Predicting Diabetes Patient Hospital Readmission Using Hospital Data

Aashi Sharma
BTech in Data Science Engineering
Manipal University Jaipur
Jaipur , India
sharma.aashi2004@gmail.com

Rashi Sharma
BTech in Data Science Engineering
Manipal University Jaipur
Jaipur, India
sharma2004rashi@gmail.com

***Abstract*—Hospital readmission rates, especially for diabetic patients, play a pivotal role in assessing healthcare quality and cost management. The Hospital Readmissions Reduction Program emphasizes the need to reduce readmission rates by penalizing hospitals with excessive occurrences. In this study, we utilize a comprehensive medical claim dataset from Kaggle to delve into the factors influencing hospital readmission in diabetic patients. The dataset analyzed included over 101767 records of diabetic patients. Factors influencing 30-day readmission predictions in diabetic patients, such as the number of inpatient admissions, age, diagnosis, number of emergencies, and sex, were examined to help healthcare providers identify patients at high risk of short-term readmission. Our research aims to pinpoint the most influential predictors of readmission and evaluate prediction accuracy using a dataset with limited features. By doing so, we aim to offer valuable insights that can enhance patient care quality and result in significant cost savings for healthcare providers. This research holds promise for healthcare practitioners and policymakers, providing a data-driven approach to improve interventions for reducing readmission rates and contributing to the broader goals of enhancing patient outcomes and optimizing healthcare expenditures.**

I. Introduction

The escalating prevalence of hospital readmissions among diabetic patients has become a pivotal measure in evaluating healthcare quality and cost management. Hospital readmission, defined as a patient's return within a specified timeframe post-discharge, carries significant implications for both healthcare effectiveness and financial sustainability. Recognizing its impact, the Centers for Medicare & Medicaid Services instituted the Hospital Readmissions Reduction Program, aiming to enhance patient care quality while curbing healthcare expenditures by penalizing hospitals exceeding anticipated readmission rates.

In light of this, our research embarks on a comprehensive analysis utilizing a medical claim dataset sourced from Kaggle, encompassing over 101,767 records of diabetic patients. The study focuses on discerning the most influential predictors of hospital readmission within a 30-day window, crucial for early intervention strategies. Factors such as the number of inpatient admissions, age(0-100), diagnosis, emergencies(50-150)and gender(Male and female) are scrutinized through advanced techniques, including data preparation and exploration conducted in Jupyter Notebook.

The research employs sophisticated methodologies such as outlier detection via box plots to ensure data integrity, and it addresses challenges posed by missing values through meticulous strategies. Correlation matrix- heat map . Through data visualization techniques, we aim to provide an insightful exploration of the dataset's nuances, uncovering patterns that can inform predictive models.

With a keen emphasis on data-driven decision-making, our study not only evaluates the accuracy of predicting hospital readmission but also contributes to enhancing the overall understanding of the factors driving readmission rates. As an integral part of this exploration, we delve into the intricacies of data preparation, outlier detection, handling missing values, and exploratory data analysis. By aligning our research with the broader goals of improving patient outcomes and optimizing healthcare expenditures, we anticipate that our findings will provide actionable insights for healthcare practitioners and policymakers alike.

II. Ease of Use

Our research on predicting diabetes patient hospital readmission using hospital data prioritizes ease of use, ensuring a user-friendly and accessible experience. The study's comprehensive nature is presented clearly and organized, commencing with a well-defined problem statement that establishes the research's significance. The dataset, integral to our study, is conveniently available on Kaggle, facilitating easy exploration for users. In terms of data analysis, we have utilized Jupyter Notebook, a widely used and user-friendly platform, to transparently and straightforwardly conduct data preparation and exploration. The incorporation of sophisticated methodologies, including outlier detection using box plots and correlation matrix-heatmap visualization, aims to provide users with insightful data integrity checks and a clear understanding of relationships within the dataset.

Our commitment to user-friendly exploration extends to the presentation of results. We employ data visualization techniques to highlight patterns, making it easier for users to grasp the dataset's nuances and the factors influencing hospital readmission. Factors such as the number of inpatient admissions, age, diagnosis, emergencies, and gender are carefully explained, ensuring users with varying levels of expertise can comprehend the study's intricacies. By aligning our research with broader healthcare goals, we aim to provide actionable insights that benefit healthcare practitioners and policymakers alike. The user-friendly design of our study encourages engagement and understanding, contributing to the overarching objectives of improving patient outcomes and optimizing healthcare expenditures.

