetes) or the cells do not respond appropriately to the insulin that is produced (type 2 diabetes).
- **BMI:** Body Mass Index of the individual. BMI is a measure of body fat based on height and weight. A higher BMI can indicate obesity, which is a significant risk factor for diabetes.
- **Diabetes Pedigree Function**: This represents a family history of diabetes disease. It's a function that scores the likelihood of diabetes based on family history. A higher score indicates a higher risk of developing diabetes.
- **Age:** The age of the individual. Risk of type 2 diabetes increases with age, especially after age 45.

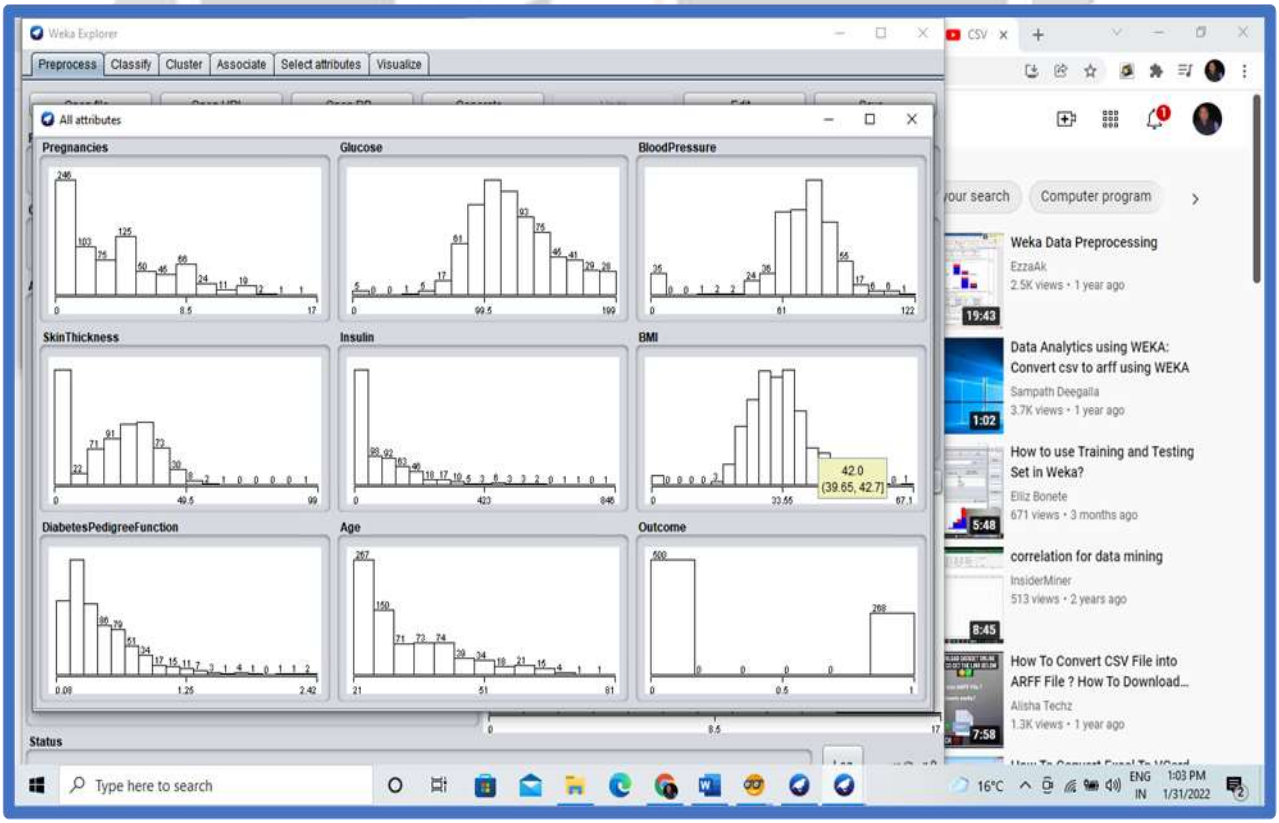

Figure 4: Data Visualization

- For continuous variables like Glucose, Blood Pressure, BMI, and Age, you can see a range of values spread out over the possible range. For example, the Glucose histogram shows a concentration of values in the middle range, suggesting that most individuals have glucose levels around this range, with fewer individuals having very low or very high levels.
- For categorical or discrete variables like Pregnancies, the data is grouped into distinct bars where each bar represents the number of times the event occurred [9]. For example, a large number of individuals have 0 or 1 pregnancies, while fewer have 8 or more.
- The Diabetes Pedigree Function, which is a score calculated based on family history, shows a skewed distribution, indicating most individuals have a low score, with fewer individuals having higher scores.

