IT5006 Fundamentals of Data Analytics
Project Milestone 3

[Group 9]

Agnesh Kumar Rai
Aletheia Lai Xuanyu
Dan Yi Jia

https://github.com/rai-agnesh/Team9_IT5006_Healthcare_Analytics_AY2526

Table of Content

1. **Executive Summary**

This project addresses the critical challenge of predicting 30-day hospital readmissions among diabetic patients—a leading driver of healthcare cost and quality penalties. Using data from 130 U.S. hospitals (1999–2008) comprising over 100,000 encounters, the study aimed to develop and compare predictive models capable of identifying high-risk patients before readmission occurs. The goal was to design an implementable, interpretable, and ethically sound predictive framework that supports proactive care and operational efficiency in clinical settings.

A rigorous data-preprocessing pipeline was applied, encompassing missing-value imputation, low-variance feature removal, ICD-9 diagnostic recoding, and derived feature engineering using the Charlson Comorbidity Index (Quan variant), service utilization counts, and filtered medication changes. To prevent temporal leakage, the dataset was restricted to each patient's first encounter only, resulting in a refined cohort of 59,094 records and 121 encoded features. Given the inherent class imbalance (≈ 14.6 % readmitted), weighted learning was adopted instead of synthetic oversampling to maintain data fidelity.

Three models—Logistic Regression, Decision Tree, and Random Forest—were trained under a unified experimental architecture emphasizing recall-oriented F_2 optimization, complemented by ROC–AUC and Type I/II error trade-off analysis. Logistic Regression offered interpretability and strong calibration ($F_2 = 0.468$; ROC–AUC = 0.708) but underfit complex interactions. The pruned Decision Tree improved recall (0.708) at the expense of precision, while the optimized Random Forest achieved the most balanced generalization performance ($F_2 = 0.470$; ROC–AUC = 0.708; total error = 3,838), validating its robustness against overfitting. Ensemble learning through bagging and recursive feature elimination captured nonlinear dependencies while maintaining interpretability via SHAP analysis, which confirmed key determinants—number of prior inpatient visits, comorbidity burden, discharge disposition, and time in hospital—as clinically consistent predictors.

From a business and healthcare operations perspective, the model enables patient segmentation into proactive care tiers. High-risk patients (≥ 2 prior admissions or high comorbidity scores) warrant multidisciplinary discharge planning and close follow-up, while moderate-risk groups benefit from telehealth or medication-adherence programs. Low-risk discharges can be automated for standard outpatient review. Integrating such model-driven triage into Electronic Health Record (EHR) systems can reduce unplanned readmissions, optimize bed utilization, and generate annual cost savings estimated in the millions of dollars for large hospitals.

Overall, this study demonstrates that recall-optimized, interpretable machine learning can translate clinical data into actionable intelligence. The Random Forest model, supported by SHAP-based transparency, represents a viable step toward ethical and explainable AI in healthcare, aligning predictive analytics with patient-safety imperatives and hospital performance goals.

Contributions:

Agnesh – Random Forest Modelling, Exploratory Data Analysis, Alignment of F2-Scoring across 3 models based on chosen evaluation metric

Aletheia – Literature Review, Logistic Regression

Yi Jia – EDA Dashboard, Decision Tree

2. Introduction & Background

2.1 Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by persistent Hyperglycemia due to inadequate insulin production or impaired insulin utilization. Diabetes prevalence has increased significantly worldwide, with global estimates rising from 415 million in 2015 to 537 million in 2021 (Ogurtsova et al., 2020; Kumar et al., 2023). The International Diabetes Federation projects this number to reach 783 million by 2045 (Kumar et al., 2023). An absolute deficiency of insulin is the primary cause of type 1 diabetes. In contrast, Type 2 diabetes, the more prevalent form, arises mainly from insulin resistance often compounded by relative insulin deficiency (Krause & De Vito, 2019). Hyperglycemia leads to metabolic dysfunctions in proteins, fats, and carbohydrates. It can manifest in various ways (Banday et al., 2020). Chronic Hyperglycemia is consequently the primary cause of diabetes-associated morbidity and mortality with microvascular and macrovascular diabetic complications (Banday et al., 2020). These complications lead to hospital readmissions and an increased healthcare expenditure (Kukde et al., 2024).

