Prediction on Hospital Readmission for Diabetic Patients

CMPUT 195

UNIVERSITY OF ALBERTA

INFO ARCHITECTS:

ANIMESH MITTAL
GARETH KUMAR
SANSKAR SINGH

TABLE OF CONTENTS

PROBLEM:

Predict whether a diabetic patient will be readmitted within 30 days

WHY THIS TOPIC

- Diabetes readmissions = major healthcare & patient burden.
- Frequent complications (poor glycemic control, comorbidities).
- Opportunity to reduce costs & improve outcomes through data insights.

WHAT DOES THIS TOPIC ADDRESS

- Using diabetic inpatient data: demographics, labs (HbA1c), medications, readmission within 30 days.
- Key question: *What factors best predict early readmission for diabetics?*

WHY DATA SCIENCE

- **Pattern Discovery**: Correlate patient characteristics, lab results, treatments with outcomes.
- **Predictive Modeling**: Estimate the probability of readmission, enabling proactive intervention.
- **Impact**: Improve care strategies, reduce costs, enhance patient health outcomes.

INITIAL EXPLORATION

- ## DATASET SCOPE
  - 100,000 diabetic hospital admissions (10-year span)
  - Each row = single inpatient encounter
  - Includes demographics, (HbA1c), admission type, discharge info, and 30-day readmission status
- ## KEY STEPS
  - Reviewed the data dictionary
  - Checked missing values, readmission rates
  - *Observations*:
    - Many missing/unknown values (e.g., weight)
    - Repeated patient encounters (decide how to handle duplicates)
    - Categorical variables with many levels (discharge disposition)
- ## EARLY FINDINGS
  - 10–12% readmissions within 30 days
  - HbA1c missing frequently → indicates potential for under-measurement
  - Medication changes varied widely among encounters

Example Dataset:

encounter_id	patient_nbr	race	gender	age	weight	admission_type_id
2278392	8222157	Caucasian	Female	[0-10)	?	6
149190	55629189	Caucasian	Female	[10-20)	?	1
64410	86047875	AfricanAmerican	Female	[20-30)	?	1
500364	82442376	Caucasian	Male	[30-40)	?	1
16680	42519267	Caucasian	Male	[40-50)	?	1
35754	82637451	Caucasian	Male	[50-60)	?	2
55842	84259809	Caucasian	Male	[60-70)	?	3
63768	114882984	Caucasian	Male	[70-80)	?	1
12522	48330783	Caucasian	Female	[80-90)	?	2
15738	63555939	Caucasian	Female	[90-100)	?	3
28236	89869032	AfricanAmerican	Female	[40-50)	?	1
36900	77391171	AfricanAmerican	Male	[60-70)	?	2
40926	85504905	Caucasian	Female	[40-50)	?	1
42570	77586282	Caucasian	Male	[80-90)	?	1
62256	49726791	AfricanAmerican	Female	[60-70)	?	3
73578	86328819	AfricanAmerican	Male	[60-70)	?	1
77076	92519352	AfricanAmerican	Male	[50-60)	?	1
84222	108662661	Caucasian	Female	[50-60)	?	1

Part I Data Cleaning

- ## DROPPING LOW-VALUE COLUMNS
  - Dropped columns with >95% missing (e.g., weight)
  - Removed sparse/unrelated attributes (e.g., payer code if highly incomplete)
  - **Payer Code & Medical Specialty:** ~50% missing → Dropped
- ## HANDLING MISSING VALUES
  - **Categorical**: Created a “Missing” category for moderately missing features (e.g., medical_specialty)
  - **Numeric**:
    - Small missing? → Rows dropped
    - High missing? → Imputed using mean/median or flagged as “No value”
- ## ENCODING CATEGORICAL VARIABLES
  - One-hot or label-encoding for admission_type, discharge_disposition, ICD-9 diagnoses, etc.
  - Merged rare categories into “Other” to avoid high dimensionality
- ## MANAGING OUTLIERS
  - Only included data within 3 standard deviations for each numerical column.
  - Aware they may affect model performance

Part II (Feature Engineering)

DISCHARGE DISPOSITION PROCESSING

- Drop rows with discharge disposition values: (representing death/hospice)
- Created new column:
  - “home” if original value is 1, otherwise “other”

ADMISSION TYPE PROCESSING

- Collapse admission type values:
  - Merge types 2 & 7 to type 1
  - Merge types 6 & 8 to type 5
- Mapped numbers to meaning:
  - Eg: 1 → “emergency”
  - 3 → “elective”

ADMISSION SOURCE PROCESSING

- Collapse 21 distinct values into three categories:Types 1,2,3 → “physician referral”
  - Remaining types → “other” (if in a defined list)
  - Others → “emergency room”

Part III (Diagnosis & Age)

PRIMARY DIAGNOSIS (diag_1) MAPPING:

- Mapped ICD-9 codes to disease categories:

  Codes beginning with "E" or "V": **Other**
  - “250”: **Diabetes**
  - In [390,459] or equals 785: **Circulatory**
  - In [460,519] or equals 786: **Respiratory**
  - In [520,579] or equals 787: **Digestive**
  - In [800,999]: **Injury**
  - In [710,739]: **Musculoskeletal**
  - In [580,629] or equals 788: **Genitourinary**
  - In [140,239]: **Neoplasms**
  - All other values: **Other**
- One-hot encode these disease types and drop the original *diag_1* column.

AGE TRANSFORMATION:

- Convert age ranges (e.g., “[0-10)”, “[10-20)”) into three classes:
  - **<30, 30-60, 60+**
- One-hot encode these age groups.

ADDITIONAL PREPROCESSING:

- Rename column: A1Cresult → HbA1c
- Drop medication columns with only one constant value: ‘citoglipton’, ‘examide’
- Remove outliers from continuous variables using the IQR method

EXPLORATORY DATA ANALYSIS (EDA)

KEY VISUALIZATIONS:

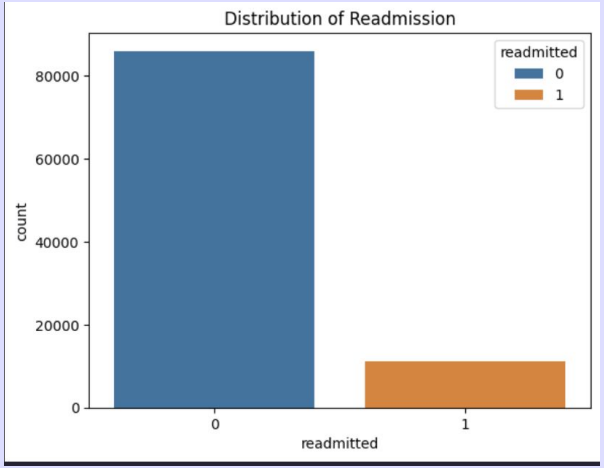
- Histograms & KDE plots of continuous variables (time in hospital, lab procedures, etc.)
- Correlation heatmap of continuous predictors
- Pairplots (selected variables colored by readmission status)

INSIGHTS:

- Observed class imbalance in target (readmitted vs. not readmitted)
- Specific variables (e.g., time in hospital, number of diagnoses) appear to differentiate readmitted patients

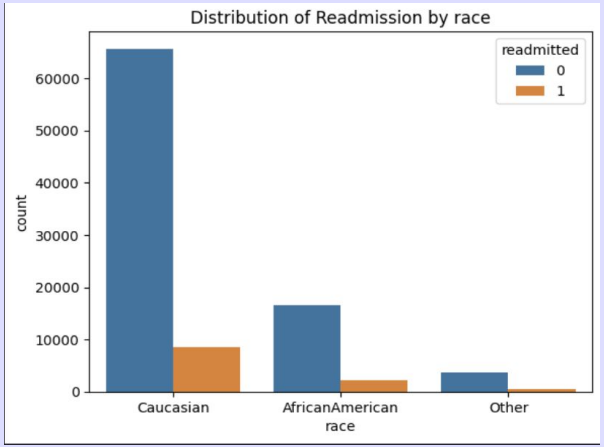
Distribution of Readmission (Bar Chart)

- **Shows the overall class imbalance**: most encounters are not readmitted (0) vs. a smaller portion readmitted (1).
- **Indicates the need for resampling** (oversampling/undersampling) to address minority class detection.



