Predicting 30-Day Readmission Among Diabetic Inpatients

Animesh Mittal, Gareth Kumar, Sanskar Singh

April 8, 2025

1 Problem Formulation & Data Acquisition

Hospital readmissions among diabetic patients represent significant healthcare burdens, leading to increased costs and negative patient outcomes. Our project aims to predict whether diabetic patients will be readmitted within 30 days after hospital discharge. Predicting readmission can facilitate timely interventions, improve patient management, and ultimately reduce healthcare costs.

Diabetes-related complications and poor glycemic control frequently contribute to hospital readmissions. According to recent healthcare reports, these readmissions significantly strain healthcare resources and elevate healthcare expenditures, underscoring the importance of accurate prediction and prevention.

We utilized a publicly available dataset containing approximately 100,000 diabetic patient admissions from the UCI Machine Learning Repository. In total, the dataset has over 50 features capturing demographics, lab results (HbA1c), medications, admission types, discharge details, and readmission status. Each data entry represents a single inpatient encounter, providing ample data for robust modeling.

Our modeling focuses on variables that include demographic information (age, gender, race), clinical data (primary diagnosis, HbA1c levels, number of diagnoses), hospitalization details (admission type/source, discharge disposition, length of stay), and medication changes during hospitalization.

2 Data Cleaning & Preprocessing

Data preprocessing involved:

- **Dropping Columns:** We dropped columns with more than 90% missing data (e.g., weight) or those deemed irrelevant (payer code, medical specialty). This practice ensured that highly sparse or uninformative features did not negatively impact the model.
- **Handling Missing Values:** Most columns had no missing values or had NaN values flagged as “Missing.” For important feature columns with minor missing values, we removed the affected rows to maintain data integrity. We also encoded columns that were incorrectly flagged, ensuring accurate categorization.
- **Encoding Categorical Variables:** Admission types, discharge disposition, admission source, and primary diagnosis codes were transformed via one-hot encoding. Rare categories were consolidated to manage dimensionality and enhance model interpretability.

- **Outlier Management:** Outliers of numeric features were identified via boxplots. Due to a high presence of outliers in some variables, we focused on data points within 3 standard deviations (F-score under 3) to bolster model robustness and performance.

3 Exploratory Data Analysis (EDA)

Our EDA provided key visual and non-visual insights:

- **Readmission Distribution:** Bar charts indicated class imbalance, with only 10–12% readmissions.
- **Demographic Insights:** Race-based bar charts uncovered disparities in readmission rates, highlighting demographic influences on healthcare outcomes.
- **Disease Category:** Primary diagnosis bar charts revealed that patients diagnosed with circulatory and diabetic conditions had higher readmission rates.
- **Duration of Hospital Stay:** KDE plots demonstrated that longer hospital stays correlated significantly with increased readmission rates.
- **Age Group Analysis:** Age distribution visualizations showed higher readmissions among older patient groups.
- **Correlation Heatmap:** Revealed relationship strengths between numerical and newly engineered categories. The generally low correlation values across the matrix suggest no single feature strongly dominates others, indicating potential non-linear relationships.

3.1 Distribution of Readmission

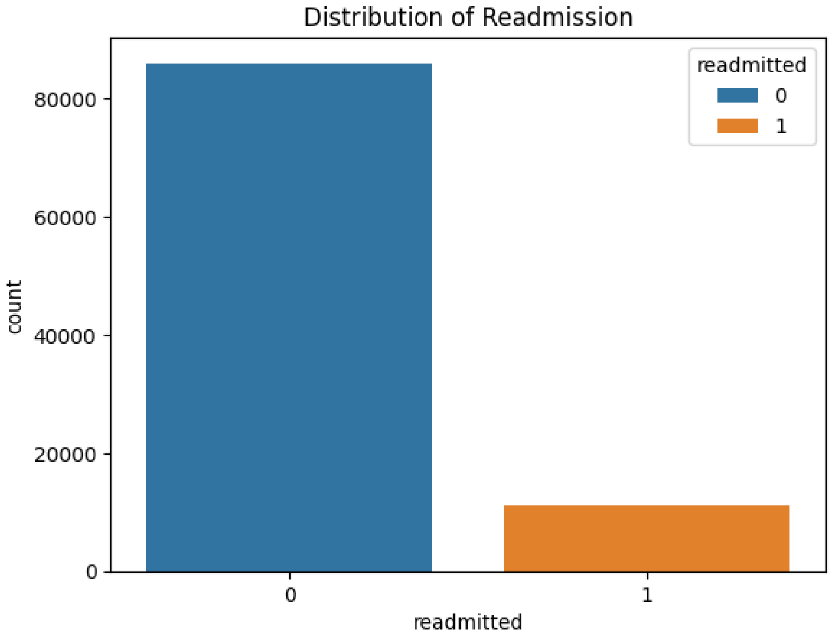

Figure 1: Overall readmission distribution across the dataset (imbalanced).

