# CSE558 Data Science Endsem Project

The Predictive Predators

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

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
**DELHI**

# Recap (1/5): 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)) |

# Recap (2/5): 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.
- Build a predictive model to determine whether a patient will be readmitted or not.

**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.

# Recap (3/5): Preprocessing

To prepare the data 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.

# Recap (4/5): Preprocessing

**Label Encoding**: Applied manual encoding to key features with ordinal relationships, like medication status (Up = 3, Steady = 2, Down = 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)

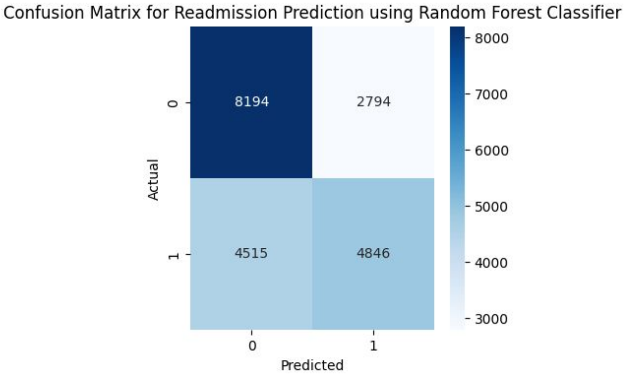
# Recap (5/5): Hypothesis Tests' results

1. The **t-test** result indicates a statistically significant difference in the average time spent in the hospital between re-admitted 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 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.
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.

# Results of ML Models on Original Data

## Random Forest Classifier:

| | Precision | Recall | F1-Score |
|---|---|---|---|
| **0** | 0.64 | 0.75 | 0.69 |
| **1** | 0.63 | 0.52 | 0.57 |
| **Accuracy** | | | 0.64 |
| **Macro Avg** | 0.64 | 0.63 | 0.63 |
| **Weighted Avg** | 0.64 | 0.64 | 0.64 |



Time Taken: 248.2 Secs

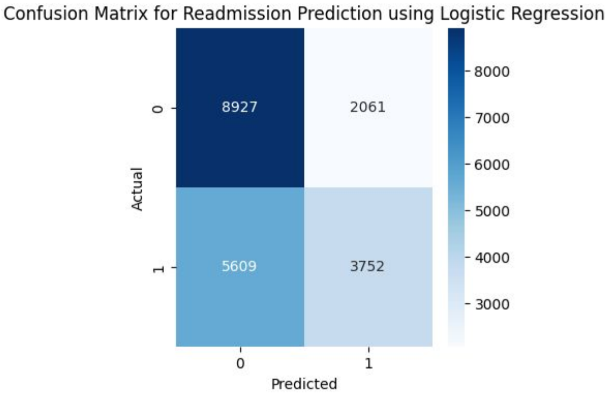
O(TNM(logN))

For NxM matrix and T stumps

# Results of ML Models on Original Data

## Logistic Regression:

| | Precision | Recall | F1-Score |
|---|---|---|---|
| **0** | 0.61 | 0.81 | 0.70 |
| **1** | 0.65 | 0.40 | 0.49 |
| **Accuracy** | | | 0.62 |
| **Macro Avg** | 0.63 | 0.61 | 0.60 |
| **Weighted Avg** | 0.63 | 0.62 | 0.61 |