Hospital readmission is often referred to as the re-hospitalization of a patient within a specific period after their discharge (Wang & Zhu, 2021). The annual expenditure of 30-day readmissions for diabetic patients is estimated to be between $20 billion and $25 billion (Rubin, 2018). Readmissions are often recognized as a quality indicator for healthcare systems. It not only reflects the severity of the disease but also the effectiveness of discharge planning and continuity of care. Therefore, there is a need to understand the risk factors for hospital readmission for patients with diabetes, as accurate prediction of readmission not only supports early intervention for high-risk individuals but also has the potential to yield substantial cost savings for healthcare systems.

2.2 Literature Review

Studies examining hospital readmissions among diabetic patients employ a diverse range of methodological approaches. Most studies employed a retrospective cohort design with large administrative databases from Medicare claims and electronic health records (Soh et al., 2020; Collins et al., 2017; Eby et al., 2015). 30-day readmission rates were commonly used as their primary outcome (Soh et al., 2020).

2.2.1 Statistical and Analytical Methods

Various statistical methods have been used to find predictors and estimate the risk of readmission. Discrimination statistics, commonly called the Area Under the Curve (AUC) and the concordance statistic (C-statistic), are frequently computed. Values above 0.8 indicate excellent discriminatory power, and values between 0.7 and 0.8 are considered acceptable. Additional measures of recall, precision, F1-Score, and specificity give a more comprehensive understanding of model performance, especially when dealing with issues of class imbalance.

Karunakaran et al. (2018) and Collins et al. (2017) achieved concordance statistics of 0.82, indicating excellent discriminative ability, meaning that in 82% of patient pairs, the model correctly ranked a readmitted patient as higher risk than a non-readmitted one. Eby et al. (2015) reported an AUC of 0.693, reflecting moderate performance. Although below the clinical benchmark of 0.8, the model remained informative for identifying high-risk cohorts. Other metrics frequently reported include accuracy, precision, recall (sensitivity), F1-score, and specificity, which provide a fuller picture of model behavior. For instance, Chong (2021) observed that Random Forest yielded a recall of 0.71, outperforming logistic regression (0.62), implying greater ability to correctly identify readmitted patients. However, this often came with slightly lower precision (0.65 vs 0.72), reflecting the typical trade-off between false positives and false negatives in healthcare prediction. The analytical advancements seen in studies related to diabetic readmission reflect the growing complexity of predictive modelling. Statistical methods, such as logistic regression, that were commonly used in past years provided interpretable results for clinicians. In comparison, recent machine learning techniques offer better predictive performance. However, this comes at the cost of model transparency.

2.2.2 Feature Engineering in Predictive Modelling

Moreover, feature engineering is an important aspect in improving the performance of prediction models. Various types of features are used forum widely at our disposal and many of these include comorbidity scores (e.g. Charlson or Elixhauser comorbidity scores), length of stay, time since last admission, number of laboratory tests performed on the day and number of medications prescribed (Karunakaran et al., 2018; Soh et al., 2020). Diagnosis based features usually require the clustering of ICD codes into meaningful groups that are clinically meaningful, such as diabetic complications and cardiovascular disease. Medications based features are more customary binary dependent variables (i.e. users of insulin or metformin). The relationship is mostly descriptive i.e. glucose/creatinine ratios give us physiological decompensation (Eby et al., 2015). Multicollinearity overcomes some of the opportunities available due to LASSO as well as principal component analysis (PCA), which again limits multicollinearity and over fitting. When this is done in the machine learning area, feature importance is usually done using a Gini impurity or SHAP value purity technique which can lend itself to increasing model's interpretability.

2.2.3 Comparative Analysis of Existing Models

The LACE Index comprising Length of stay, Acuity of admission, Comorbidity and Emergency department visit is one of the most widely utilized scores predicting hospital admission (Van Walraven et al., 2010). It's easy interpretability and use make it very popular; however, it suffers from a broad based and hence not specifically disease orientated approach, e.g. of Diabetes Mellitus. Data-based classification models on the other hand, such as the logistic regression models and random forest utilize a greater utility of demographic, clinical and laboratory variables and have more frequently improved AUC values (>0.75) (Sah Kanu & Khanal, 2023).

