CSE558 Data Science Midsem Project

The Predictive Predators

Manan Aggarwal (2022273)
Manit Kaushik (2022277)
Mohd. Masood (2022299)
Shobhit Raj (2022482)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Dataset

Dataset (Link): Contains data points for 10 years of clinical care for 130 US hospitals relating to diabetes patients.

Size: (101766 x 48)

Some Features are:

Demographics	race, gender, age, weight
Clinical Info	Time_in_hospital (days), num_procedures, num_medications, A1Cresult (Result of HbA1c test which measures the blood glucose level), etc.
Diagnosis	diag_1, diag_2, diag_3 (diagnosis coded as first three digits of ICD9)
Medication Change	Metformin, repaglinide, nateglinide, etc. (up, down, steady, no)
Label	Readmitted (If the patient was readmitted (<30, >30, No))

Problem Statement

Problem Statement:
What are the key factors contributing to readmission in diabetic patients?

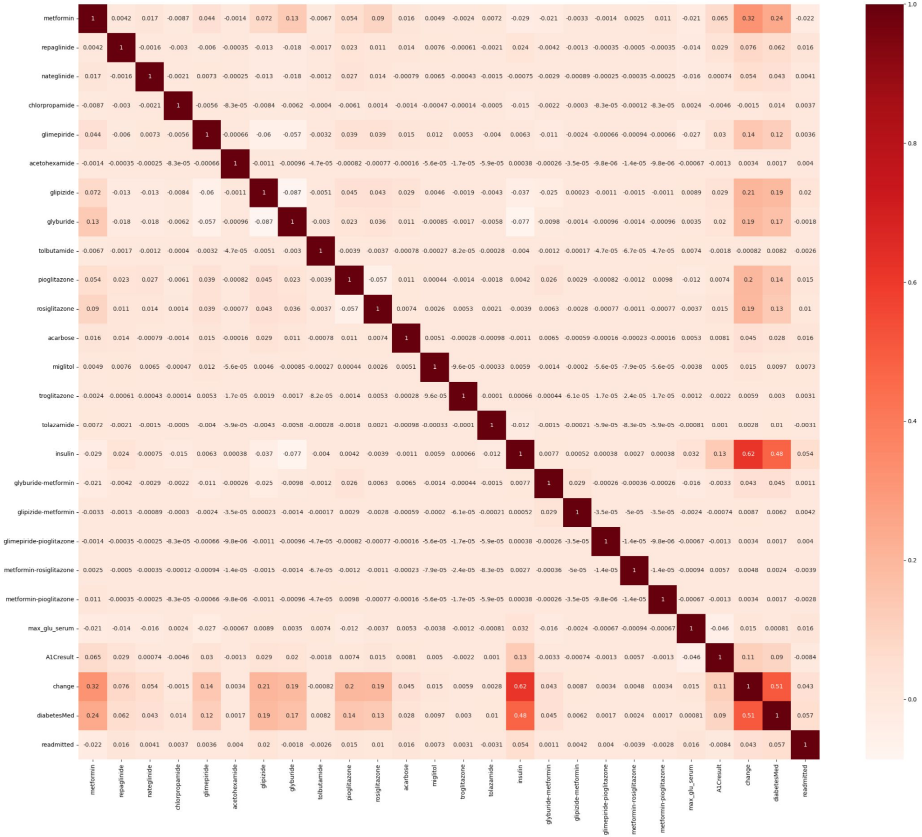
Objective:

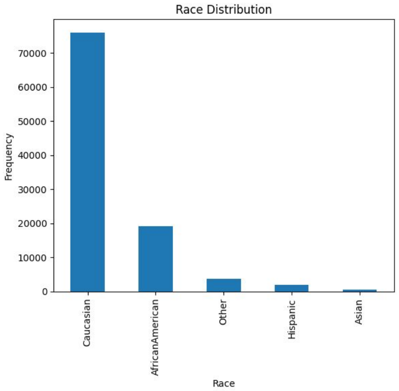
- Use **data analysis and feature importance methods** to identify factors (e.g., HbA1c levels, medication, length of stay, demographics) linked with readmission.
- Provide actionable insights for **customizing patient care** to minimize readmission risks.