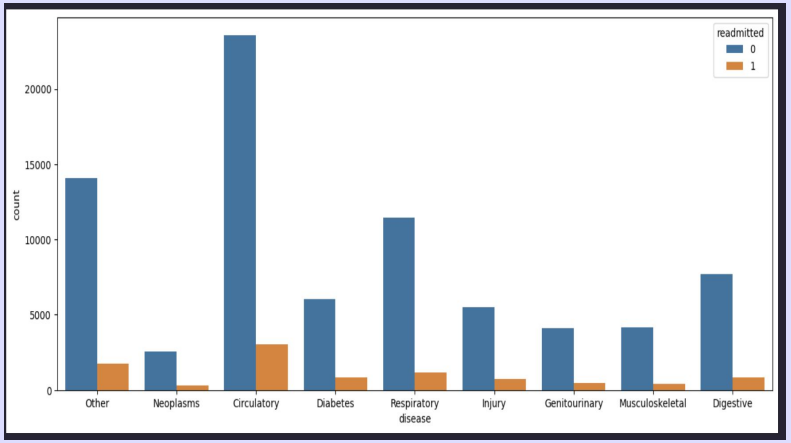
Distribution of Readmission by Race (Bar Chart)

- **Breaks down readmission** rates among different racial groups.
- **Identifies if any group** has notably higher or lower readmission.
- Provides insight into **demographic disparities** and model fairness concerns.



Disease category vs. Readmission status:

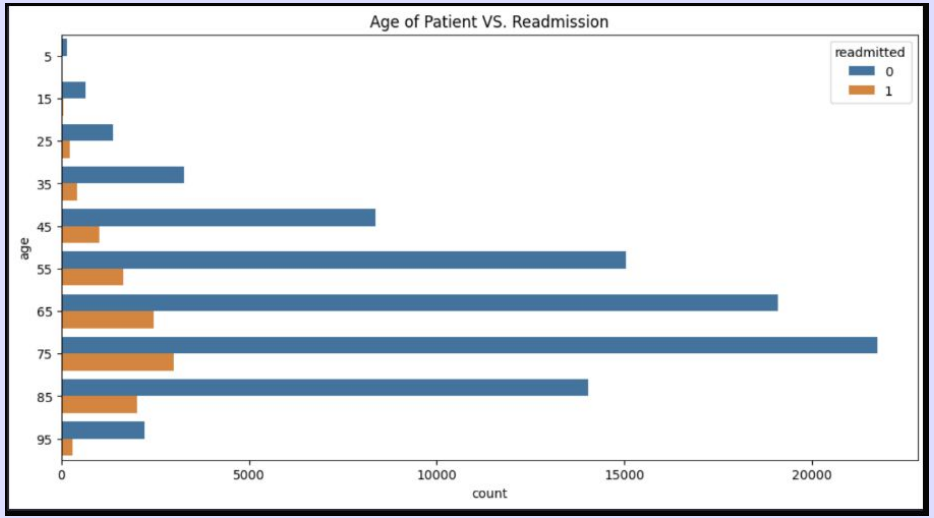
- **Shows primary diagnosis categories** (e.g., Circulatory, Diabetes, etc.) and how often patients in each group were readmitted (0 = no, 1 = yes).
- **Highlights dominant categories:** For example, Circulatory might have the highest encounter volume overall.
- **Exposes relative readmission rates:** Some categories (e.g., Diabetes) may show a higher proportion of readmitted patients than others.
- **Reveals potential focus areas:** Categories with a notable share of readmissions could benefit from targeted interventions.



Time in Hospital vs. Readmission (KDE Plot)

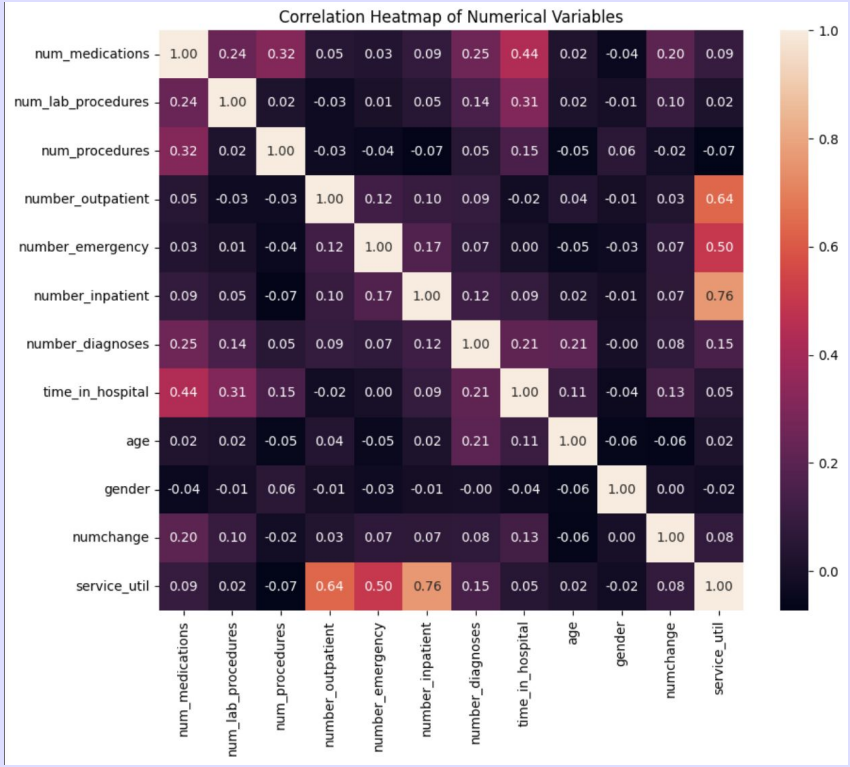
- **Visualizes hospital stay duration** distribution for readmitted vs. not readmitted.
- **Suggests longer stays** may be linked to higher readmission likelihood.
- Highlights a key predictor in early readmission modeling.

- **Displays age distribution** for readmitted (1) vs. not readmitted (0).
- **Highlights which age groups** have higher overall hospital visits.
- **May indicate increased readmission** among certain older age categories.



Correlation Heatmap: Key Insights

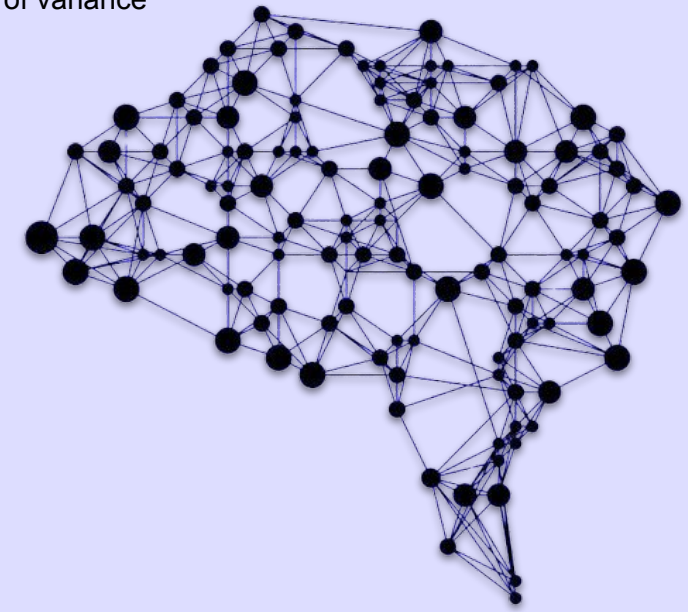
- **Highlights pairwise relationships:** Reveals which numerical features (e.g., hospital stay, lab procedures, medications, diagnoses) are strongly interrelated.
- **Identifies multicollinearity:** Very high positive correlations suggest some features might be redundant; supports the case for dimensionality reduction or feature selection.
- **Informs model development:** Strong correlations among clinical measures may indicate underlying patient severity; these relationships can guide feature grouping and refinement.
- **Aids in interpretation:** Understanding correlated predictors helps in deciphering the model's decision process and ensures more robust analyses.