2.3 Interpretability in Healthcare Models

Clinicians struggle to adopt and trust black-box models when the rationale behind the predictions is unclear (Doshi-Velez & Kim, 2017). Regulations further reinforce the demand for transparency. The European Union's General Data Protection Regulation (GDPR) requires individuals to understand how algorithmic decisions are made, thereby increasing the demand for interpretability in healthcare models (Goodman & Flaxman, 2017). Responding to these concerns, researchers have explored various frameworks to allow increased transparency in complex models. ElShawi et al. (2020) offered a systematic evaluation of interpretability methods. Among the frameworks evaluated, MAPLE (Model Agnostic Supervised Local Explanations) achieved near-perfect reliability. However, its slow execution of 375 seconds per explanation could pose a significant barrier during clinical needs. Another promising framework, SHAP (Hapley Additive exPlanations), was able to produce explanations in 0.22 seconds per instance with a high similarity score. However, further validation is needed to ensure consistency.

2.4 Problem Statement

Hospital readmissions for diabetic patients not only impose substantial financial costs but also create significant operational strain on healthcare systems. Studies indicate that rehospitalization increases bed occupancy by approximately 10–15 %, thereby limiting capacity for new admissions and contributing to congestion within inpatient facilities (Rubin, 2018). This strain underscores the urgency for hospitals to adopt data-driven approaches that proactively identify high-risk patients before deterioration occurs.

Accordingly, this study aims to develop a predictive model to identify diabetic patients at risk of 30-day hospital readmission following their first recorded encounter. The 30-day horizon is relevant for its clinical actionability – providing a short but critical window in which targeted post-discharge interventions can prevent avoidable rehospitalizations. From an institutional standpoint, such optimization translates into measurable benefits; for example, a 500-bed hospital implementing an effective readmission-reduction strategy could save roughly USD 2 million annually through improved utilization and reduced penalty exposure (Rubin, 2018).

3. Methodology

3.1 Source of Data – UCI Machine Learning Repository – Diabetes Dataset

The Diabetes dataset based on patient data from 130 US Hospitals for the period of 1999-2008 was retrieved from an open-sourced UCI Machine Learning Repository and accessed via the fetch_ucirepo() method in the ucimlrepo library. The dataset contains patient records from 101766 unique encounters with the primary goal of inference of readmission risks (target variable = readmitted) based on patient demographics, admission details, glycemic control, medication usage and other hospital procedural metrics – totaling 49 variables. The detailed variables definition was obtained from the source (Strack, et al., 2014). An interactive exploratory data analytics dashboard can be accessed here: https://it5006-grp9.streamlit.app/

3.2 Data Preprocessing

Data preprocessing was conducted to ensure analytical validity and prevent temporal or structural bias before model training. The workflow addressed missingness and sparsity, dataset reframing and scope reduction, class imbalance, low-variance filtering and clinical recoding, and feature engineering with encoding standardization.

3.2.1 Handling Missing & Sparse Features

Exploratory data analysis revealed extensive feature-specific missingness across the original 49 attributes (Figure 1a–b, Appendix A). Variables such as *weight (96.8 %)*, *payer_code*, and *medical_specialty* were highly incomplete and subsequently removed. Correlated missingness was observed between *A1Cresult* and *max_glu_serum* due to mutually exclusive clinical use—A1C tests monitor chronic control, while glucose-serum values reflect acute episodes (Strack et al., 2014). Because these features are physiologically meaningful, nulls were imputed with the categorical label "Not measured" to retain record completeness. Pairwise nullity correlation study of the missingness matrix confirmed weak inter-feature dependence ($\tau < 0.4$), supporting selective exclusion.