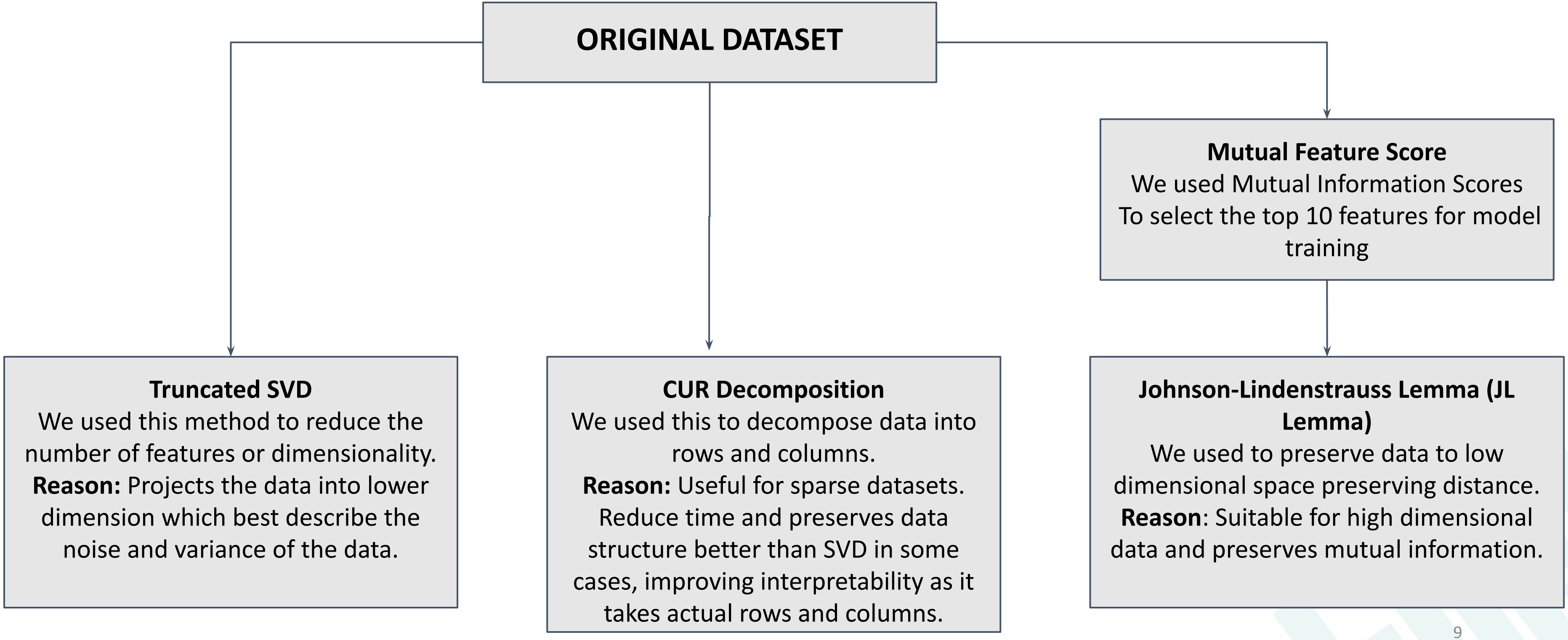
Time Taken: 272.5 Secs

O(NMI)

For NxM matrix and I Iterations

# Reducing Feature Set

We applied randomized scaling techniques to improve the training efficiency of our machine learning models.



**ORIGINAL DATASET**

**Mutual Feature Score**
We used Mutual Information Scores
To select the top 10 features for model training

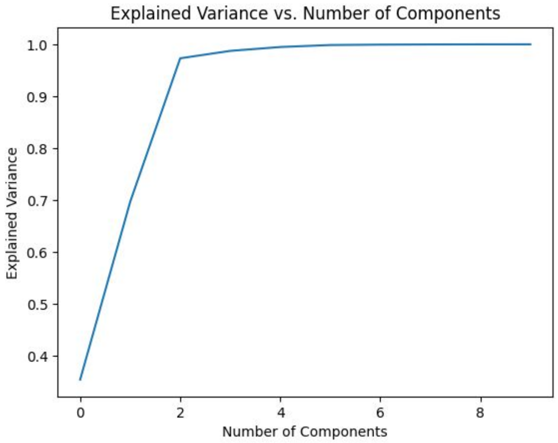
**Truncated SVD**
We used this method to reduce the number of features or dimensionality.
**Reason:** Projects the data into lower dimension which best describe the noise and variance of the data.

**CUR Decomposition**
We used this to decompose data into rows and columns.
**Reason:** Useful for sparse datasets. Reduce time and preserves data structure better than SVD in some cases, improving interpretability as it takes actual rows and columns.

**Johnson-Lindenstrauss Lemma (JL Lemma)**
We used to preserve data to low dimensional space preserving distance.
**Reason**: Suitable for high dimensional data and preserves mutual information.

# Reducing Feature Set

The number of components was obtained from explained variance vs number of components, where it was found that n=10 is a pretty good point which explains much of the variance in the data



We applied mutual_info_classif (calculated as the reduction in uncertainty of one variable given knowledge of the other) on the dataset with the given target which gives us the columns which best describe the target variable and took 10 out of these which were the following:

**A1Cresult, diag_3, number_diagnoses, number_emergency, num_procedures, diag_1, discharge_disposition_id, number_outpatient, diag_2, number_inpatient**

# Inferences from the Reduced Feature Set

1. African American patients often face acute diabetes issues like ketoacidosis (1440), hyperosmolarity (993), and uncomplicated type II diabetes (819), highlighting the need for acute care improvements.
2. Asian patients primarily present with septicemia (41) and hypertension complications (25, 24), emphasizing infection control and blood pressure management.
3. Caucasian patients frequently deal with peptic ulcer hemorrhage (5343), coronary atherosclerosis (5085), and renal complications from diabetes (2973), indicating a need for specialized chronic care.
4. Hispanic patients commonly show circulatory (150) and neurological (117) diabetes complications and hypertension (91), pointing to integrated chronic disease management needs.
5. Patients of Other races present with coagulation issues (286), cancer (198), and diabetes (175), showing diverse healthcare needs.
6. Overall, specialties such as 'Internal Medicine' and 'Emergency/Trauma' account for the highest number of diagnoses, reflecting their critical role in managing complex and acute cases.