DIMENSIONALITY REDUCTION & RESAMPLING STRATEGIES

PRINCIPAL COMPONENT ANALYSIS (PCA):

- Applied PCA (post scaling) to reduce dimensionality while retaining 95% of variance
- Reported the number of principal components obtained



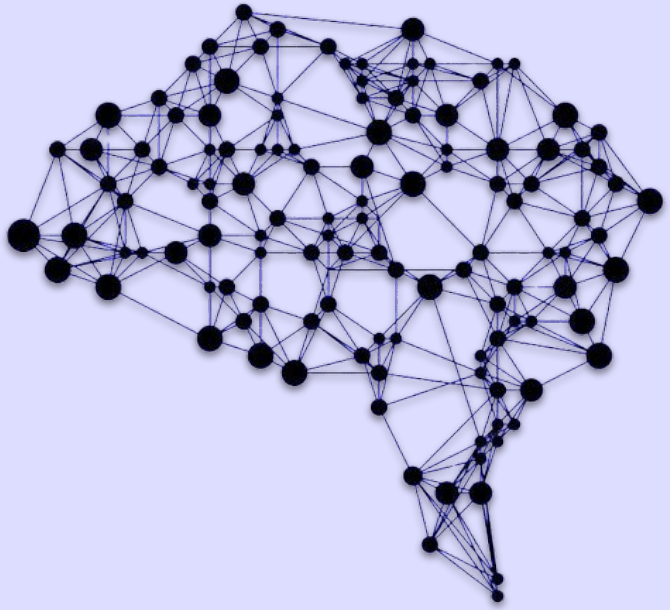
RESAMPLING TECHNIQUES:

- **Random Oversampling:** Augment minority class
- **Random Undersampling:** Reduce majority class
- Models are run on:
    - Original imbalanced data
    - Oversampled dataset
    - Undersampled dataset

MODELING - ALGORITHMS & IMPLEMENTATION

MODELS TRAINED

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- K-Nearest Neighbors



IMPLEMENTATION DETAILS

Models evaluated using classification reports and confusion matrices

- Separate evaluations are conducted for:
    - Original (imbalanced) training set
    - Oversampled training set
    - Undersampled training set

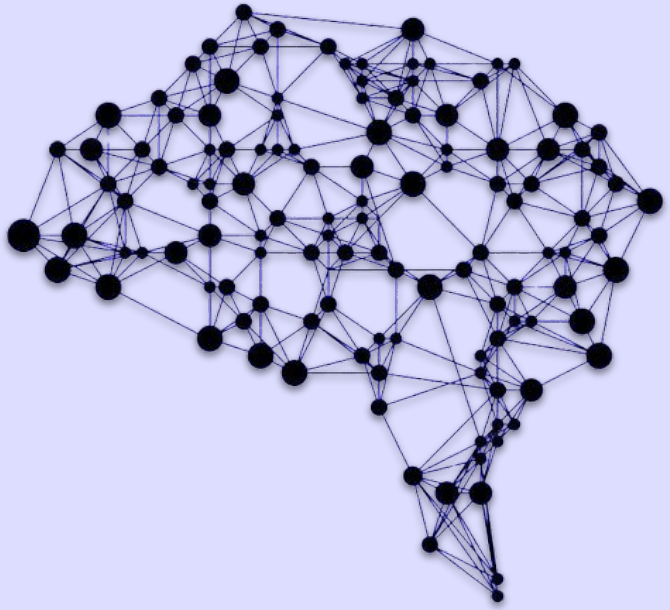
PERFORMANCE METRICS VISUALIZED

- Bar charts comparing Accuracy, Precision, and Recall across models and sampling methods

FEATURE IMPORTANCE & SUBSET MODELING

FEATURE IMPORTANCE

- Random Forest used to extract importance scores for all features
- Top features (e.g., time in hospital, number of diagnoses, key demographic/encoded predictors) are highlighted in a bar plot



SUBSET MODELING

- Selected the most important features
- Re-ran the four models using only this subset
- Compared performance of subset models versus full-feature models

Final Results & Comparative Analysis

The experiment first applies PCA for dimensionality reduction, then evaluates various classification models under different sampling conditions. It reports performance metrics for each model configuration, including warnings about undefined precision for minority classes. The main modeling approaches include:

- Modeling on **original (imbalanced) data** without any resampling.
- Modeling after **random oversampling** of the minority class.
- Modeling after **random undersampling** of the majority class.
- Modeling using a **selected subset of important features** (feature subset modeling).

PCA Dimensionality Reduction

- **Initial Feature Set:** 57 features.
- **After PCA:** The feature set is reduced to 43 principal components.
  This reduction is aimed at retaining the most significant variance while potentially improving model performance and reducing overfitting.

Final Results & Comparative Analysis

Modeling on Original Data (No Resampling)

In this stage, the models are trained and evaluated using the imbalanced dataset.

Logistic Regression

- **Performance Summary:**
  - **Class 0:** F1-Score: 0.94
  - **Class 1:** F1-Score: Reported as 0.00

- **Overall Metrics:**
  - **Macro Average:** F1-Score: 0.47
  - **Weighted Average:** F1-Score: 0.84

Decision Tree Classifier

- **Performance Summary:**
  Similar trends are observed:
  - **Class 0:** Good scores F1: 0.94
  - **Class 1:** Very low performance F1: 0.02

- **Overall Metrics:**
  - **Accuracy:** 0.89
  - **Macro Average:** F1-Score: 0.48
  - **Weighted Average:** F1-Score: 0.84

Final Results & Comparative Analysis

Random Forest Classifier

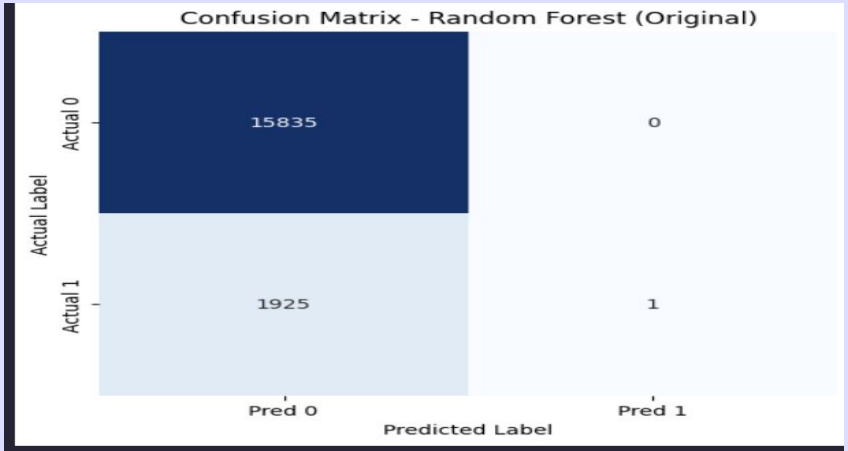
- **Performance Summary:**
  The report shows similar performance to the Logistic Regression and Decision Tree for Class 0, whereas Class 1 remains problematic.
  - **Class 0:** F1-score (approximately 0.89–0.94).
  - **Class 1:** F1-score of 0.00.