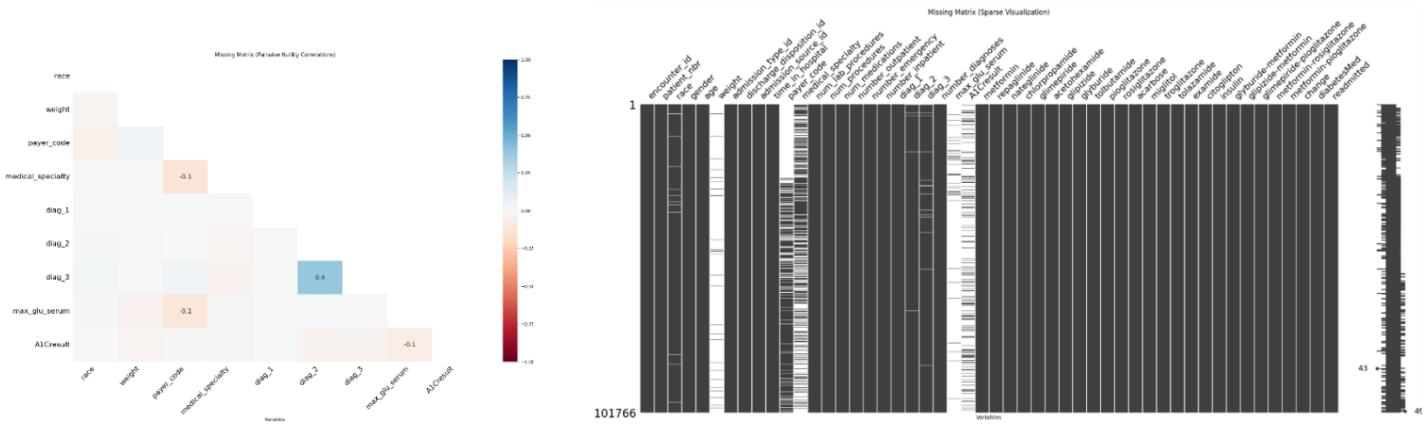


Figure 1. Pairwise Nullity Correlation Heatmap between Features with Missing Data (left) & Missingness Matrix (Right)

3.2.2 Patient Profile Analysis for Scope Reduction

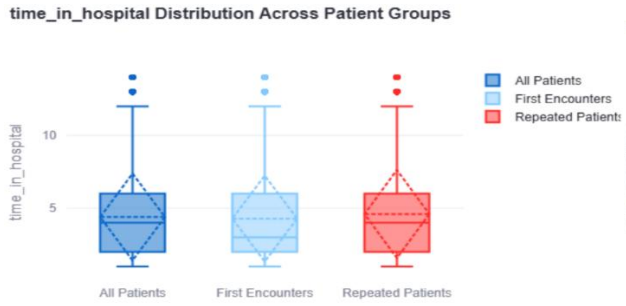


Figure 2 Distribution of time_in_hospital across patient groups

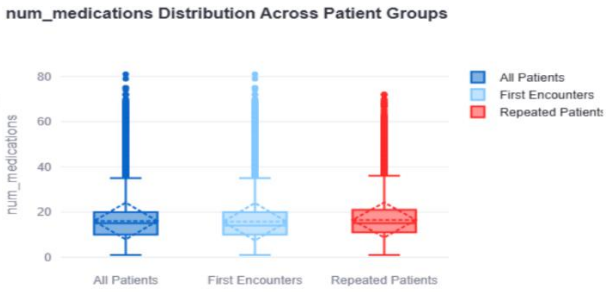


Figure 3. Distribution of num_medications across patient groups

A patient-profile analysis revealed distributional differences for key features between *repeated encounters* and *first encounters* (Figure 2 & 3) – longer mean average hospital days (4.6 vs 4.3 days) and greater clinical complexity. The objective was to detect potential temporal leakage—instances where later readmissions might inadvertently inform earlier cases within the same patient record, biasing model learning outcomes.

The Kruskal–Wallis and χ^2 tests confirmed statistically significant distributional differences ($p < 0.05$) across 31 of 34 features, validating that repeated-encounter records represented a systematically distinct population segment. Patients with repeated encounters demonstrated a substantially higher 30-day readmission risk of 19.5 %, compared to the baseline cohort (14.6%). Repeated patients also underwent more lab procedures and medication changes on average. These observed discrepancies indicated that the inclusion of multiple encounters per patient risked embedding outcome information from future hospitalizations, thereby violating temporal independence. To mitigate this, the dataset was restricted to each patient's first recorded encounter, ensuring that all predictors preceded the outcome event and were not confounded by subsequent care episodes. Additionally, encounters labeled for "> 30 days" readmission were excluded to maintain focus on the clinically actionable 30-day window against non-readmission cases, a standard quality-of-care metric endorsed in prior literature (Rubin, 2018).