3.2 Distribution of Readmission by Race

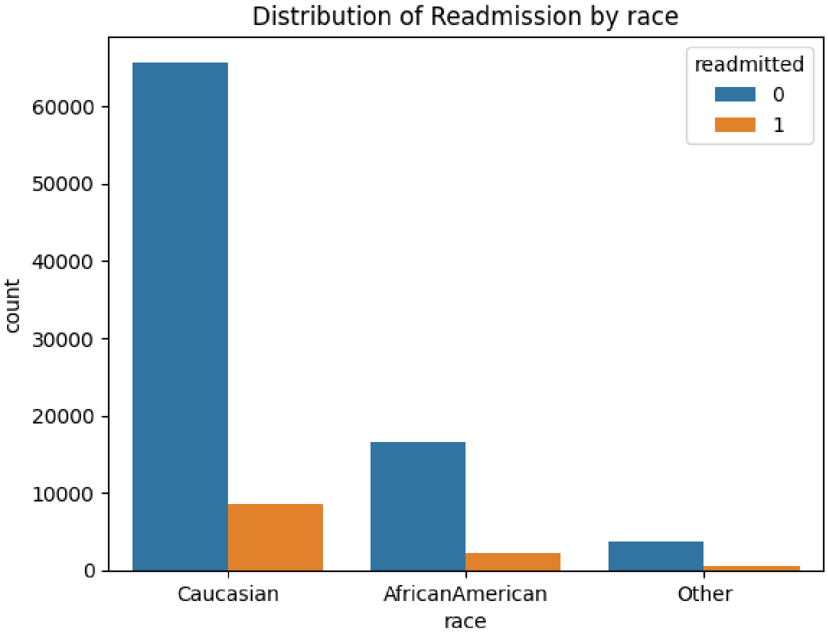

Figure 2: Readmission rates broken down by race, highlighting demographic disparities.