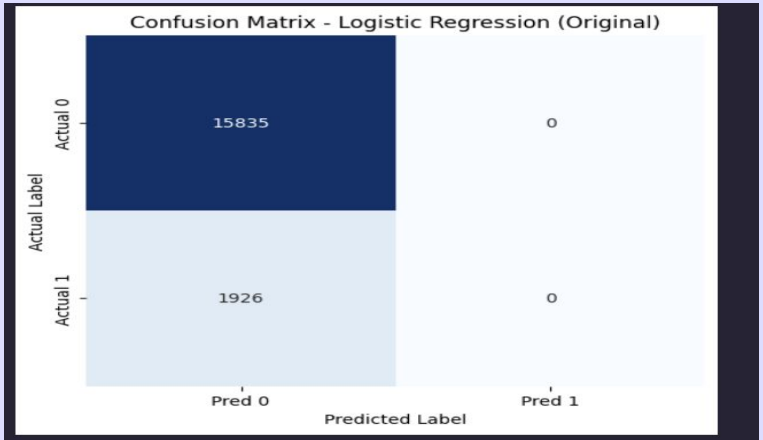
- **Overall Metrics:**
  - **Accuracy:** 0.89
  - **Macro Average:** F1-Score: 0.47
  - **Weighted Average:** F1-Score: 0.84

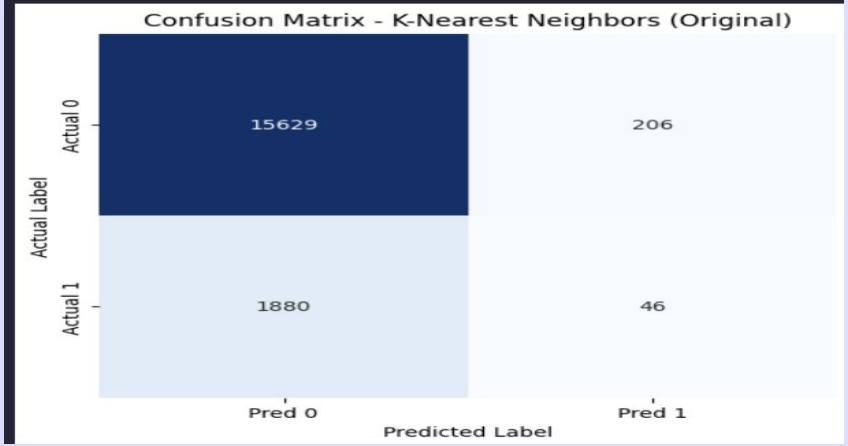
K Nearest Neighbour

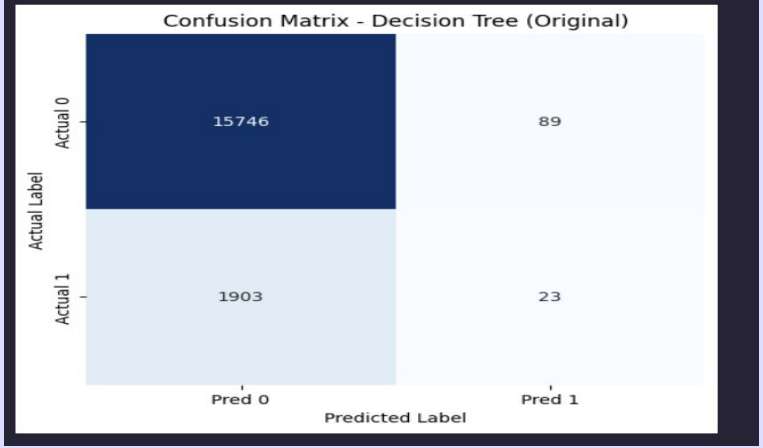
- **Performance Summary:**
  The report shows similar performance to Decision Tree for Class 0, whereas Class 1 remains problematic.
  - **Class 0:** F1-score (approximately 0.8–0.9).
  - **Class 1:** F1-score of 0.00.

- **Overall Metrics:**
  - **Accuracy:** 0.79
  - **Macro Average:** F1-Score: 0.37
  - **Weighted Average:** F1-Score: 0.74


Confusion Matrix - Random Forest (Original)
Actual Label
Actual 0
Actual 1
15835
0
1925
1
Pred 0
Pred 1
Predicted Label


Confusion Matrix - Logistic Regression (Original)
Actual Label
Actual 0
Actual 1
15835
0
1926
0
Pred 0
Pred 1
Predicted Label


Confusion Matrix - K-Nearest Neighbors (Original)
Actual Label
Actual 0
Actual 1
15629
206
1880
46
Pred 0
Pred 1
Predicted Label


Confusion Matrix - Decision Tree (Original)
Actual Label
Actual 0
Actual 1
15746
89
1903
23
Pred 0
Pred 1
Predicted Label

Final Results & Comparative Analysis

Modeling on Oversampled Data

After performing **random oversampling**, the minority class is increased to balance with the majority:

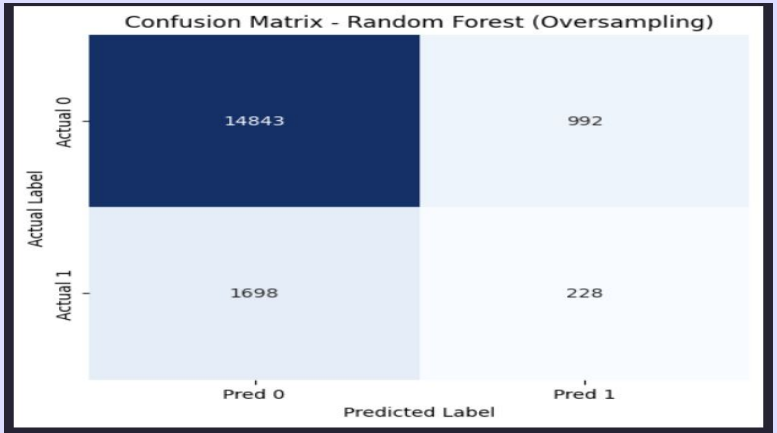
- **New Minority and Majority Counts:** [63,335 each].

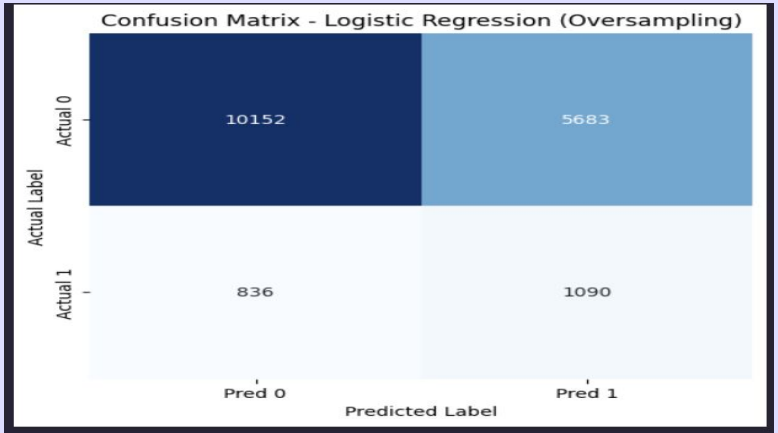
Random Forest

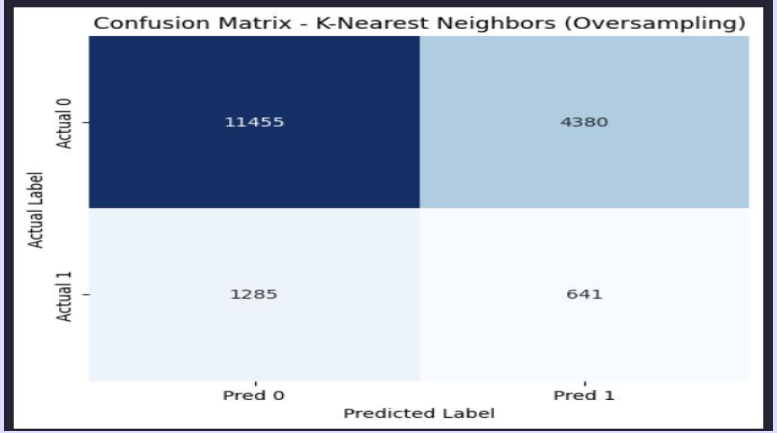
- **Performance Summary:**
  - **Class 0:** F1-Score: 0.94
  - **Class 1:** F1-Score: 0.25
- **Overall Metrics:**
  - **Accuracy:** 0.63
  - **Macro Average:** F1-Score: 0.50
  - **Weighted Average:** F1-Score: 0.70