Principal Component Analysis (PCA):

Principal Component Analysis (PCA) is one of the design acknowledgment approaches, and one of its uses is to explore high-dimensional information, which is not anything but difficult to grasp by just looking at a large amount of information. This is one of the applications of PCA. In order to do an analysis of the data (shown in figure 4.3), we must first transform high-level information into low-level measurement, following which we must plot the data and interpret the results. PCA is applied to simplify the presentation of the relevant data into a few basic plots, particularly the score plot and the stacking plot. When it comes to research, it might be challenging to summarize a massive amount of material. The principal component analysis (PCA) computation is used to find out the link between the enormously correlated informative index.

This Figure 5, shows a screenshot of a software tool used for Principal Component Analysis (PCA), which is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Component Variance (red line): This line descends sharply, reflecting the amount of variance that each principal component captures from the data. The first principal component has a high value, indicating that it captures a substantial amount of the total variance in the dataset [10]. The exact values are not visible for each point, but the first point on the red line appears to be very close to 1, which would mean that the first component captures nearly 100% of the variance. This is an indicator of a strong dominant component.

Cumulative Variance (yellow line): This line increases as more components are added, showing the total variance captured cumulatively by all principal components considered together. The first point on the yellow line, corresponding to the first principal component, has a value of approximately 0.958 or 95.8%. This means that the first principal component alone accounts for about 95.8% of the total variance in the dataset.

Subsequent Components: After the first component, there is a marked flattening of the curve, indicating that additional components contribute significantly less to capturing the variance. For instance, the second principal component adds a very small amount to the cumulative variance (the exact value is not visible but is very close to 0.958), suggesting that it provides little new information.

Interpretation: The values suggest that the dataset is dominated by the first principal component, which captures most of the variance. This could imply that one underlying factor or a combination of factors strongly influences the dataset. When the cumulative variance line flattens and additional components do not significantly increase the percentage of the variance captured, it often indicates that those components are less informative.

