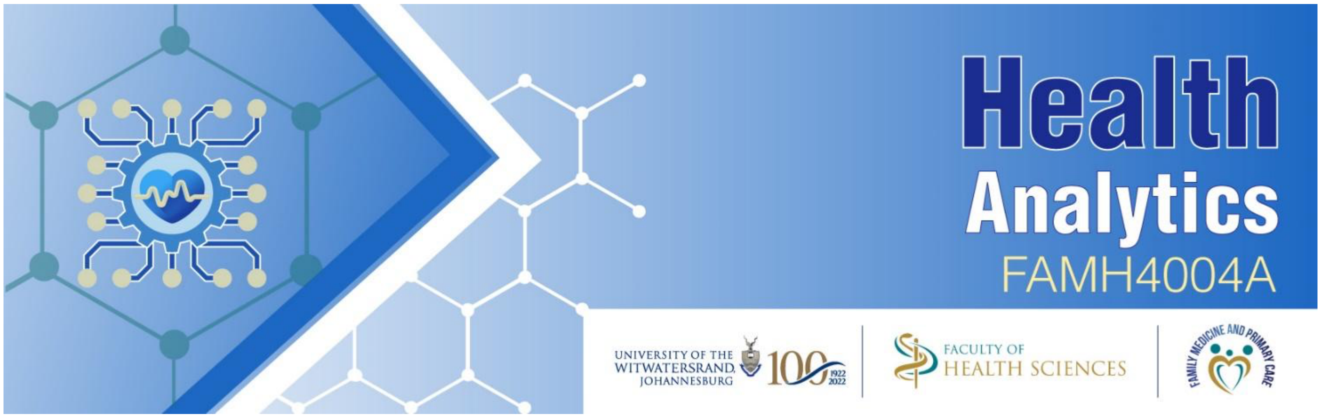
Health
Analytics
FAMH4004A
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
100 1922 2022
FACULTY OF HEALTH SCIENCES
FAMILY MEDICINE AND PRIMARY CARE

MAPPING OUT THE ECOSYSTEM

FAMH4004A
[DATE: 01 - 09 - 2023]

Please complete the following table. The assignment mark will be adjusted according to the contribution percentage shared here.

Name	Surname	Student Number	Contribution (Up to 25% - 4 group members)
Chuene	**Komane**	**1632562**	
Remofilwe	**Semaushu**	**1837926**	
Christopher Molemo	**Theys**	**1834424**	

Comprehensive Analysis of Diabetes Patient Data

Preprocessing, Visualizations, and Model Evaluation for Exploring Factors Affecting Readmission and Patient Outcomes

The current study delves into a comprehensive dataset encompassing a decade's worth of data on diabetic patients across 130 US healthcare facilities, scrutinizing factors pertaining to readmissions and other outcomes. Over 50 features, including demographic details, diagnostic reports, and medication data have been inspected meticulously to distill insights and craft predictive models.

Methodology

Data Preprocessing

The initial stage of the investigation necessitated the importation of essential libraries, notably Pandas and NumPy, to facilitate data handling and mathematical computations, respectively. Subsequently, the dataset underwent rigorous cleaning procedures to uphold data privacy principles, a process that saw the removal of individual identifiers and the elimination of columns suffering from substantial missing entries, surpassing a 10% threshold. Furthermore, columns demonstrating singular unique values were discarded to avoid skewing the predictive model.

Feature Engineering

A concerted effort was undertaken to re-engineer several features to enhance their predictive power. This involved the aggregation of various categories into more broad features, such as “readmission” and “number_of_visits”, aiming to encapsulate richer information on patients' readmission statuses and visit frequencies.

Results

Visualizations

To convey the nuanced patterns within the data clearly, a series of visualizations were conceived. These encompassed examinations into readmission rates based on gender and prior medication status, highlighting a higher propensity for readmissions among females and those with a history of diabetes medication. Moreover, the race-distribution visualization revealed that a substantial majority of the patients were Caucasian, albeit the readmission rates did not vary significantly across different racial groups.

Predictive Model

The developed predictive model exhibited a modest accuracy of 62.21%, with a precision rate of 64%. Despite leveraging the SelectKBest method to optimize the model, it did not yield the anticipated improvements, suggesting potential shortcomings in the pre-processing phase. The results propose that further advancements could be attained through refined pre-processing strategies, including a more astute feature selection grounded in domain knowledge.

Discussion

The analysis portrayed a rich tapestry of the intricacies involving diabetic patients' readmissions. The substantial representation of the Caucasian demographic in the dataset foregrounded a notable trend in doctor visits, which warrants a more detailed analysis in subsequent research to unravel the underlying causes.

The model's performance, while acceptable, identifies a clear avenue for enhancements, particularly through the incorporation of expert domain knowledge in the preprocessing and feature engineering stages, possibly leading to a more robust predictive tool in the future.

Conclusion

This research embarked on a profound analysis of a decade-long dataset aiming to glean insights into the dynamics of readmission among diabetic patients. While the predictive model constituted in the study offers a reasonable starting point, there exists a substantial scope for refinement, with a pivotal role envisaged for domain knowledge in elevating the model's performance to a higher echelon. Future directives should gravitate towards the rigorous exploration of preprocessing techniques, including thoughtful imputation strategies and a more nuanced understanding of the feature space, to construct a model with enhanced predictive accuracy and reliability.