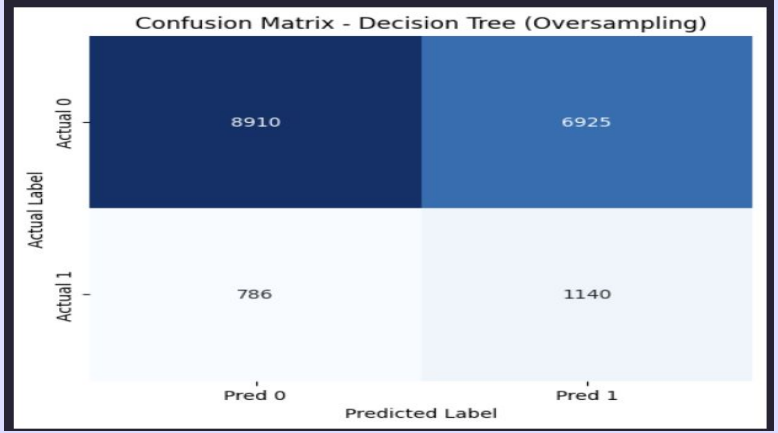
Decision Tree Classifier

- **Performance Summary:**
  - **Class 0:** F1-Score: 0.70
  - **Class 1:** F1-Score: 0.23
- **Overall Metrics:**
  - **Accuracy:** 0.68
  - **Macro Average:** F1-Score: 0.49
  - **Weighted Average**: F1-Score: 0.73


Confusion Matrix - Random Forest (Oversampling)
Actual Label
Actual 0
Actual 1
14843
992
1698
228
Pred 0
Pred 1
Predicted Label


Confusion Matrix - Logistic Regression (Oversampling)
Actual Label
Actual 0
Actual 1
10152
5683
836
1090
Pred 0
Pred 1
Predicted Label


Confusion Matrix - K-Nearest Neighbors (Oversampling)
Actual Label
Actual 0
Actual 1
11455
4380
1285
641
Pred 0
Pred 1
Predicted Label


Confusion Matrix - Decision Tree (Oversampling)
Actual Label
Actual 0
Actual 1
8910
6925
786
1140
Pred 0
Pred 1
Predicted Label

Modeling on Undersampled Data

With **random undersampling**, the majority class is reduced to balance the classes:

- **New Counts:** [7,705 for both the majority and minority classes].

Logistic Regression (Undersampled Data)

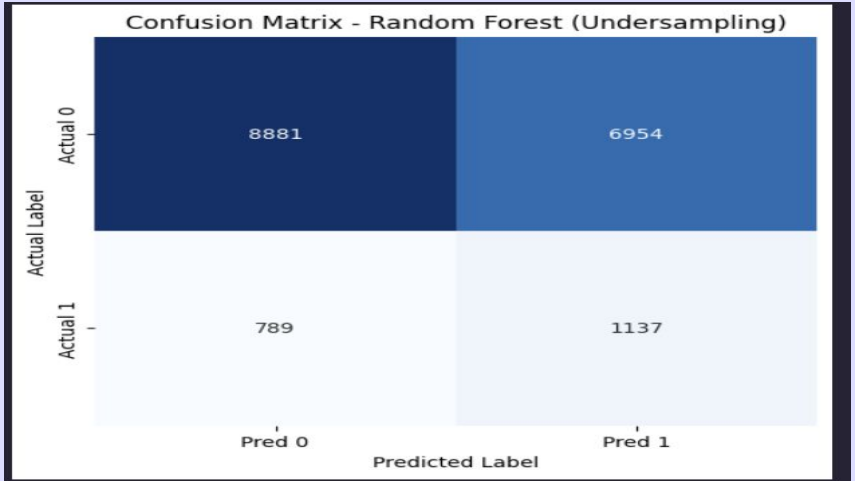
- **Performance Summary:**
  - **Class 0:** F1-Score: 0.75
  - **Class 1:** F1-Score: 0.25

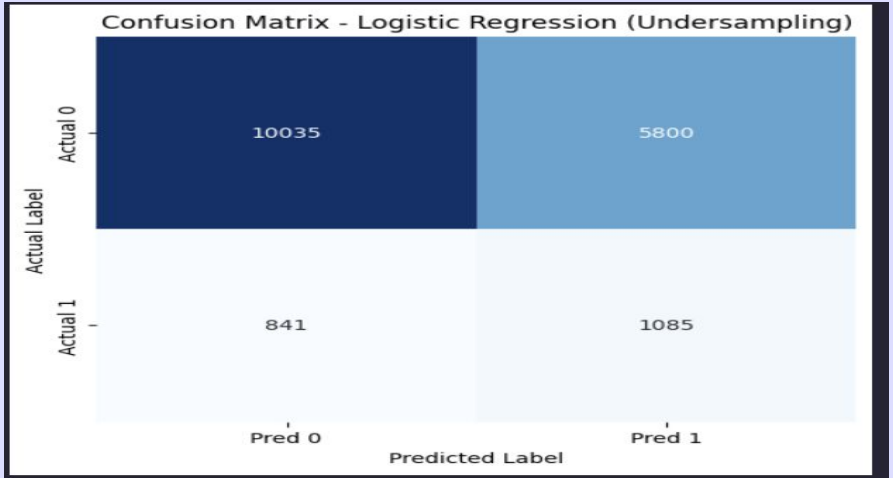
- **Overall Metrics:**
  - **Accuracy:** 0.63
  - **Macro Average:** F1-Score: 0.50
  - **Weighted Average:** F1-Score: 0.70