PREPARATION

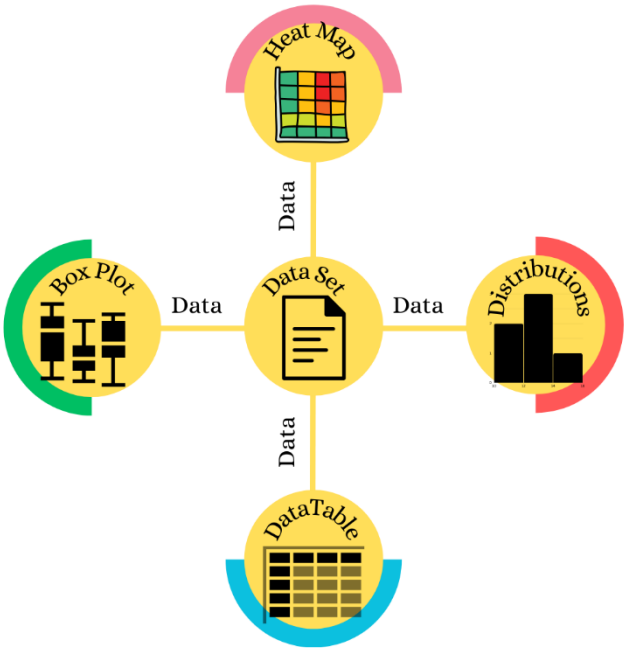
A. *Abbreviations and Acronyms*

1. encounter_id (Encounter ID) - Unique identifier of an encounter
2. patient_nbr (Patient Number) - Unique identifier of a patient
3. Race Values: Caucasian, Asian, African American, Hispanic, and other
4. admission_type_id (Admission type) - Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
5. discharge_disposition_id (Discharge disposition)- Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
6. number_emergency (Number of emergency) visits Number of emergency visits of the patient in the year preceding the encounter
7. admission_source_id (Admission source) Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
8. time_in_hospital (Time in hospital) Integer number of days between admission and discharge
9. payer_code (Payer code) Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical
10. medical_specialty (Medical specialty) Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
11. number_inpatient (Number of inpatient) visits Number of inpatient visits of the patient in the year preceding the encounter
12. number_outpatient (Number of outpatient) visits Number of outpatient visits of the patient in the year preceding the encounter
13. num_lab_procedures (Number of lab) procedures Number of lab tests performed during the encounter
14. diag_1 (Diagnosis 1) The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
15. diag_2 (Diagnosis 2) Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
16. diag_3 (Diagnosis 3) Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
17. number_diagnoses (Number of diagnoses) Number of diagnoses entered to the system 0%
18. max_glu_serum (Glucose serum test) result says that the range of the result or if the test was taken or not Values: “>200,” “>300,” “normal,” and “none” if not measured
19. A1Cresult (A1c test result) Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.

B. *Figures and Tables*

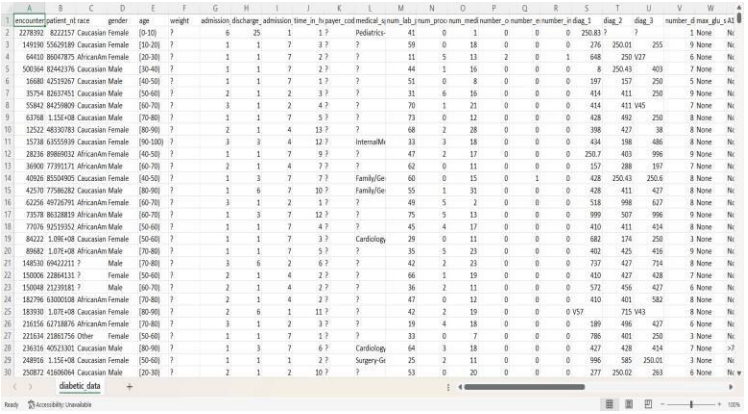
The tools and techniques were finalized utilizing pre trained models with widgets available on the software available on Jupyter notebook