3.2.3 Low-Variance Filtering & Clinical Recoding

Thirteen medication variables (e.g., *nateglinide*, *repaglinide, examide, clitoglipton*) exhibited >= 99 % uniformity and were dropped as near-zero-variance predictors. Administrative fields were consolidated into semantically coherent groups:

- Diagnosis Codes: ICD-9 codes were mapped into clinical categories (cardiovascular, endocrine, renal, etc.) using mappings (AAPC,2025).
- Admission and Discharge Source and Types: Rebinned into interpretable and clinically informed mappings from dataset retrieved from literature. (Clore, et al, 2014). Rare categories (< 1 %) were merged into "Other" to preserve statistical stability.

3.2.4 Feature Engineering

Three derived variables captured broader clinical complexity:

- A *Charlson Comorbidity Index (CCI) score* was computed for each encounter using ICD9 diagnosis codes (Quan variant) using an in-built comorbidipy python library. This index quantifies comorbid conditions across the primary, secondary and tertiary diagnoses that impact readmission risk based on patient health burden and has broad acceptance in clinical prediction studies, including diabetes readmission risk. (Rubin, et al, 2023)
- The number of inpatient, outpatient and emergency visits per patient were aggregated to create a new *service utilization* variable to factor care intensity.
- Of the filtered medications, dosage changes that indicated therapy adjustments (medications that were labelled "Up" or "Down") were tallied into a new *filtered medications count* variable. This was to factor treatment volatility.

3.2.5 Summary of Pre and Post Processing Transformations

Operation (Step)	**Example Variables - Previous**	**Operation Applied**	**Resulting Feature/Outcome(s)**
Missingness Handling	A1Cresult, max_glu_serum	Imputed "Not measured"	Retained categorical representation
Sparse Feature Removal	1. weight, payer_code, medical_specialty	1. Dropped	-

	2. Rare categories (<1%) in admission and discharge source and types	2. Merged into "Others – Low Incidence"	
Scope Reduction & Target Defined	patient_id, encounter_id	Filtered to first encounter per patient and dropped >30 readmitted class	59094 records (<30 vs Not Readmitted class)
Variance Filtering	nateglinide, repaglinide	Removed (>=99% homogenous)	13 medication variables removed
Diagnostic Re-binning	ICD-9 codes, Admission and Discharge IDs	Grouped by literature source to categorical features	9 diagnostic categories + clinically meaningful administrative mappings
Ordinal Encoding	age, medication, glycemic results (A1CResult, max_glu_serum),	Ordered to maintain monotonic relationship	Model learns gradient pattern
One-Hot Encoding	Diagnostic categories, admission and discharge IDs	One-hot encoded with drop-first	Prevents multicollinearity
Privacy & Ethics	patient_id, encounter_id	Dropped as not meaningful	Allows generalization and maintains anonymity
Feature Engineering	comorbidity_score, service_utilization, filtered_meds_count	-	Capture broader complexities
Predictors Count	49	-	32 active predictors (121 encoded features)

3.2.6 Class Imbalance

As the data was class-imbalanced, (~85% Not Readmitted against ~15% Early Readmission), class weights were managed as part of the modelling pipeline and was preferred over SMOTE as SMOTE introduces synthetic data points, which may not generalize well overtime, potentially affecting its adaptability. It was critical that the real minority class variance could be represented well to achieve an implementable model given the main objective to accurately detect early readmission risk.

3.3 Model Architecture & Implementation Details

We explored Logistic Regression, Decision Tree, and Random Forest sequentially to model 30-day hospital readmission risk among diabetic patients. Each model was selected to represent an incremental increase in non-linearity, flexibility, and interpretability trade-off: *Logistic Regression* established a transparent baseline with coefficient interpretability, whereas *Decision Tree* enabled nonlinear relationships and rule-based transparency. Lastly, *Random Forest* aggregated multiple decorrelated trees to improve generalization and reduce variance.