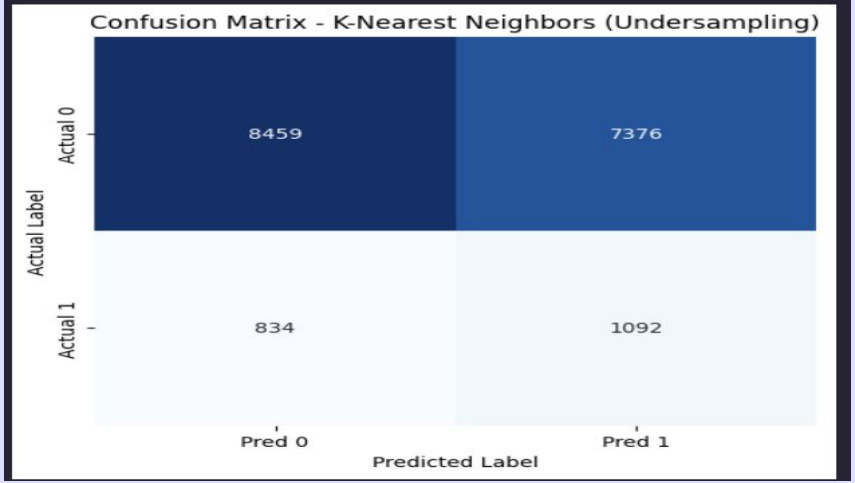
Final Results & Comparative Analysis

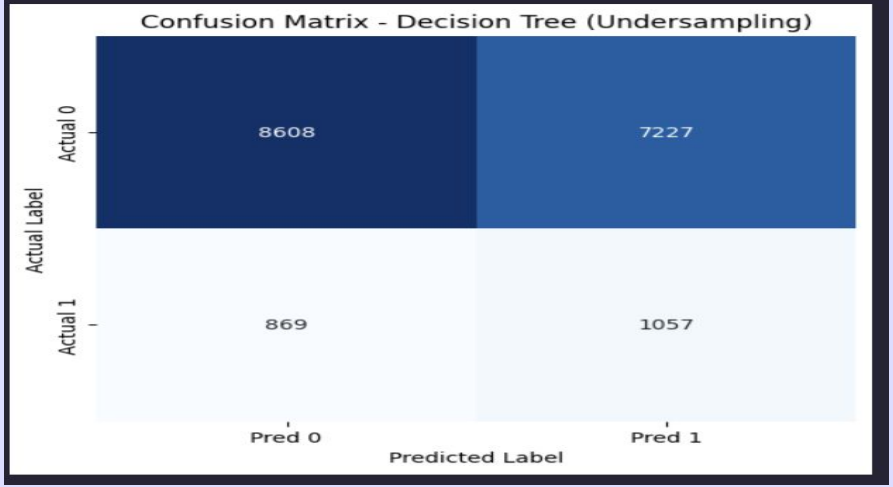
Decision Tree Classifier (Undersampled Data)

- **Performance Summary:**
    - **Class 0:** F1-Score: 0.68
    - **Class 1:** F1-Score: 0.21
- **Overall Metrics:**
    - **Accuracy:** 0.54
    - **Macro Average:** F1-Score: 0.44
    - **Weighted Average:** F1-Score: 0.62


Confusion Matrix - Random Forest (Undersampling)
Actual Label
Actual 0
Actual 1
8881
6954
789
1137
Pred 0
Pred 1
Predicted Label


Confusion Matrix - Logistic Regression (Undersampling)
Actual Label
Actual 0
Actual 1
10035
5800
841
1085
Pred 0
Pred 1
Predicted Label


Confusion Matrix - K-Nearest Neighbors (Undersampling)
Actual Label
Actual 0
Actual 1
8459
7376
834
1092
Pred 0
Pred 1
Predicted Label


Confusion Matrix - Decision Tree (Undersampling)
Actual Label
Actual 0
Actual 1
8608
7227
869
1057
Pred 0
Pred 1
Predicted Label

Feature Subset Modeling

Instead of reducing dimensions via PCA, a **feature subset** was chosen based on importance. The selected features include:

- **Selected Features:**
  `['num_lab_procedures', 'num_medications', 'time_in_hospital', 'age', 'num_procedures', 'number_diagnoses', 'service_util', 'number_inpatient', 'gender', 'A1Cresult', 'number_outpatient', 'insulin', 'source_emergency room', 'number_emergency', 'Caucasian', 'diag_Circulatory', 'source_physician referral', 'change', 'metformin', 'admission_emergency']`

The modeling is repeated using these features.

Final Results & Comparative Analysis

Logistic Regression (Feature Subset)

- **Performance Summary:**
  - **Class 0:**
    - F1-Score: 0.94
  - **Class 1:**
    - F1-Score: 0.00 (indicating that the minority class is not captured)

- **Overall Metrics:**
  - **Accuracy:** 0.89
  - **Macro Average:**
    - F1-Score: 0.47
  - **Weighted Average:**
    - F1-Score: 0.84

Final Results & Comparative Analysis

Decision Tree Classifier (Feature Subset)

- **Performance Summary:**
  Similar to the original data results:
  - **Class 0:** High metrics (precision ~0.89, recall ~0.99, F1 ~0.94).
    Class 1: Very low performance (precision ~0.22, recall ~0.01, F1 ~0.02).
- **Overall Metrics:**
  - Weighted average metrics closely follow those seen in previous models (precision ~0.82, recall ~0.89, F1 ~0.84).

Random Forest Classifier and K-Nearest Neighbors Classifier

- **Random Forest:**
  - Similar issues occur with Class 1 having undefined or very low performance; warning messages about undefined precision continue to appear.
- **K-Nearest Neighbors:**
  - **Class 0:** High precision and recall (around 0.89–0.94).
  - **Class 1:** Performance remains very low (precision ~0.16, recall ~0.02, F1 ~0.03).
- **Overall Accuracy:**
  - KNN reports an overall accuracy of 0.88 with similar macro and weighted averages indicating significant class imbalance effects.

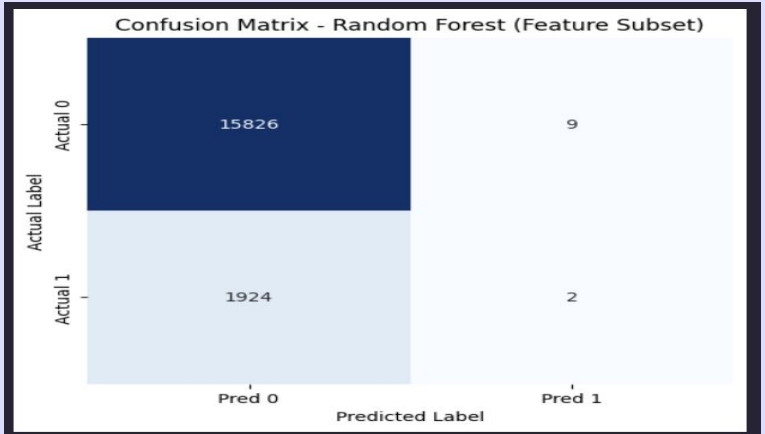
Concluding Remarks

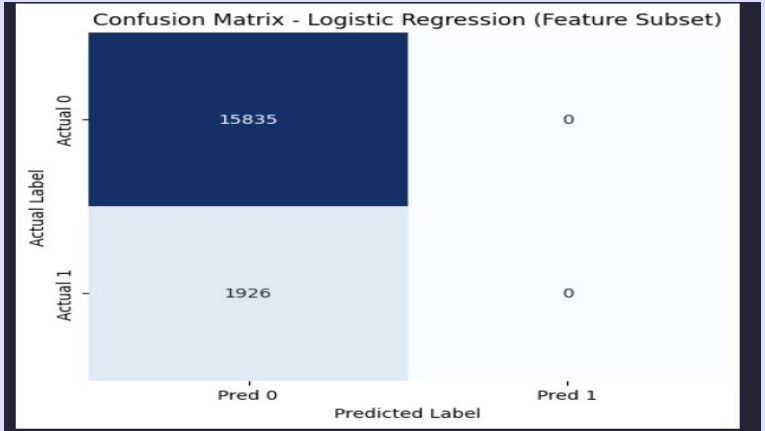
- **Class Imbalance Impact:**
  The experiments consistently show that while the models perform very well on the majority class (Class 0), the minority class (Class 1) suffers from very low recall and F1-scores. This issue is a common challenge in imbalanced classification tasks.

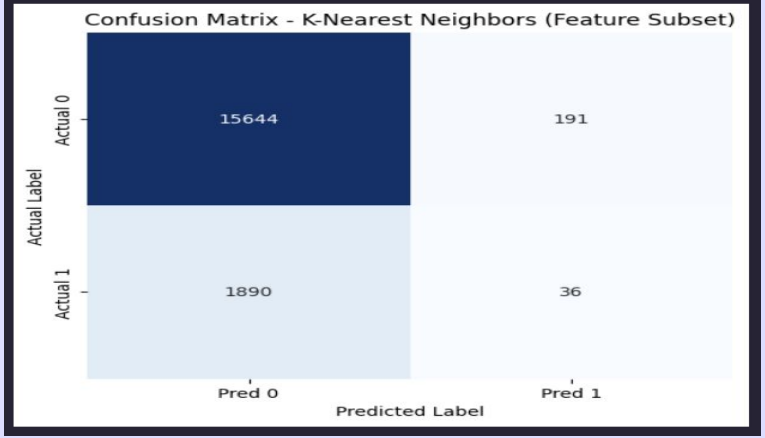
- **Effect of Resampling:**
  - **Oversampling** improved the recall of the minority class, but overall accuracy dropped.
  - **Undersampling** balanced the class counts but resulted in a reduced overall accuracy (as more data from the majority class was removed).