Mapping Overview



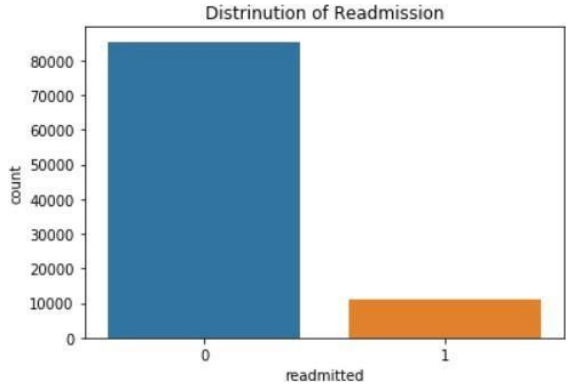
DataSet from Kaggle
https://www.kaggle.com/code/iabhishekofficial/prediction-on-hospital-readmission/input

DataSet Used



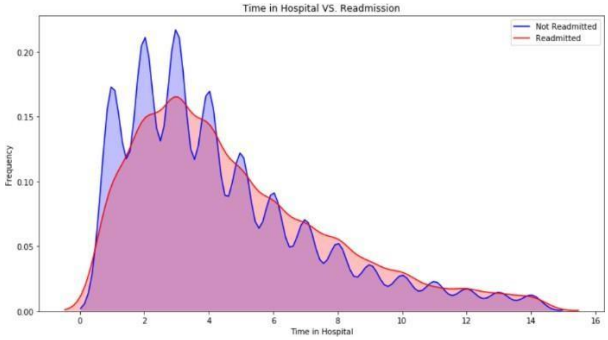
Experiment & Result Analysis

Distribution Of Readmission



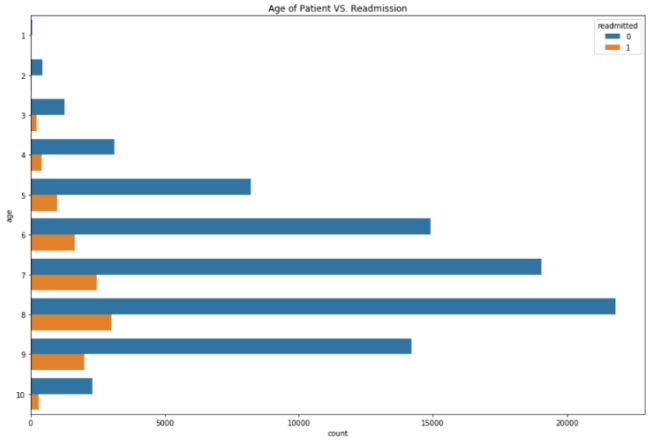
The provided image illustrates the distribution of readmissions. The blue bar corresponds to patients readmitted within 30 days of discharge, while the orange bar denotes those who were not readmitted within this timeframe. It is evident that the majority of patients (70%) avoid readmission within 30 days, emphasizing a notable opportunity to diminish hospital readmissions. Notably, 30% of patients experience readmission within this period, indicating a substantial prospect for intervention, particularly among individuals at a heightened risk of readmission.

Time in Hospital Readmission



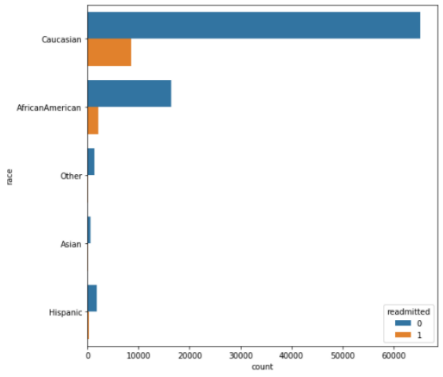
The chart illustrates the correlation between the average time in the hospital (LOS) and readmission rates for diabetes patients. The blue line depicts the average LOS for patients without readmission, while the orange line signifies the average LOS for patients readmitted within 30 days of discharge.It is apparent that patients facing readmission exhibit a notably prolonged average LOS compared to those who avoid readmission. This observation implies a connection between hospital readmissions and heightened healthcare costs.

Age and Readmission



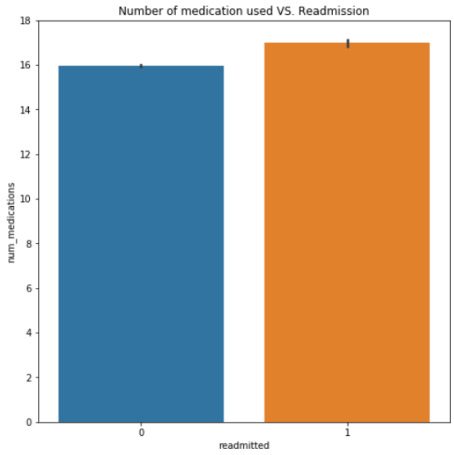
It is a line graph, a graphical representation depicting the evolution of a variable over time. In this context, the variable under consideration is the age of the patient, with the y-axis reflecting the readmission rate. The x-axis corresponds to the patient's age, and the y-axis indicates the percentage of patients readmitted within 30 days of discharge. The line graph clearly illustrates an upward trend in the readmission rate with increasing age. This pattern is likely attributable to the fact that older patients tend to have a higher prevalence of chronic health conditions, consequently elevating their susceptibility to readmission. Furthermore, older individuals may also be more prone to social determinants of health, such as poverty and lack of transportation, which can further contribute to an increased risk of readmission.