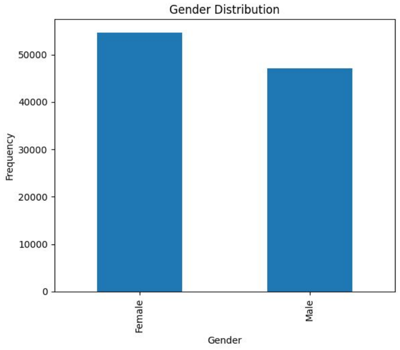
Importance:

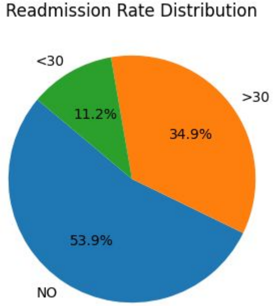
- **Improves Patient Outcomes**: Reduces risk of complications by focusing on critical readmission factors.
- **Optimizes Resource Use**: Helps hospitals allocate resources effectively, targeting high-risk areas.
- **Enables Targeted Interventions**: Insights guide adjustments in discharge planning, medication management, and follow-up care.

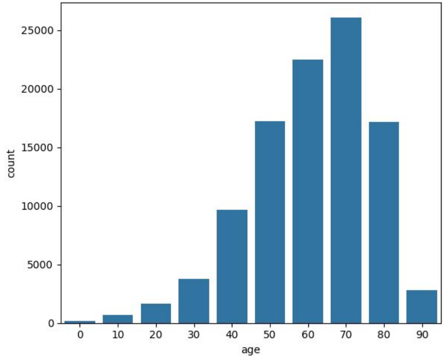
Exploratory Data Analysis











Exploratory Data Analysis

Insulin & Readmission: Moderate positive correlation with readmission, suggesting higher readmission likelihood for insulin users.

Insulin & Medication Change: Positive correlation with "change" and "diabetesMed," indicating frequent adjustments in insulin patients.

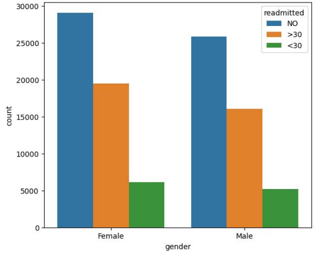
Weak Correlations: Factors like insulin usage contribute to readmission risk but are not standalone predictors.

Age Distribution: Higher concentration in 50-80 age group, especially 70-80, with higher readmission rates, particularly >30 days.

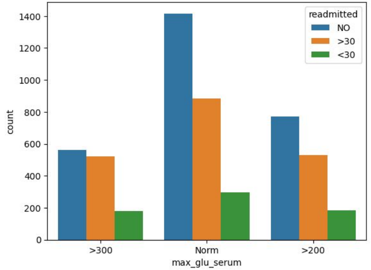
Gender Distribution: Slightly more females; similar readmission patterns for both genders with most patients not readmitted.

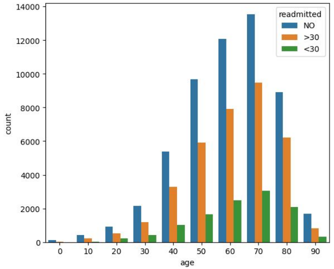
Race Distribution: Majority Caucasian, followed by African American; similar readmission trends across races with low <30-day readmissions.

Age Distribution by readmission: The distribution of all the readmitted classes seem to follow similar pattern.