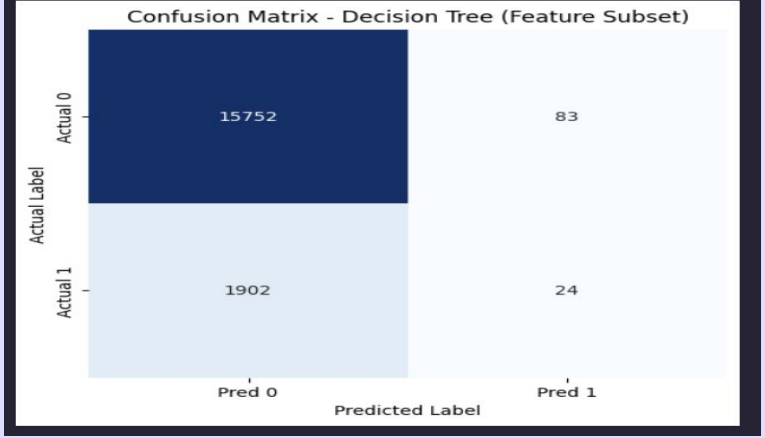
- **Feature Selection vs. PCA:**
  - Using **PCA** reduces the feature dimensionality but might mask individual feature effects.
  - The **feature subset** approach leverages domain-relevant features but still faces challenges in capturing the minority class in the models.

- **Warning Messages:**
  Throughout the experiments, warnings related to undefined precision for labels with no predicted samples highlight potential pitfalls when dealing with rare events in classification tasks.


Confusion Matrix - Random Forest (Feature Subset)
Actual Label
Actual 0
Actual 1
15826
9
1924
2
Pred 0
Pred 1
Predicted Label

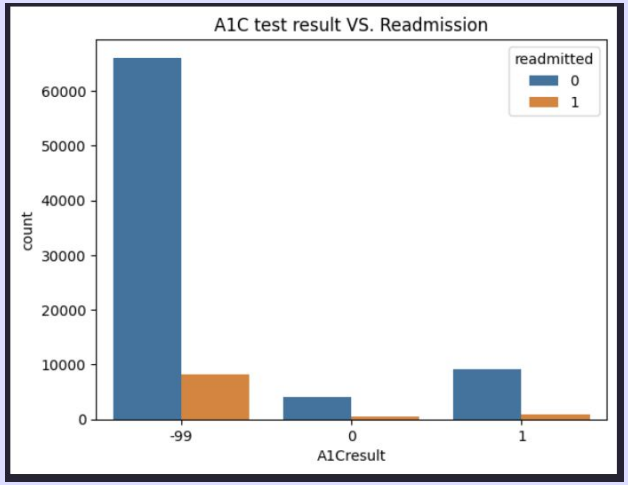

Confusion Matrix - Logistic Regression (Feature Subset)
Actual Label
Actual 0
Actual 1
15835
0
1926
0
Pred 0
Pred 1
Predicted Label


Confusion Matrix - K-Nearest Neighbors (Feature Subset)
Actual Label
Actual 0
Actual 1
15644
191
1890
36
Pred 0
Pred 1
Predicted Label


Confusion Matrix - Decision Tree (Feature Subset)
Actual Label
Actual 0
Actual 1
15752
83
1902
24
Pred 0
Pred 1
Predicted Label

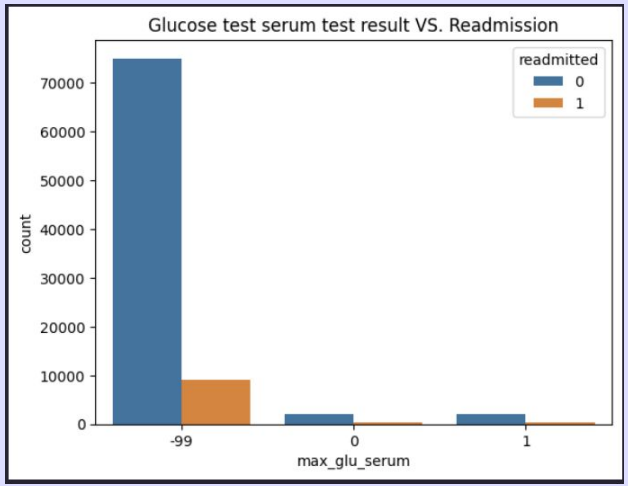
Graph 1: A1C Test Result vs. Readmission

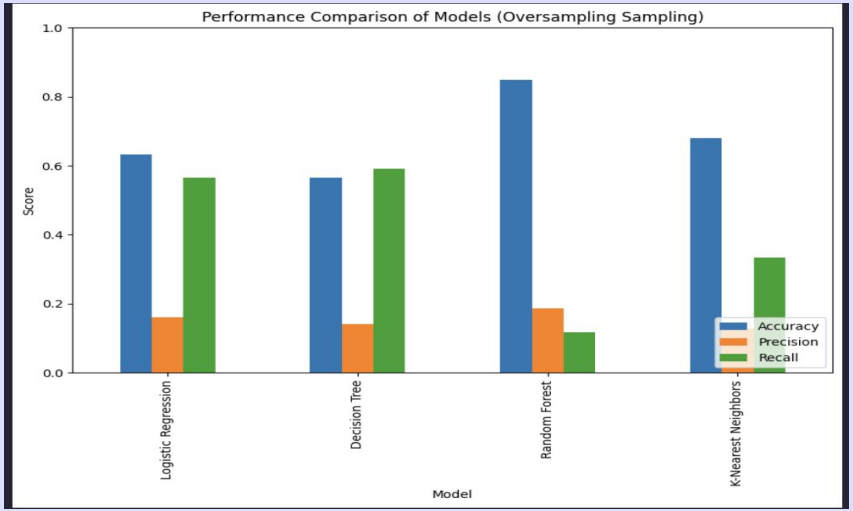
- **Depicts how A1C test outcomes** (test values of >7 and >8 are grouped into value 1) relate to readmission rates.
- **Shows small counts for "-99"** (no test), indicating many patients never had A1C measured.
- **Reflects possible missed opportunities** to assess and manage glycemic control before discharge, which might affect readmission risk.