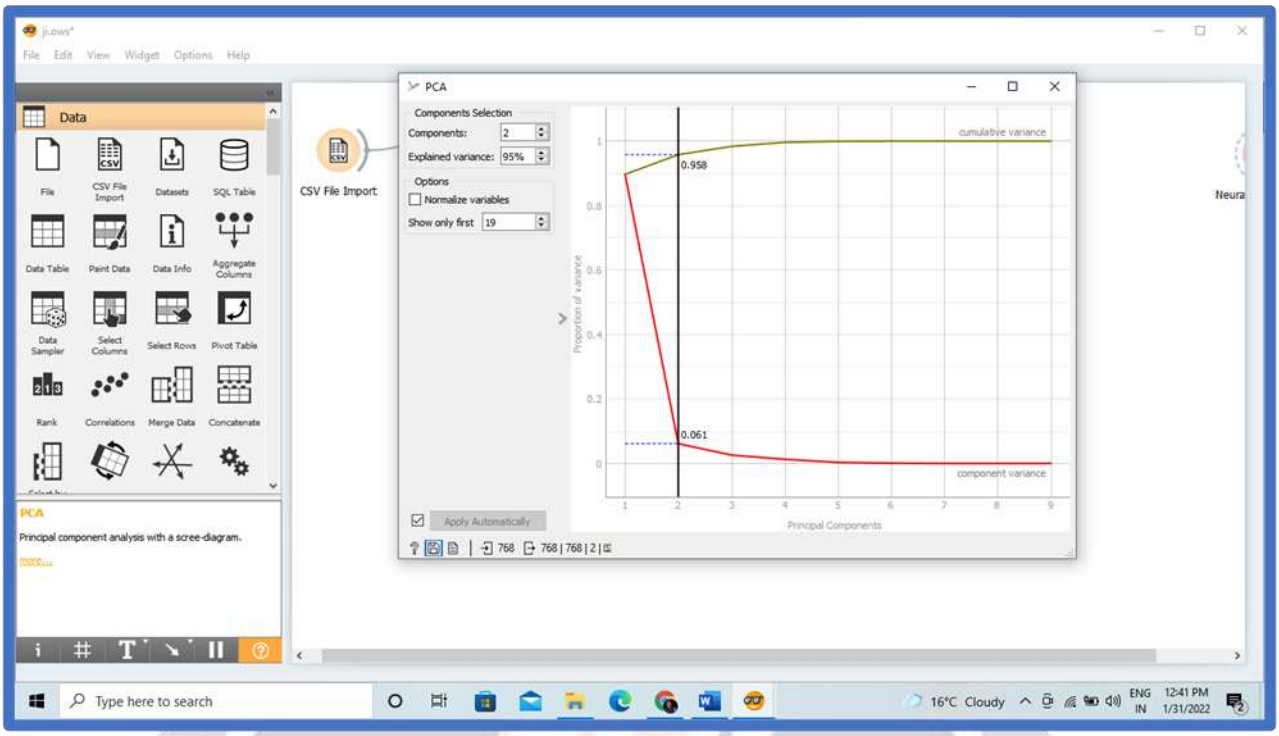

Figure 5: PCA

Discussion

The study in question employs a robust approach to predicting the likelihood of an individual transitioning from a state of normal glycaemic control to prediabetes, and from prediabetes to diabetes within the span of a year. By leveraging a substantial dataset, the research applies ensemble machine learning techniques which are known for their ability to improve prediction accuracy through the combination of multiple models to address the complexity and heterogeneity of medical data. The researchers have emphasized the importance of temporal changes in medical data, indicating that the progression of these metabolic states is not only a matter of static, current values but also their change over time. The study reveals that cumulated medical data—essentially the medical history of a patient—can significantly impact the accuracy of predictions. This suggests that the trajectory of change in medical parameters is crucial in understanding the progression towards diabetes. In the process of data analysis, the research has employed data-driven feature selection methods to pinpoint the most influential predictors within the dataset. This process of feature selection is pivotal as it determines the efficiency and accuracy of the predictive models. The identification of nine key characteristics that, when added to the predictive models, outperform the standard predictors, underscores the value of these features. Such features may include but are not limited to, changes in glucose levels, BMI variations over time, and other metabolic indicators that have been traditionally used to assess diabetes risk. The study's conclusion emphasizes the critical role of dynamic monitoring of glucose levels and BMI in clinical settings. It suggests that doctors should not only rely on static, absolute values of these measures but also consider their progression over time to make more informed predictions about a patient's health trajectory. This could mean more regular monitoring and the use of sophisticated analytical tools that can track and interpret these changes over time. Extending the conclusion, it is evident that the incorporation of longitudinal data analysis in predictive modelling represents a significant advancement in medical diagnostics. The study's findings could pave the way for the development of more nuanced and personalized health monitoring systems, potentially allowing for earlier interventions and more effective management of prediabetes and diabetes. Additionally, the research highlights the potential for machine learning models to assimilate complex and time-series data in a way that could revolutionize predictive healthcare, leading to more proactive and preventive approaches in medicine. It also calls for a revaluation of current clinical practices, encouraging a shift towards a more holistic and dynamic assessment of health risks rather than a reliance on static snapshots of patient data.

References

1. Meyer, “Support Vector Machines – The Interface to libsvm in package e1071”, August 2015
2. S. S. Shwartz, Y. Singer, N. Srebro, “Pegasos: Primal Estimated sub - Gradient Solver for SVM”, Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007
3. X. L. Dong and F. Naumann. Data fusion: Resolving data conflicts for integration. PVLDB, 2(2):1654–1655, Aug. 2009.
4. X. L. Dong and D. Srivastava. Big data integration. Synthesis Lectures on Data Management, 7(1):1–198, 2015.
5. P. Fellegi and A. B. Sunter. A theory for record linkage. Journal of the Americal Statistical Association, 64(328):1183–1210, 1969.
6. H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. pages 371–380, 2001.
7. J. Gao, Q. Li, B. Zhao, W. Fan, and J. Han. Mining reliable information from passively and actively crowdsourced data. In KDD, pages 2121–2122, 2016.
8. L. Getoor and A. Machanavajjhala. Entity resolution: theory, practice & open challenges. PVLDB, 5(12):2018–2019, 2012.
9. C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu. Corleone: Hands-off crowdsourcing for entity matching. In SIGMOD, pages 601–612, 2014.
10. Mining in Continuous Data for Diabetes Prediction. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1209-1214). IEEE.