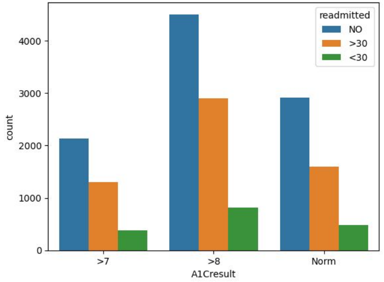


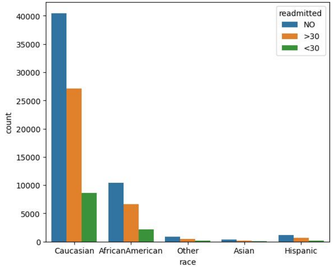
Max Glucose Serum vs. Readmission





Max Glucose Serum vs. Readmission





Preprocessing

To prepare the dataset for analysis and model building, we applied the following tasks to handle missing values :-

1. **race**: Missing values were categorized as "others" to retain data diversity without creating imbalance.
2. **weight**: Dropped due to ~97% missing values, as it would contribute minimal usable information.
3. **payer_code**: Dropped because it had limited relevance to readmission prediction (~40% missing)
4. **medical_specialty**: Filled missing values as "missing" to maintain the sample count without loss of potentially useful information. (~49% missing) [1]
5. **diag_1**: Primary diagnosis (21 rows) with missing values dropped, as these represented a very small portion of the data.
6. **diag_2** & **diag_3**: Filled missing diagnosis codes with values from the previous diagnosis column to retain continuity in patient health history.
7. **readmitted_binary:** New feature made to identify if a patient was ever readmitted.

Preprocessing

Label Encoding: Applied manual encoding to key features with ordinal relationships, like medication status (Up = 3, Down = 2, Steady = 1, No = 0) and test results for max_glu_serum & A1Cresult. For features like change and diabetesMed, we encoded binary values (Yes/No) to 1 and 0, and we encoded readmitted values as: >30 = 2, <30 = 1, and NO = 0 to reflect increasing levels of readmission concern.

Effects of Preprocessing:

1. Improved data consistency by addressing missing values systematically.
2. Retained a higher number of records by filling or encoding instead of excessive data removal.
3. Reduced dimensionality by removing irrelevant or highly sparse features (e.g., weight and payer_code).
4. Manual encoding preserved interpretive continuity, allowing the model to understand and differentiate levels of medication change or test results.

Final Dataset Size: (101742 x 45)

Hypothesis Tests

1. **t-test** to compare mean of time_in_hospital of those readmitted and those who were not readmitted. Null Hypothesis: The mean time spent in the hospital is equal for patients who were readmitted and those who were not readmitted.
2. **Chi-Square test of independence** to determine if age distribution varies across different readmission categories.

   Null Hypothesis: Age distribution is independent of readmission status, meaning the age distribution is similar across different readmission categories.
3. **Chi-Square goodness of independence** to check if A1C Result and Readmission are related. Null Hypothesis: A1C results (whether normal, >7, >8, etc.) are independent of the readmission status.
4. **Chi-Square goodness of independence** to check if insulin medicine and Readmission are related. Null Hypothesis: The use of insulin is independent of the readmission status.

Hypothesis Tests' results

1. The **t-test** result indicates a statistically significant difference in the average time spent in the hospital between readmitted and non-readmitted patients, suggesting that readmitted patients tend to have longer hospital stays, potentially reflecting more complex health issues.
2. The **Chi-Square** test result indicates a statistically significant difference in age distribution across different readmission categories, suggesting that certain age groups may exhibit unique patterns of readmission.
3. The **Chi-Square** test result indicates a statistically significant association between A1C results and readmission status, suggesting that poor blood sugar control (elevated A1C levels) may be linked to a higher likelihood of readmission.
4. The **Chi-Square** test result indicates a statistically significant association between insulin usage and readmission status, suggesting that patients on insulin may be at a higher risk of readmission, possibly due to more severe diabetes requiring closer monitoring.

References

[1] Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. Biomed. Res. Int. 2014;2014:781670.

Thank You