3.3 Correlation Heatmap

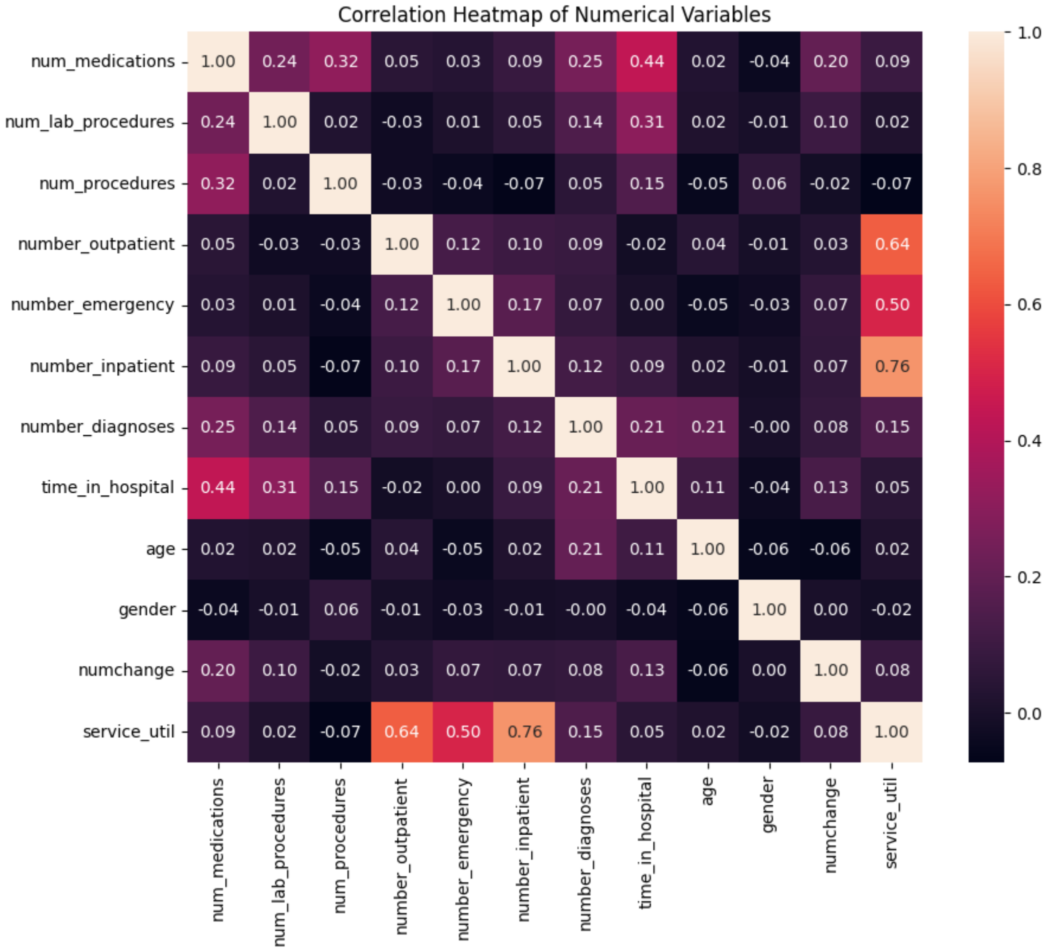

Figure 3: Heatmap illustrating correlations among key numerical and engineered variables. Lower correlation values might indicate non-linear relationships.

3.4 Number of Lab Procedures vs. Readmission

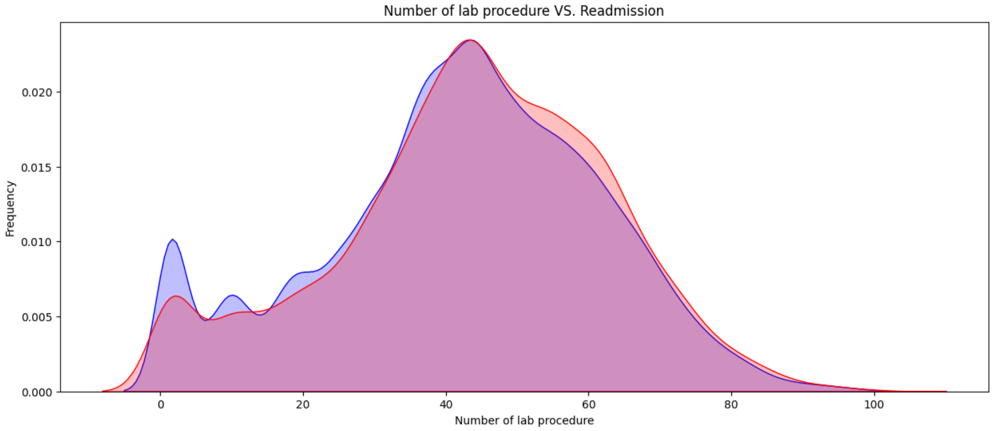

Figure 4: Relationship between the number of lab procedures and readmission outcomes.

4 Dimensionality Reduction & Resampling Strategies

We performed Principal Component Analysis (PCA) to reduce dimensionality, retaining 95% of the variance. PCA effectively condensed data features, mitigating multicollinearity and enhancing computational efficiency without significant information loss.

To address class imbalance, we employed:

- **Random Oversampling:** Augmented the minority class (readmitted patients), enhancing model sensitivity.
- **Random Undersampling:** Reduced majority class instances, balancing the dataset and facilitating unbiased model training.

5 Modeling

We evaluated Logistic Regression, Decision Tree, K-Nearest Neighbors, and Random Forest algorithms, applying hyperparameter tuning via grid search and cross-validation. Random Forest demonstrated superior predictive accuracy, robustness, and effective handling of complex interactions and non-linear relationships.

Dependent Variable: Readmission within 30 days (binary)

Independent Variables: Selected features derived from PCA and importance analysis: demographics, HbA1c, primary diagnosis, number of diagnoses, admission source, and hospital stay duration.

Random Forest mathematically aggregates predictions from multiple decision trees, each trained on bootstrapped samples:

$$\hat{y} = \text{majority vote}\{\text{Tree}_1(x), \text{Tree}_2(x), \ldots, \text{Tree}_n(x)\}$$

6 Model Evaluation

We validated our model using a sizeable unseen test set (30% of data), employing metrics:

- **Accuracy:** Evaluating overall prediction correctness.
- **Precision:** Assessing the accuracy of positive readmission predictions, reducing false positives.
- **Recall:** Measuring the model's ability to capture actual readmissions, crucial for timely interventions.