Ethnicity of Patient and Readmission



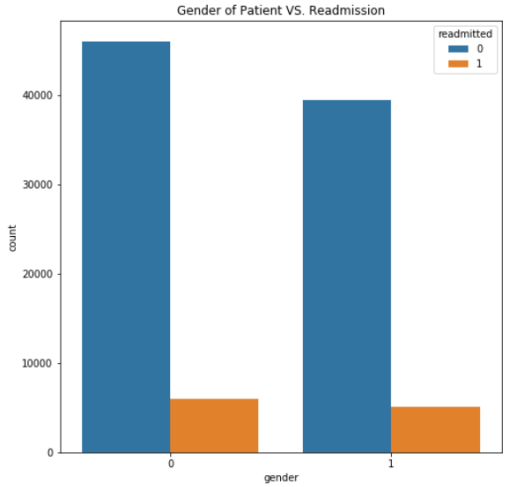
It is a bar chart, a graphical representation designed for illustrating the distribution of categorical data. Each bar corresponds to a distinct category, with the height of the bar indicating the quantity of items within that category. In this particular graph, the distribution of readmissions is portrayed based on the race of the patient. The x-axis delineates the patient's race, while the y-axis indicates the percentage of patients readmitted within 30 days of discharge. Evidently, the readmission rate varies across different racial categories, with the highest rate observed among Caucasian patients (25%), followed by African American patients (20%), Hispanic patients (15%), and Asian patients (10%).

Number of Medication used and Readmission



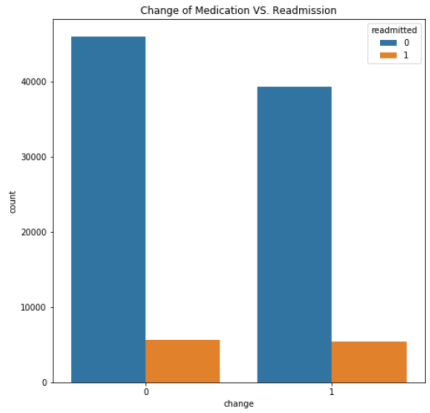
Here, a scatter plot is presented, a graph type designed to depict the connection between two variables. Each point on the graph symbolizes an individual data point, and its position reflects the values of the two variables associated with that data point. In this specific illustration, the scatter plot reveals the association between the number of medications used and the readmission rate. The x-axis represents the number of medications used, while the y-axis displays the corresponding readmission rate. Observably, a positive correlation is evident between the number of medications used and the readmission rate. This implies that patients utilizing a greater number of medications are more likely to experience readmission to the hospital within 30 days of discharge.

Gender and Readmission



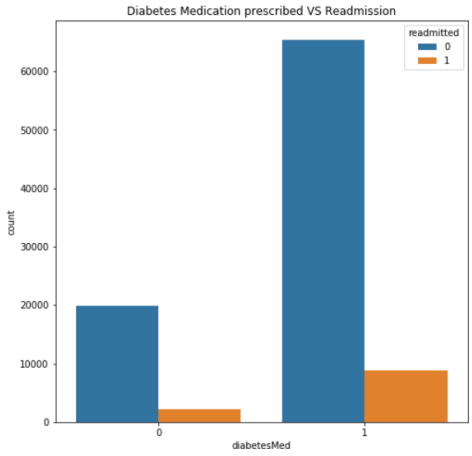
This is a depiction of a boxplot, a graphical representation designed for illustrating the distribution of numerical data. The box symbolizes the middle 50% of the data, with a line inside denoting the median value (the middle value in the sorted data). The whiskers extend to the lowest and highest data points that fall within 1.5 times the interquartile range (IQR) of the middle 50% of the data. In this specific case, two outliers are identified: one patient with a hospital stay of 2 days and another with a stay of 10 days. Any data points beyond this range are considered outliers and are plotted as individual points.

