Performance evaluation of real time statistical analysis of NoSQL databases and big data - a case study with clinical data.

1st Given Name Surname

*dept. name of organization (of Aff.) name of organization (of Aff.)*

City, Country

email address or ORCID

4th Given Name Surname

*dept. name of organization (of Aff.) name of organization (of Aff.)*

City, Country

email address or ORCID

2nd Given Name Surname

*dept. name of organization (of Aff.) name of organization (of Aff.)*

City, Country

email address or ORCID

5th Given Name Surname

*dept. name of organization (of Aff.) name of organization (of Aff.)*

City, Country

email address or ORCID

3rd Given Name Surname

*dept. name of organization (of Aff.) name of organization (of Aff.)*

City, Country

email address or ORCID

6th Given Name Surname

*dept. name of organization (of Aff.) name of organization (of Aff.)*

City, Country

email address or ORCID

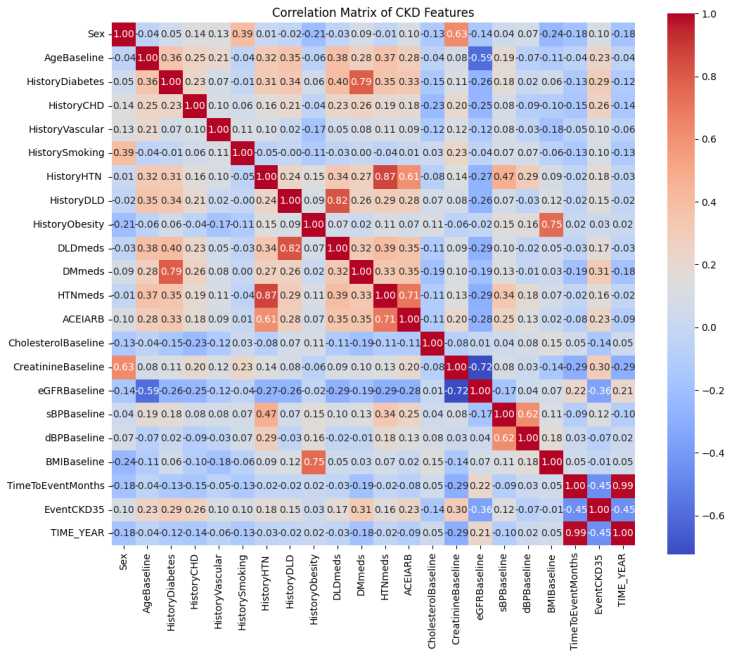
*Abstract*—This document is a model and instructions for LATEX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. \*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

*Index Terms*—component, formatting, style, styling, insert.

Introduction

Chronic Kidney Disease (CKD) presents significant challenges in early-stage prediction; however, machine learning models show substantial promise for facilitating early detection. While numerous machine learning approaches can aid this process, it is crucial to select the most effective model, as accuracy is fundamental to generating reliable predictions. This study aims to evaluate the real-time performance and predictive accuracy of various machine learning models for early CKD detection, utilizing a large clinical dataset stored within IndexedDB, a NoSQL database environment. By analyzing how effectively these models manage big data, this research sheds light on the advantages and limitations of real-time predictive analysis, ultimately enhancing clinical decision-making related to CKD detection.The analysis is conducted using a dataset comprising 900,000 entries generated in IndexedDB, which serves as the foundation for this study. One of the main challenges is identifying the suitable machine learning algorithm, as performance typically diminishes when the dataset expands. This decline is often attributed to the requirement for simultaneous processing of multiple tasks, which can slow down computations and negatively impact accuracy, particularly in large NoSQL datasets. Moreover, the presence of missing values can contribute to

inaccuracies in predictions, frequently arising from high data volumes or potential errors during the data generation process in IndexedDB. To ensure precise performance evaluation and timely analysis, it is essential to rigorously validate the completeness of data entries. Counting records and confirming data integrity are critical steps in verifying that all entries are accurately represented and that no data is absent, which is vital for achieving dependable results in real-time performance assessments.In our research, two machine learning algorithms—Logistic Regression and Support Vector Machines (SVM)—yielded encouraging results when analyzing clinical data stored in a NoSQL framework. However, as real-time data updates and expands, these models frequently encounter obstacles in maintaining high accuracy and effective performance due to the complexities introduced by the dynamic nature of big data. Visualization tools such as bar plots, line charts, and other graphical representations play an essential role in analyzing the timing and performance metrics of these models, enabling quick insights. Nevertheless, complications such as missing or inconsistent data can lead to unexpected visual patterns and complicate the analysis process. If the model performance does not correlate with observed data patterns, clinicians and engineers may struggle to accurately predict the prevalence of conditions like Chronic Kidney Disease (CKD) and other health metrics. This emphasizes the importance of regularly updating and refining the models to account for new features and evolving data structures within the NoSQL database. As data is continuously updated, careful attention to feature selection and data management is crucial to ensure meaningful insights and maintain the accuracy of real-time prediction



.

**Figure-1**

The correlation matrix illustrates the relationships among features related to Chronic Kidney Disease (CKD), helping us understand how they interact and influence the performance of machine learning models. Each feature has an estimated execution time that affects overall performance. In a NoSQL dataset, performance values can vary based on data size and complexity. Analyzing execution time and related graphs is essential for accurate performance estimation, as prolonged execution times may signal issues like missing values or errors, which can impact model accuracy. Thus, monitoring these times is vital for identifying potential problems.

**Motivation:**

To check the performance of time taken to estimate the detection of Chronic Kidney Disease. There are total five clinical queries used to check the performance. We have used Indexed DB, Mongo DB, Neo4j and Redis. For indexed DB Patient Data Retrieval Query works perfect because the performance time works perfectly for retreiving the patients data. For Mongo DB the query of find high risk CKD patients will work well because mongo db performance is fast when compared to indexed DB. Creating Patient Nodes query will be perfect for Neo4j as it is effective in storing and query information of patients based on their health conditions and risk factors.