Random Forest, combined with PCA and resampling techniques, exhibited excellent predictive capabilities, minimal variance between training and testing performances, and strong generalization potential.

6.1 Model Performance Comparison

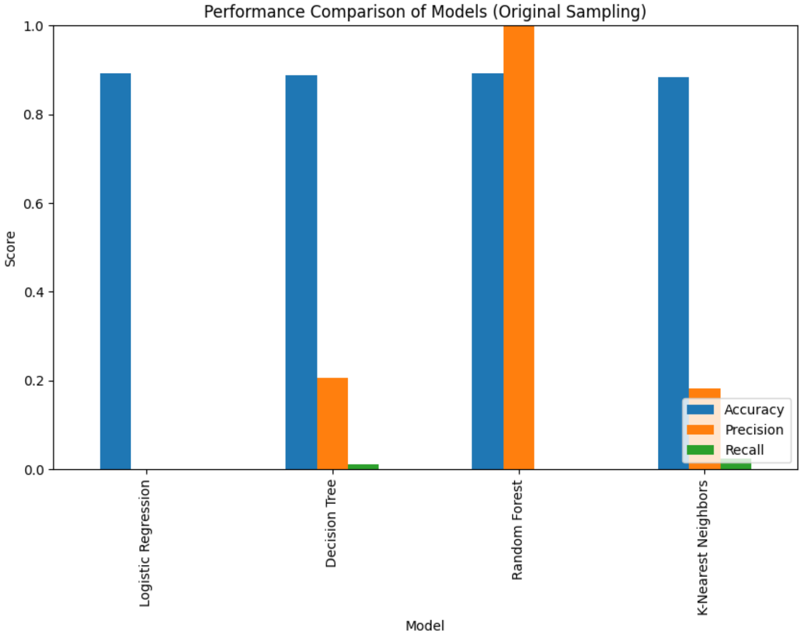

Figure 5: Model performance metrics (Accuracy, Precision, Recall) on original sampling.

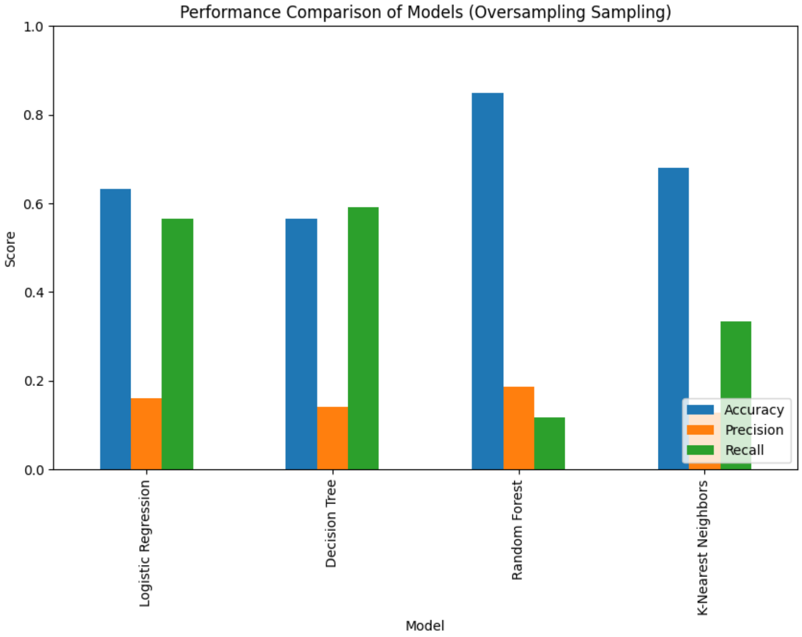

Figure 6: Model performance with oversampling approach.