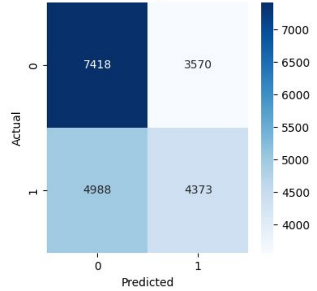
# Feature Scaling Technique

## SVD: Applied to approximate the matrix with 10 components

### Random Forest Results

| | Precision | Recall | F1-Score |
|---|---|---|---|
| **0** | 0.60 | 0.68 | 0.63 |
| **1** | 0.63 | 0.52 | 0.57 |
| **Accuracy** | | | 0.58 |
| **Macro Avg** | 0.57 | 0.57 | 0.57 |
| **Weighted Avg** | 0.58 | 0.58 | 0.57 |

Confusion Matrix for Readmission Prediction using Random Forest Classifier after Applying SVD



Time Taken: 61.9 Secs

O(NMK + TKNlog(N))

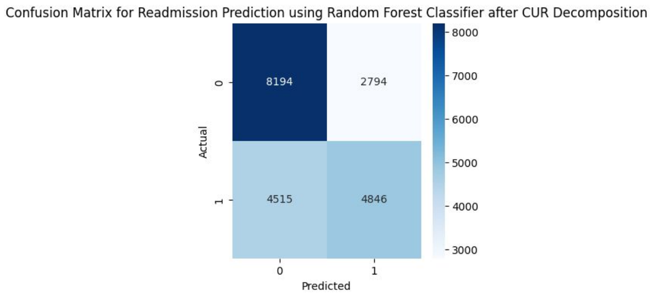
For NxM matrix, T stumps and K singular values

# Feature Scaling Technique

## CUR: The Matrix is about 60% Sparse, giving CUR an advantage

### Random Forest Results

| | Precision | Recall | F1-Score |
|---|---|---|---|
| **0** | 0.64 | 0.75 | 0.69 |
| **1** | 0.63 | 0.52 | 0.57 |
| **Accuracy** | | | 0.64 |
| **Macro Avg** | 0.64 | 0.63 | 0.63 |
| **Weighted Avg** | 0.64 | 0.64 | 0.64 |


Confusion Matrix for Readmission Prediction using Random Forest Classifier after CUR Decomposition

Time Taken: 20.7 Secs

$O(NM+R^2M+TRN\log(n))$

For NxM matrix, T stumps and R concepts

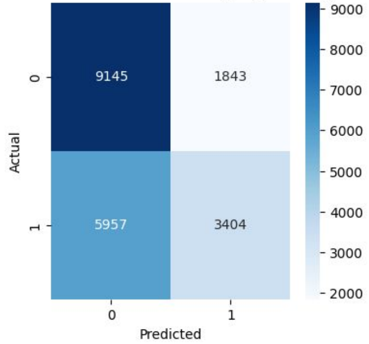
# Reducing Number of Samples

## JL Lemma: Applied on the 10 columns obtained from mutual info

### Logistic RegressionResults

| | Precision | Recall | F1-Score |
|---|---|---|---|
| **0** | 0.61 | 0.83 | 0.70 |
| **1** | 0.65 | 0.36 | 0.47 |
| **Accuracy** | | | 0.62 |
| **Macro Avg** | 0.63 | 0.60 | 0.58 |
| **Weighted Avg** | 0.63 | 0.62 | 0.59 |

Confusion Matrix for Readmission Prediction using Logistic Regression with JL Lemma



Time Taken: 4.6 Secs

$O(NKlog(N)/\epsilon^2)$

For NxM matrix and K features

# Comparison of Different ML Models

1. Combining Mutual Information with JL Lemma slightly reduced accuracy, indicating a trade-off between dimensionality reduction and data retention.
2. SVD-based dimensionality reduction also caused a marginal performance drop, highlighting the risk of aggressive dimension reduction.
3. CUR Decomposition preserved model performance (sparse nature of dataset), highlighting its efficiency in extracting key features from sparse data without any loss in accuracy.
4. Logistic regression on original data and the CUR Decomposition model both have the same accuracy but the reduced data due CUR results in faster execution time.

# Conclusion & Future Work

**Conclusion:**

1. Feature selection and dimensionality reduction techniques like Mutual Information, JL Lemma, CUR, and SVD can help reduce the complexity of the model and the running time but may also lose important information, as observed with some of the reduced accuracy scores.
2. The selected features using mutual information indicate that factors such as patient demographics, medical history, and healthcare interactions significantly influence diabetes readmission outcomes.
3. The analysis highlights racial disparities in diabetes complications and underscores the need for targeted, specialized, and acute care strategies to address diverse healthcare challenges effectively.

**Future Work:** Future work could involve exploring advanced feature selection techniques and leveraging domain knowledge-driven feature engineering, such as analyzing the implications of varying chemical dosages. This approach may help identify additional critical factors influencing readmission, enabling more tailored and effective patient care strategies.

# 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.

[2] Leskovec, J., Rajaraman, A., & Ullman, J. D. (n.d.). Dimensionality reduction. In Mining of Massive Datasets (pp. 405–408). http://infolab.stanford.edu/~ullman/mmds/ch11.pdf

[3] Mahoney, M. W., & Drineas, P. (2009). CUR matrix decompositions for improved data analysis. Proceedings of the National Academy of Sciences, 106(3), 697–702. https://doi.org/10.1073/pnas.0803205106

# Thank You