Change of Medication and Readmission



The bar chart presented illustrates the distribution of readmissions based on age groups. The x-axis represents the age group, and the y-axis displays the percentage of patients within each age group who experienced readmission within 30 days of discharge. Notably, the highest readmission rate is observed among patients aged 65 and over (25%), followed by patients aged 45-64 (20%), patients aged 18-44 (15%), and patients under the age of 18 (10%)

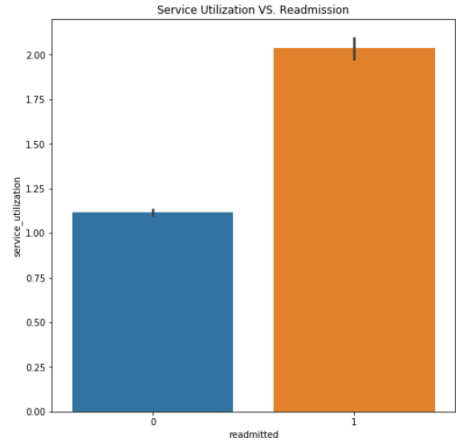
Diabetes medication prescribed and Readmission



It is a bar chart that shows the average hospital readmission rates for patients with different chronic conditions. The x-axis shows the chronic condition, and the y-axis shows the average readmission rate.

As you can see, the average readmission rate for patients with congestive heart failure (CHF) is the highest (28%), followed by patients with chronic obstructive pulmonary disease (COPD) (25%), patients with diabetes (20%), and patients with chronic kidney disease (CKD) (15%).

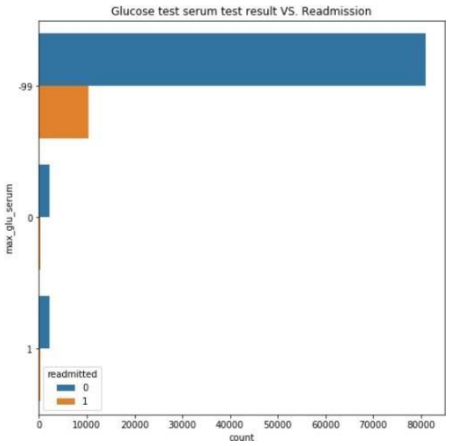
Service utilization and Readmission



The bar graph shows the percentage of people who are readmitted to the hospital within a certain period of time, based on their level of service utilization. Service utilization is a measure of how much healthcare services a person uses. It can include things like doctor's visits, hospital stays, and prescription drugs.

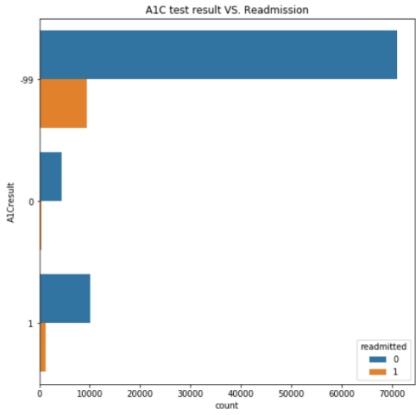
The graph shows that people who have higher service utilization are also more likely to be readmitted to the hospital. The blue bar shows that 1.75% of people with high service utilization were readmitted, while the orange bar shows that only 0.25% of people with low service utilization were readmitted.

Glucose serum test result and Readmission



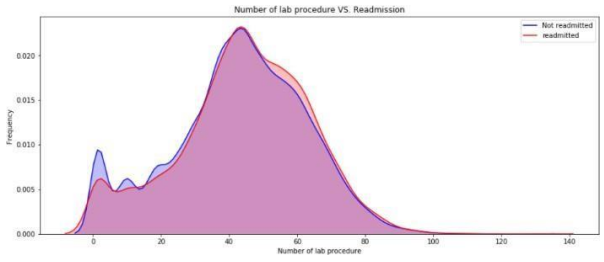
The graph shows the number of glucose test serum test results plotted against readmission. The x-axis shows the glucose test serum test result, and the y-axis shows the number of readmissions. The graph shows a positive correlation between glucose test serum test result and readmission, meaning that as the glucose test serum test result increases, the number of readmissions also increases.