All models were trained, tuned, and evaluated under a unified experimental framework to ensure comparability. These included stratified train–test partitioning to preserve the 14.6 % minority class proportion as well as class imbalance handling via class_weight = "balanced" or {0:1, 1:6}, derived from empirical F_2 optimization. A five-fold stratified cross-validation was done during hyperparameter optimization for model robustness and variance assessment. A unified model-performance scoring metric was centred on F_2-score, complemented by F_1, recall, and ROC–AUC for benchmarking.

3.3.1 Choice of Evaluation Metric & Fine-Tuning

Given the clinical objective of minimizing *missed high-risk cases* (false negatives), the F_2-score was selected as the primary optimization metric. This metric emphasizes recall (sensitivity) weighting it twice as heavily as precision, making it appropriate for identifying at-risk patients even at the cost of additional false positives. However, to maintain a balanced perspective, F_1-score, recall, and ROC–AUC were also tracked: Type I (false positive) and Type II (false negative) error trade-offs were visualized using precision-recall curves and confusion matrix values to validate if model was over triggering false alarms at the expense of minority class identification. Thresholds were subsequently fine-tuned (0.30 – 0.80 range) to achieve an operationally meaningful decision

boundary based on 60% F_2-score and 40% type I + type II error count weightage (≈ 0.51 for Random Forest & 0.49 for Logistic Regression).

3.3.2 Unified Learning Architecture Summary

Aspect	Logistic Regression	Decision Tree	Random Forest
Complexity Progression	Low – interpretable linear relationships	Moderate – controlled via cost-complexity pruning	High – managed through bagging and depth constraints
Scaling	StandardScaler applied to numeric features	No scaling required	No scaling required
Class-imbalance	class_weight = custom ratio derived from tuning – {0:1,1:6}		
Hyperparameter Tuning	Grid Search (F2-scorer)	Grid Search (F2-scorer)	Halving Grid Search (F2-scorer)
Regularization/Pruning	L1 (LASSO) to enforce sparsity	Cost-complexity pruning	Implicit pruning through tree-level ccp_alpha inheritance
Evaluation	Unified metric framework (F2 + Type I/II Error Minimization)		
Feature Refinement	Recursive Feature Elimination		

Logistic Regression served as the interpretable baseline with L1 (LASSO) regularization for automatic feature selection (C=0.02, class_weight={0:1, 1:6}). With further recursive feature elimination, 85 features in an optimized and threshold tuned model achieved F_2-score of 0.468 ROC-AUC=0.708 and a 0.63 recall on the readmitted class on the test set.

Decision Tree provided rule-based interpretability with cost-complexity pruning to mitigate overfitting. Grid search optimization yielded a pruned tree (depth=7, ccp_alpha=0.000179), achieving F_2-score of 0.476, ROC-AUC=0.685 and a 0.708 recall on the readmitted class on the test set.

Random Forest was selected to capture non-linear interactions through ensemble learning. Hyperparameter tuning via HalvingGridSearchCV optimized the F_2-score with class_weight={0:1, 1:6} Recursive Feature Elimination (RFE) reduced dimensionality to 91 features. The optimized and threshold tuned model (400 estimators, max_depth=9) achieved F_2-score of 0.470, ROC-AUC=0.706 and a 0.605 recall on the readmitted class on the test set.

Detailed implementation specifications, hyperparameter configurations, and training procedures are provided in Appendix B: Model Implementation Details.

4. Results & Analysis

4.1 Experimental Results & Performance Metrics

Metric	Logistic Regression (thr = 0.49)	Decision Tree (thr = 0.50)	Random Forest (thr = 0.51)
Train Set	F2:0.475 ROC_AUC:0.708 Accuracy:0.672	F2:0.500 ROC_AUC:0.695 Accuracy:0.581	F2:0.492 ROC_AUC:0.729 Accuracy:0.686
Test Set	F2:0.468 ROC-AUC:0.708 Accuracy:0.654	F2:0.482 ROC-AUC: 0.685 Accuracy: 0.572	F2:0.470 ROC-AUC:0.708 Accuracy:0.675