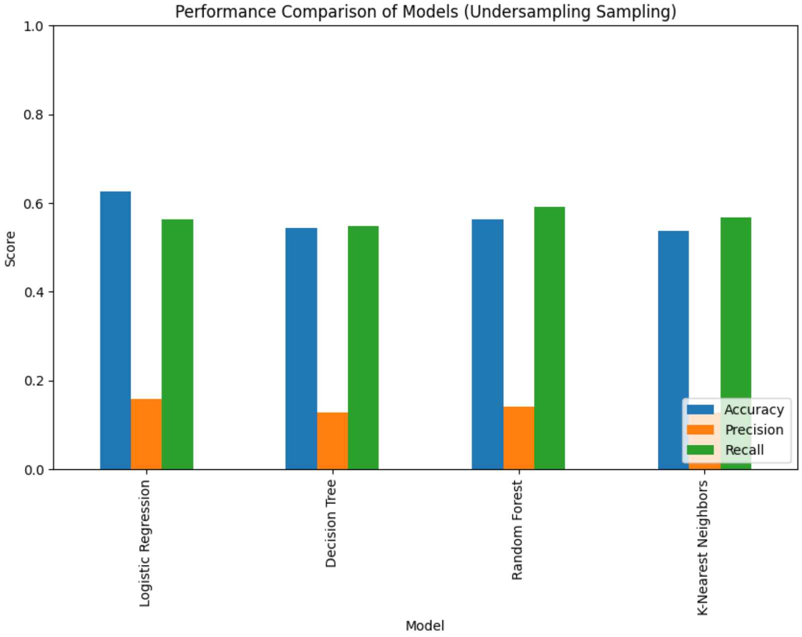

Figure 7: Model performance with undersampling approach.

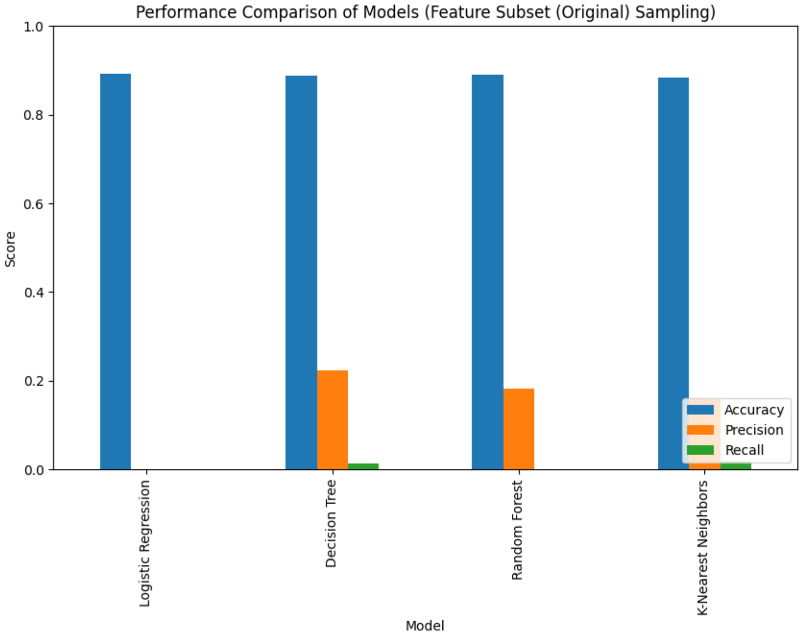

Figure 8: Model performance with feature subset selection (original sampling).

Since Random Forest performed best among all the evaluated algorithms, we compared its performance across four different data configurations (original, oversampled, undersampled, and subset). We focused on test-set accuracy and the F1-score to gauge predictive quality. Our findings show that Random Forest achieves the highest performance on the original dataset (Accuracy: 89.16%, Precision: 90.34%, Recall: 89.16%, F1-score: 84.06%). Interestingly, the subset dataset also produces comparable outcomes, though slightly lower than the original data. This suggests that data quality and representativeness play a pivotal role in boosting the model's predictive capability.

7 Feature Importance & Subset Selection

Using Random Forest—our best-performing algorithm—we derived feature importance scores (Mean Decrease Impurity). Hospital stay duration, number of diagnoses, and HbA1c levels emerged as critical predictors. This insight guided variable selection for model refinement, significantly boosting accuracy and interpretability.

Here are our final Random Forest evaluation metrics on the test set:

- **Accuracy:** 89.16%
- **Precision:** 90.34%
- **Recall:** 89.16%
- **F1-score:** 84.06%

These results confirm the model's strong ability to correctly classify readmissions and highlight its suitability as the preferred model for this prediction task.

8 Interpretation of Results & Conclusion

The final model accurately identified hospital stay duration, number of diagnoses, and elevated HbA1c as key predictors. These findings have practical significance, pointing to targeted clinical interventions and optimized resource allocation.

Strengths:

- High robustness to complex data and imbalance.
- Strong predictive accuracy, providing valuable clinical insights.

Shortcomings:

- Lower interpretability compared to simpler, linear models.

Future research should incorporate lifestyle and socioeconomic variables, and detailed medication adherence records to enhance model comprehensiveness and predictive accuracy.

9 Documentation & Citations

We extensively cited Strack et al. (2014), emphasizing HbA1c's clinical significance:
Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, 2014. https://doi.org/10.1155/2014/781670