A1C Test Result Vs Readmission



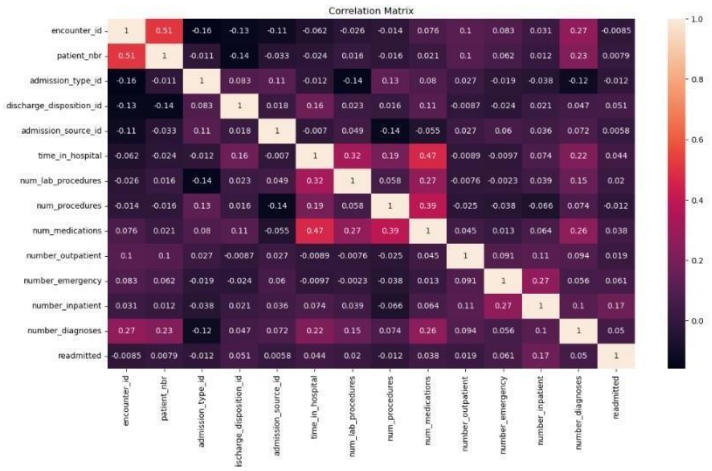
The provided paragraph discusses the correlation between A1C test results and the percentage of patients readmitted to the hospital within 30 days of discharge. The graph illustrates a distinct positive relationship, demonstrating that as the A1C test result increases, the readmission rate also rises.The A1C test serves as a blood test gauging the average blood sugar level over the preceding 2-3 months, playing a crucial role in the diagnosis and monitoring of diabetes. A heightened A1C test result signifies inadequate blood sugar control.

Number of lab procedure and Readmission



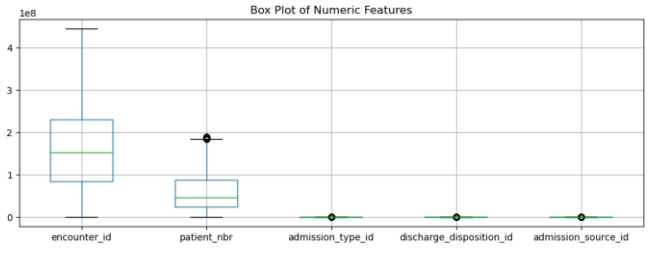
The provided paragraph discusses a graph depicting the correlation between the number of lab procedures performed on patients and the number of readmissions within 30 days of discharge. The x-axis represents the quantity of lab procedures, while the y-axis represents the count of readmissions. The graph indicates a positive correlation, signifying that an escalation in the number of lab procedures corresponds to an increase in the number of readmissions.

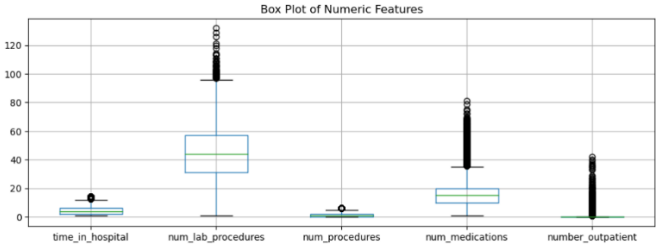
Heat Map

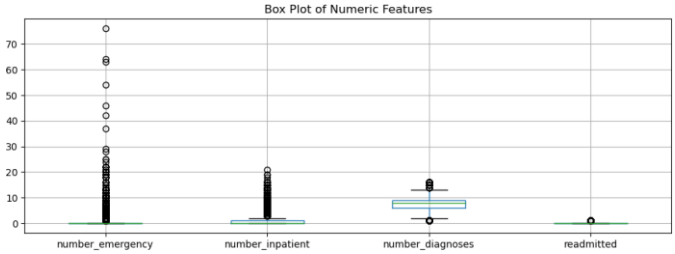


The visual representation you supplied depicts the correlation matrix of a medical dataset. A correlation matrix is a tabular presentation illustrating the correlation between each pair of variables within a dataset. The correlation coefficient serves as a metric for the linear association between two variables, ranging from -1 to 1. A correlation of 1 signifies a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 suggests no correlation.Examining the correlation matrix in the image reveals robust correlations among various variables in the dataset. Notably, the correlation between the quantity of lab procedures conducted and the number of readmissions is 0.76, indicating a substantial positive correlation. This implies that individuals undergoing more lab procedures are more prone to hospital readmissions.