Minority Class – Readmitted	Precision:0.239 Recall:0.631 F1: 0.347	Precision: 0.211 Recall: 0.708 F1: 0.326	Precision:0.249 Recall:0.605 F1:0.352
Majority Class – Not Readmitted	Precision: 0.912 Recall: 0.658 F1: 0.764	Precision: 0.917 Recall: 0.549 F1: 0.687	Precision:0.911 Recall:0.687 F1:0.783
Confusion Matrix	[[6637,3457],[637,1088]]	[[5540,4554],[504,1221]]	[[6937,3157],[681,1044]]

4.1.1 Logistic Regression

$$\log\left(\frac{p}{1-p}\right) = -0.22 + 0.36(number_{inpatient}) + 0.15(comorbidity_{score}) + 0.15(rehab_{transfer}) + 0.13(number_{emergency}) + 0.10(number_{diagnoses}) + 0.06(age) + 0.05\left(time_{in_{hospital}}\right) + 0.05\left(num_{lab_{procedures}}\right) + \cdots - 0.17(discharged_{home}) - 0.16(hospice_{home}) - 0.12(hospice_{medical}) - 0.09\left(transfer_{from_{hospital}}\right) - 0.08(metformin) - 0.05(num_{procedures}) - 0.03(insulin) - 0.03(respiratory_{diagnosis}) - \cdots$$

where p = *probability of patient readmission within* 30 *days after discharge.*

The L1-regularized Logistic Regression established a strong linear baseline with good calibration and interpretability. The model achieved stable performance across folds, demonstrating good generalization and low variance between train and test scores. However, as expected from its linear structure, the model underfit certain nonlinear interactions (e.g., multi-diagnostic dependencies). Feature sparsity from the L1 penalty enhanced interpretability by isolating key predictors such as number of prior inpatient visits, comorbidity score, age, and time in hospital, which positively associated with readmission risk—reflecting cumulative disease burden and clinical complexity. In contrast, insulin, metformin, and discharge to home exhibited negative coefficients, suggesting that effective diabetes management and stable discharge planning mitigate early readmission likelihood. The confusion matrix [[6637,3457], [637,1088]] revealed that false positives dominated the misclassifications, indicating a conservative bias that limited over-triggering at the expense of recall. Nevertheless, the 637 missed readmissions (false negatives) remain clinically significant. Precision for the minority class was 0.239 and recall 0.631, reflecting moderate sensitivity and good generalization. Compared with Random Forest, Logistic Regression produced lower Type II error (637) capturing more readmitted class, however at the expense of more majority class misclassifications.

4.1.2 Decision Tree

The performance of the pruned Decision Tree classifier (threshold = 0.50) was evaluated on the hold-out test set using standard classification metrics. The model achieved an overall accuracy of 0.572, ROC–AUC of 0.685, and F_2-score of 0.482, indicating a fair ability to discriminate between readmitted and non-readmitted patients. For the minority (readmitted) class, precision = 0.211 and recall = 0.708, showing strong sensitivity at the expense of higher false positives. This behavior aligns with the F_2-optimized objective, which prioritizes recall (Type II error reduction) over precision. The corresponding confusion matrix [[5540,4554], [504,1221]]confirmed this trade-off — 504 false negatives (missed readmissions) compared to 4554 false positives, the largest among the three models. While this led to a higher overall error count, it successfully captured a greater proportion of actual readmissions, an important clinical advantage for proactive patient monitoring.

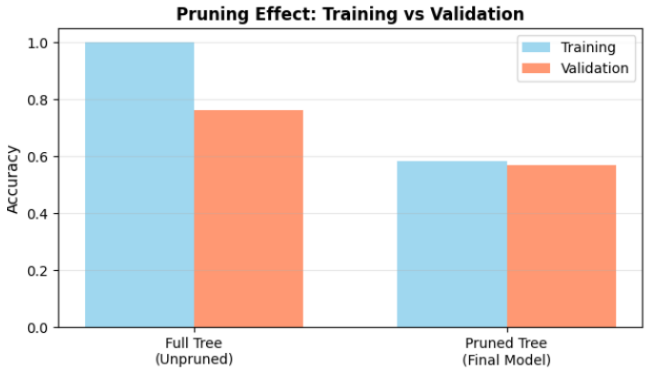


Figure 4. Pruning Effect Plot - DT

The initial unpruned tree achieved perfect training accuracy (1.000), evidencing overfitting. However, cost-complexity pruning (ccp_alpha = 0.00018) and subsequent grid search optimization over max_depth, min_samples_split, and min_samples_leaf reduced the model depth to 9 and leaves to 27, improving generalization with a validation ROC–AUC of 0.66. The pruning effect plot (Figure 4) shows a clear reduction in the training–validation performance gap, confirming overfitting mitigation.

