CSE558 - Data Science Endsem Project
Team Name - The Predictive Predators

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

1. Abstract

The study investigates factors influencing readmission among diabetic patients. Exploratory analysis and dimensionality reduction techniques identified critical predictors of readmission. Machine learning models, including Logistic Regression and Random Forests, were implemented to forecast outcomes, offering insights to improve patient care and optimize healthcare resource allocation. Github repository for this project: Link [1]

2. Dataset

We used the data set titled [**Diabetes 130-US Hospitals for Years 1999-2008**] [2] provided by Strack et al. [3] representing ten years (1999-2008) of clinical care in 130 US hospitals and integrated delivery networks. Each row in the data set corresponds to the hospital records of diabetic patients who underwent laboratory tests, received medications, and stayed in the hospital for up to 14 days. The dataset contains **101,766 instances and 47 features**. The following table outlines the different features of the data set, grouped into specific categories.

Group	Features
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)

Table 1. Features of the Diabetes 130-US Hospitals dataset, grouped by category.

The label being the readmitted attribute representing if a patient was readmitted and if readmitted then how much time did it take for the patient to be readmitted.

3. Problem Statement

Readmissions among diabetic patients pose a significant challenge to healthcare systems, resulting in higher costs and poorer outcomes. Identifying the factors contributing to readmission is crucial for addressing this issue effectively. This project aims to **investigate the key factors associated with readmission**, focusing on medical, demographic, and behavioral factors such as HbA1c levels, medication, length of stay, and patient demographics.

By understanding these factors, we **aim to build a predictive model** that can forecast readmission, ultimately improving patient outcomes, optimizing resource allocation, and enabling targeted interventions.

Exploratory Data Analysis

The following key observations were made during the exploratory data analysis:

- **Insulin & Readmission**: Moderate positive correlation with readmission, suggesting higher likelihood among insulin users.
- **Insulin & Medication Change**: Positive correlation with "change" and "diabetesMed", indicating frequent adjustments in insulin patients.
- **Weak Correlations**: Insulin usage contributes to readmission risk but is not a standalone predictor.
- **Age Distribution**: Higher concentration in the 50-80 age group, especially 70-80, with higher readmission rates for those readmitted after 30 days.
- **Gender Distribution**: Slightly more females; readmission patterns similar across genders.
- **Race Distribution**: Majority Caucasian, followed by African American; similar readmission trends across races with lower less than 30-day readmissions.

- **Age Distribution by Readmission**: Similar distribution patterns across readmission categories for all age groups.

Below are some visualizations (graphs) that represent these findings from the exploratory data analysis:

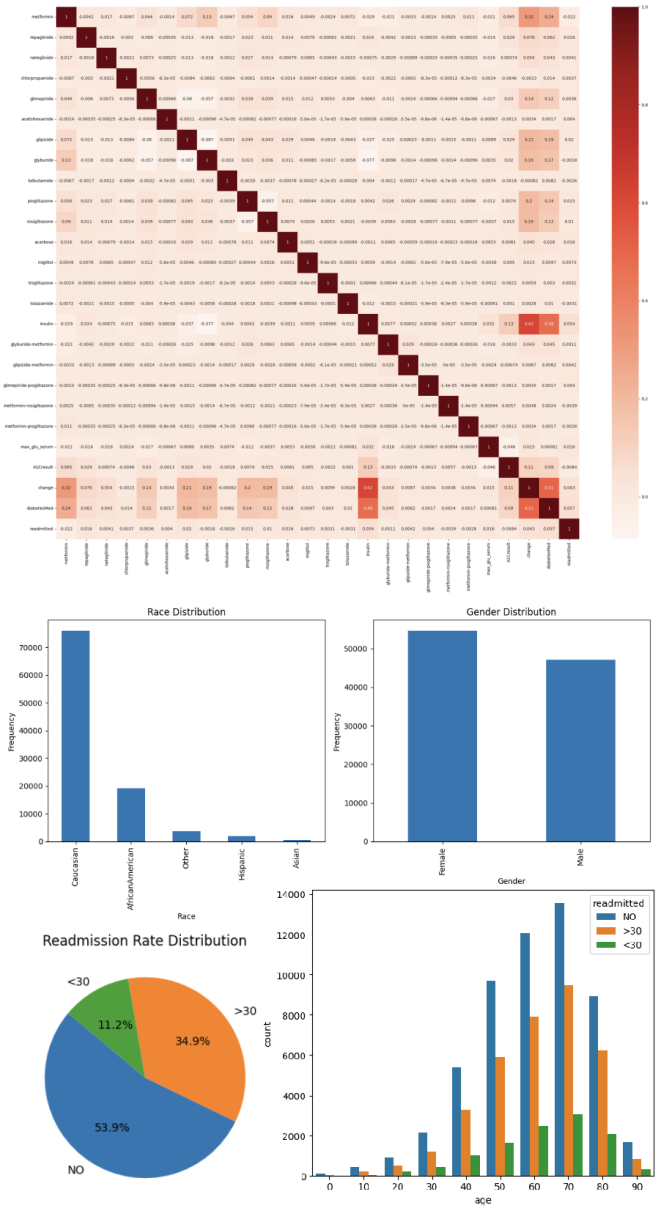

Figure 1. EDA Graphs

4. Challenges of Dataset

4.1. Missing Values

To prepare the dataset for analysis and model building, the following tasks were applied to handle missing values as also used by Strack et al. [3]:

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

4.2. Label Encoding

Manual encoding was applied to key **features with ordinal relationships**, such as `medication status` (Up = 3, Steady = 2, Down = 1, No = 0) and test results for `max_glu_serum` and `A1Cresult`. For features like `change` and `diabetesMed`, binary values (Yes/No) were encoded as 1 and 0, respectively. Additionally, `readmitted` values were encoded as follows: `>30 = 2, <30 = 1, & NO = 0`, reflecting increasing levels of readmission concern. We also made a new feature `readmitted_binary` which is 1 if the patient was readmitted and 0 otherwise. Hereafter, for the purpose of this project we only focus on this attribute as target.

4.3. Effects of Pre-processing

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

Final dataset size after pre-processing: 101742 x 45

5. Hypothesis Tests

The following hypothesis tests were conducted on 5% of the dataset to assess the statistical significance of various factors related to readmission:

- **Two tailed t-test to compare mean of `time_in_hospital`**:
  - **Reason**: Assess whether the average time spent in the hospital differs between patients who were readmitted and those who were not.
  - **Null Hypothesis**: The mean time spent in the hospital is equal for patients who were readmitted and those who were not readmitted.
- **Two tailed Chi-Square test of independence for age distribution across readmission categories**:
  - **Reason**: As observed in figure 2, the distribution of age across the readmission categories appears to follow a similar Gaussian pattern. Assess if this is true.
  - **Null Hypothesis**: Age distribution is independent of readmission status, meaning the age distribution is similar across different readmission categories.
- **Two Tailed Chi-Square goodness of independence for A1C result and readmission status**:
  - **Reason**: Examines whether A1C levels are associated with readmission status.
  - **Null Hypothesis**: A1C results (whether normal, `>7`, `>8`, etc.) are independent of the readmission status.
- **Two Tailed Chi-Square goodness of independence for insulin medicine and readmission status**:
  - **Reason**: Evaluate whether the use of insulin is associated with readmission status.
  - **Null Hypothesis**: The use of insulin is independent of the readmission status.

5.1. Conclusion

- 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.
- 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.
- The **Chi-Square test** results indicate **no statistically significant association** between A1C results and readmission status, suggesting that elevated A1C levels may not have a direct impact on the likelihood of readmission.
- 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.

6. Results of ML Models on Original Data

The tables below show the performance of Logistic Regression & Random Forest Classifier on the original dataset:

Class	Precision	Recall	F1-Score
0	0.64	0.75	0.69
1	0.63	0.52	0.57
Macro Avg	0.64	0.63	0.63
Weighted Avg	0.64	0.64	0.64
Accuracy	0.64		
Time Taken	248.2 sec		

Table 2. Random Forest

Class	Precision	Recall	F1-Score
0	0.61	0.81	0.70
1	0.65	0.40	0.49
Macro Avg	0.63	0.61	0.60
Weighted Avg	0.63	0.62	0.61
Accuracy	0.62		
Time Taken	272.5 sec		