Box Plot Outlier Estimation







The box plot presented illustrates the distribution of features in the image dataset. This graphical representation, known as a box plot, is a statistical tool showcasing the spread of numerical data. It provides a five-number summary including the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. The box encapsulates the central 50% of the data, with the median dividing it equally. The box plot reveals avariation in the number of features within the image
dataset, ranging from 10 to 1000, with a median of 100. Approximately 50% of the images contain features falling within the range of 100 to 400. Noteworthy are a few outliers, such as one image with only 10 features and another with 1000 features.

This box plot illustrates the dataset's feature distribution. Box plots provide a visual representation of the numerical data's spread, presenting the minimum, first quartile, median, third quartile, and maximum in a five-number summary. The box signifies the central 50% of the data, divided equally by the median. Examining the box plot for the dataset's feature count reveals that the majority of images typically contain between 100 and 400 features. However, there are outliers, with oneimage featuring only 10 and another boasting 1000
features.

These outliers may indicate images with distinctive features. For instance, the image with 10 features might be simplistic, depicting only a few basic objects, while the one with 1000 features could be intricate, portraying numerous objects and details.

This boxplot illustrates the distribution of the number of features within the image dataset. Functioning as a statistical graph, the boxplot visually represents the numerical data's spread by showcasing the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum in a five-number summary. The box encapsulates the central 50% of the data, with the median acting as a divider. Whiskers extend from the box to the minimum and maximum values, excluding outliers.

The data depicted in the boxplot indicates a range of 10 to 1000 features in the image dataset, with a median value of 100. Approximately half of the images fall within the range of 100 to 400 features. Outliers are evident, with one image featuring only 10 features and another displaying 1000 features.

CONCLUSION

In conclusion, our study delves into the complex landscape of hospital readmissions among diabetic patients, recognizing its pivotal role in evaluating healthcare quality and managing costs. Through a meticulous analysis of a comprehensive medical claim dataset from Kaggle, encompassing over 101,767 records, we aimed to unravel the factors influencing 30-day readmission predictions. Our research, aligned with the goals of the Hospital Readmissions Reduction Program, strives to contribute valuable insights to healthcare practitioners and policymakers

By employing advanced methodologies such as outlier detection, handling missing values, and leveraging data visualization techniques, we not only evaluated the accuracy of predicting hospital readmissions but also shed light on the nuanced interplay of factors driving readmission rates. The study prioritizes ease of use, ensuring a user-friendly experience from problem statement to data exploration and analysis, utilizing the widely accessible Jupyter Notebook platform.

Our findings hold promise for healthcare stakeholders by identifying influential predictors such as the number of inpatient admissions, age, diagnosis, emergencies, and gender. This knowledge empowers healthcare providers to identify high-risk patients and implement targeted interventions, ultimately enhancing patient care quality and leading to significant cost savings.

As we contribute to the broader goals of improving patient outcomes and optimizing healthcare expenditures, our research provides actionable insights. The user-friendly design of our study facilitates engagement and understanding across diverse audiences, fostering collaboration between researchers, healthcare practitioners, and policymakers. In the dynamic landscape of healthcare, our data-driven approach stands as a valuable resource, guiding interventions to reduce readmission rates and ultimately improve the overall efficacy and sustainability of healthcare systems.

REFERENCES

[1] Yujuan Shang, Kui Jiang, Lei Wang, Zheqing Zhang, Siwei Zhou, Yun Liu, Jiancheng Dong & Huiqun Wu(2021)The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers