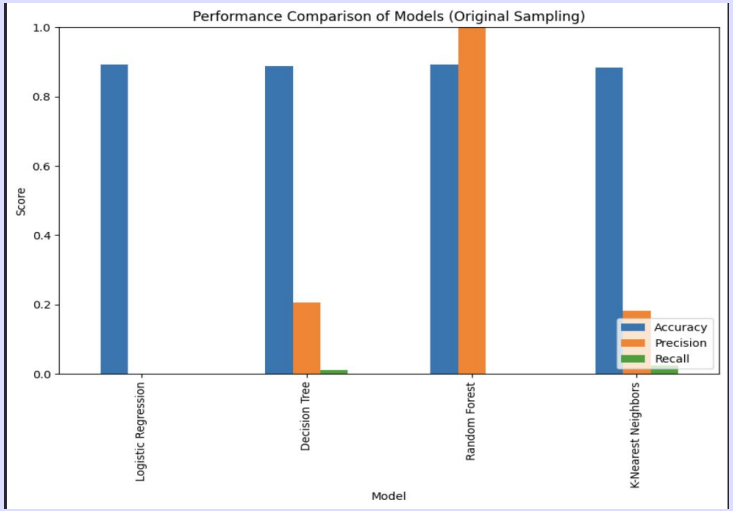


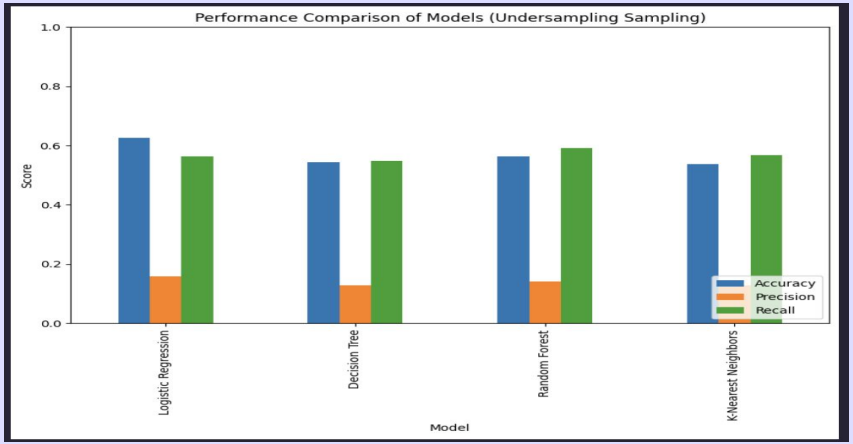
Graph 2: Glucose Serum Test Result vs. Readmission

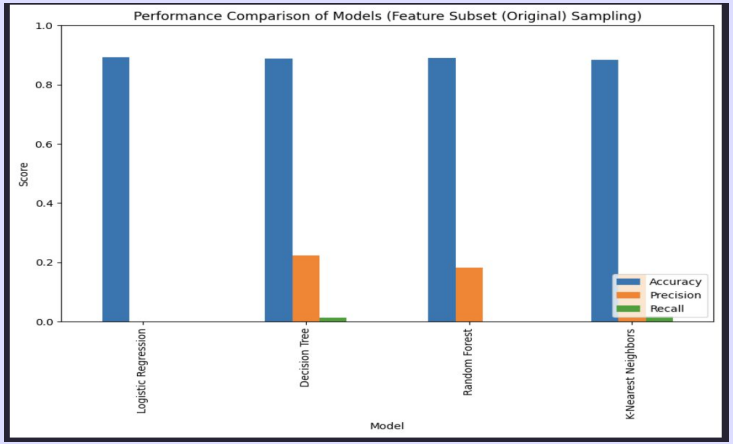
- **Visualizes results** of glucose serum tests: normal (0), high (1), or not taken (-99).
- **Similar story to A1C**: not testing (shown by -99) is common and may correlate with readmission outcomes.
- **Supports the notion** that more rigorous monitoring (testing) could impact early readmission rates.

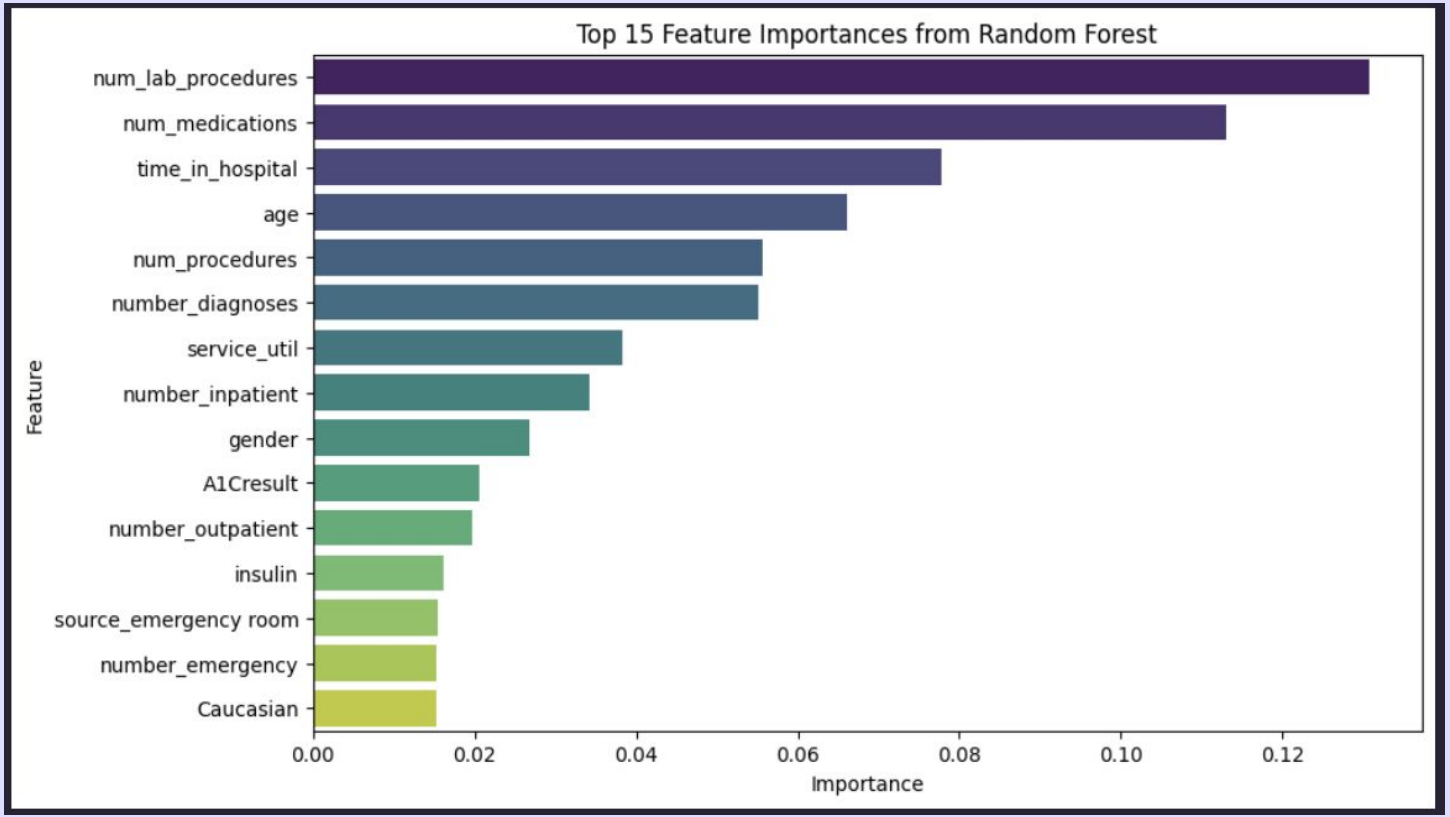



Performance Comparison of Models (Oversampling Sampling)
Score
Model
Logistic Regression
Decision Tree
Random Forest
K-Nearest Neighbors
Accuracy
Precision
Recall


Performance Comparison of Models (Original Sampling)
Score
Model
Logistic Regression
Decision Tree
Random Forest
K-Nearest Neighbors
Accuracy
Precision
Recall


Performance Comparison of Models (Undersampling Sampling)
Score
Model
Logistic Regression
Decision Tree
Random Forest
K-Nearest Neighbors
Accuracy
Precision
Recall


Performance Comparison of Models (Feature Subset (Original) Sampling)
Score
Model
Logistic Regression
Decision Tree
Random Forest
K-Nearest Neighbors
Accuracy
Precision
Recall


Top 15 Feature Importances from Random Forest
num_lab_procedures
num_medications
time_in_hospital
age
num_procedures
number_diagnoses
service_util
number_inpatient
gender
A1Cresult
number_outpatient
insulin
source_emergency room
number_emergency
Caucasian
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance

THANK YOU FOR LISTENING TO US