Table 3. Logistic Regression

7. Results of ML Models on Scaled Data

7.1. Maximizing Variance

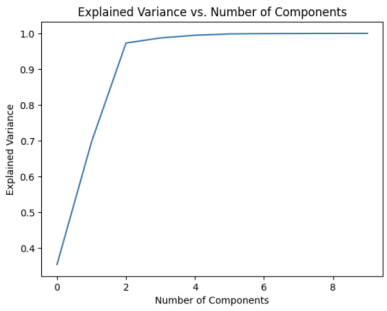

Figure 2. Explained Variance v/s Number of Components

The optimal number of components was determined by analyzing the explained variance against the number of components. It was found that `n=10` provides a good balance, capturing a significant portion of the variance in the data.

7.2. Truncated SVD

We applied this method to reduce the dimensionality of the data to 10 features. The rationale behind this approach is that it projects the data into a lower-dimensional space that best captures the underlying variance and minimizes noise.

Class	Precision	Recall	F1-Score
0	0.60	0.68	0.63
1	0.63	0.52	0.57
Macro Avg	0.57	0.57	0.57
Weighted Avg	0.58	0.58	0.57
Accuracy	0.58		
Time Taken	61.9 sec		

Table 4. Random Forests on Truncated SVD

7.3. CUR Decomposition

We applied this technique to decompose the data into rows and columns. This approach is particularly advantageous for handling sparse datasets (our dataset is 60% sparse), as it reduces computation time and, in some cases, better preserves the data structure compared to SVD. By maintaining the original row and column format, it enhances interpretability. Additionally, the number of features was reduced to 10 while keeping the number of samples constant.

Class	Precision	Recall	F1-Score
0	0.64	0.75	0.69
1	0.63	0.52	0.57
Macro Avg	0.64	0.63	0.63
Weighted Avg	0.64	0.64	0.64
Accuracy	0.64		
Time Taken	20.7 sec		

Table 5. Random Forests on CUR Decomposition

7.4. Mutual Information & JL Lemma

Firstly, we applied the **mutual_info_classif** method to assess the relationship between features and the target variable, identifying the most informative features. The top 10 selected features were: `A1Cresult`, `diag_3`, `number_diagnoses`, `number_emergency`, `num_procedures`, `diag_1`, `discharge_disposition_id`, `number_outpatient`, `diag_2`, & `number_inpatient`.

Next, we used the **Johnson-Lindenstrauss Lemma** to project the data into a lower-dimensional space while preserving distances. This technique is particularly suitable for high-dimensional data and effectively preserves mutual information.

Class	Precision	Recall	F1-Score
0	0.61	0.83	0.70
1	0.65	0.36	0.47
Macro Avg	0.63	0.60	0.58
Weighted Avg	0.63	0.62	0.59
Accuracy	0.62		
Time Taken	4.6 sec		

Table 6. Logistic Regression on JL Lemma

7.5. Analysis

- Mutual Information with JL Lemma slightly reduced accuracy, reflecting a trade-off between dimensionality reduction and data retention.
- SVD-based reduction caused a minor performance drop, emphasizing the risks of aggressive dimensionality reduction.
- CUR Decomposition maintained accuracy while efficiently handling sparse data, preserving key features without performance loss.
- Logistic regression on original and CUR-reduced data yielded identical accuracy, with CUR-reduced data enabling faster execution.

8. Conclusion

- Feature selection and dimensionality reduction techniques such as Mutual Information, JL Lemma, CUR, and SVD effectively reduce model complexity and runtime but may lead to minor accuracy losses due to information reduction.
- Mutual Information highlights the influence of patient demographics, medical history, and healthcare interactions on diabetes readmission outcomes.
- The analysis reveals racial disparities in diabetes complications, emphasizing the need for targeted and specialized care strategies to address diverse healthcare challenges.

9. Future Work

Future efforts could explore advanced feature selection techniques and domain knowledge-driven feature engineering. For instance, analyzing the effects of varying chemical dosages may uncover additional critical factors influencing readmission, enabling more precise and effective patient care strategies.

References

[1] Github link for DSc-Project https://github.com/Sparta9000/DSc-Project.

[2] UCI Machine Learning Repository — archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008. [Accessed 15-12-2024].

[3] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *Biomed Res. Int.*, 2014:781670, Apr. 2014.