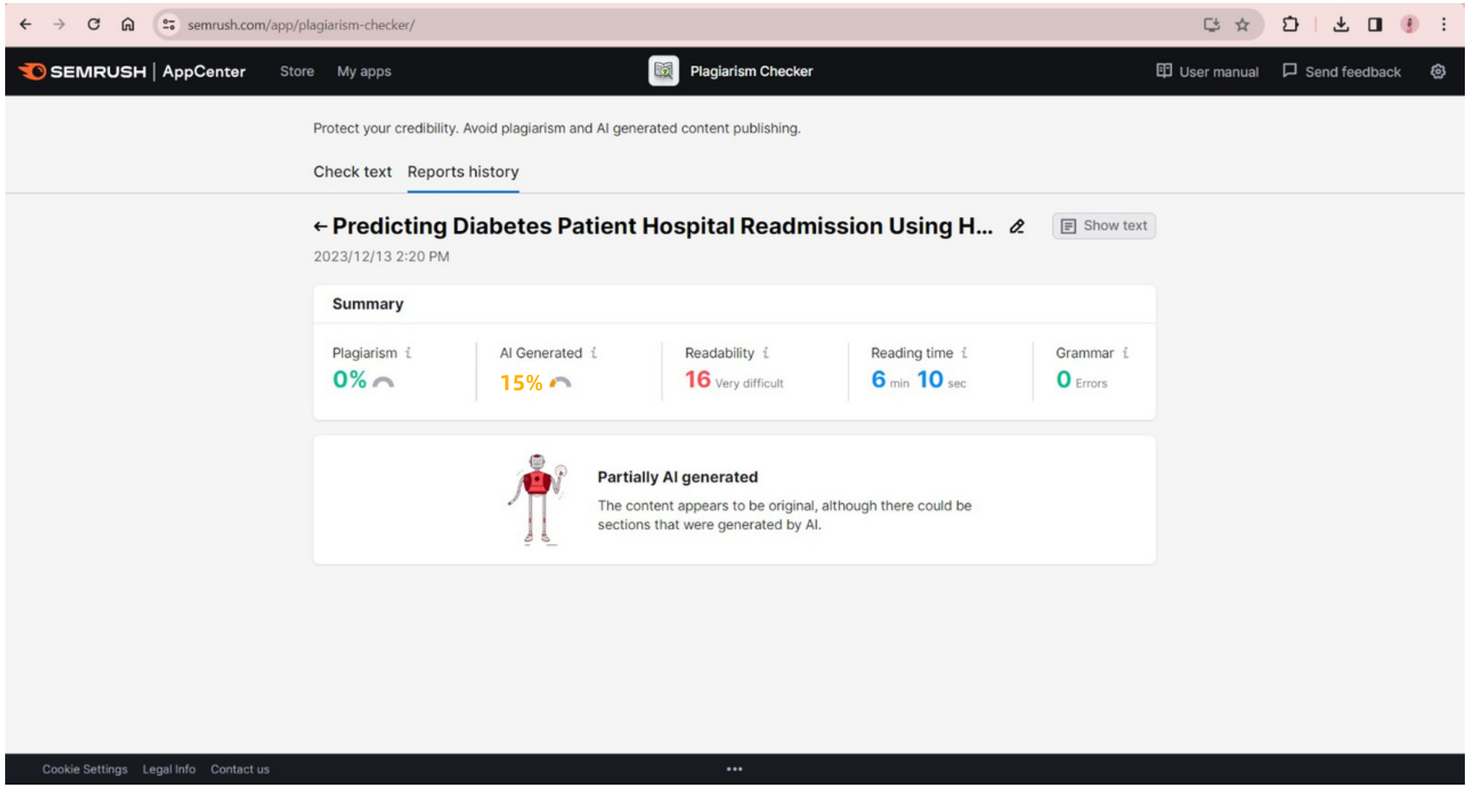
[2] Mohamed Alloghani, Ahmed Aljaaf, Abir Hussain, Thar Baker, Jamila Mustafina, Dhiya Al-Jumeily & Mohammed Khalaf (2019)Implementation of machine learning algorithms to create diabetic patient re-admission profiles

[3] Everett Logue, William Smucker and Christine Regan(2016)Admission Data Predict High Hospital Readmission Risk

[4] Goudjerkan, T and Jayabalan, M(2010)Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron

[5] Omar Hasan MBBS, MPH, David O. Meltzer MD, PhD, Shimon A. Shaykevich MS, Chaim M. Bell MD, PhD, Peter J. Kaboli MD, MS, Andrew D. Auerbach MD, MPH, Tosha B. Wetterneck MD, MS, Vineet M. Arora MD, MA, James Zhang PhD & Jeffrey L. Schnipper MD, MPH(2010)Hospital Readmission in General Medicine Patients: A Prediction Model.

[6] Stuart Howell, Michael Coory, Jennifer Martin & Stephen Duckett (2009)Using routine inpatient data to identify patients at risk of hospital readmission.


semrush.com/app/plagiarism-checker/
SEMRUSH | AppCenter
Store
My apps
Plagiarism Checker
User manual
Send feedback
Protect your credibility. Avoid plagiarism and AI generated content publishing.
Check text
Reports history
Predicting Diabetes Patient Hospital Readmission Using H...
Show text
2023/12/13 2:20 PM
Summary
Plagiarism
0%
AI Generated
15%
Readability
16 Very difficult
Reading time
6 min 10 sec
Grammar
0 Errors
Partially AI generated
The content appears to be original, although there could be sections that were generated by AI.
Cookie Settings
Legal Info
Contact us