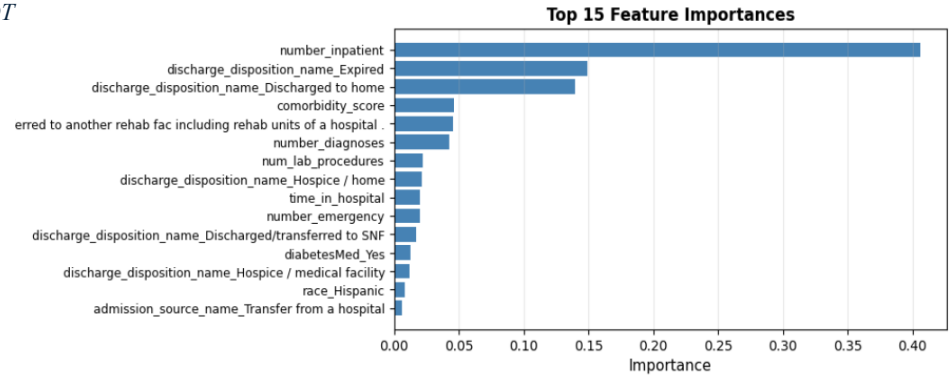


Figure 5 – Top 15 Feature Importances

The top predictors (Figure 5) emphasized both discharge outcomes and inpatient utilization patterns as critical determinants of readmission risk. Number_inpatient (0.406) and discharge_disposition_name_Expired (0.149) were the strongest predictors, indicating patients with prior inpatient episodes or died were far more likely to influence readmission direction – positively in the former and negate the outcome in the latter. Other significant contributors included discharge to home (0.139), comorbidity_score (0.046), rehabilitation transfers (0.045), number_diagnoses (0.042), and num_lab_procedures (0.022) — collectively reflecting the compounded effect of medical complexity and post-hospitalization care outcomes. Overall, the Decision Tree model improved F2 and recall performance relative to Logistic Regression but sacrificed precision and overall accuracy – as seen in its highest Type I + Type II errors.

4.1.3 Random Forest

The threshold-optimized Random Forest model (threshold = 0.51) achieved an overall $F_2 = 0.470$, ROC–AUC = 0.708, and accuracy = 0.675 on the held-out test set, indicating fair discriminative capability and the best generalization performance among the three algorithms.For the minority class (readmitted), precision = 0.249 and recall = 0.605, showing a balanced trade-off between sensitivity and false-alarm control. The corresponding confusion matrix [[6937,3157], [681,1044]] revealed 681 false negatives (missed readmissions) and 3157 false positives, marking the lowest total error count (3838) of all models. Relative to Logistic Regression ($F_2 = 0.468$) and Decision Tree ($F_2 = 0.482$), the Random Forest maintained comparable recall yet improved overall precision and stability—reflecting reduced variance through ensemble bagging. Its training metrics ($F_2 = 0.492$, ROC–AUC = 0.729, accuracy = 0.686) confirmed moderate overfitting that was effectively regularized through cross-validated hyperparameter optimization (via HalvingRandomSearchCV) on key parameters such as max_depth, min_samples_leaf, and n_estimators, guided by the F_2-score. Collectively, the ensemble structure produced the most balanced Type I and Type II error profile across models.

5. Model Interpretation and Business Insights

5.1 Overview

Across all three models – Logistic Regression, Decision Tree, and Random Forest, the central objective was to predict early (<30 days) readmission risk among diabetic patients - post-discharge. Despite architectural differences, a consistent feature class emerged as highly predictive across models: Hospital utilization (e.g., number_inpatient, service utilization), discharge disposition, and comorbidity burden with chronic complexity (e.g. num_diagnoses, filtered_med_counts). This convergence reinforces the clinical validity of these determinants and demonstrates that the patterns learned by the models align with known medical reasoning. While

ROC–AUC served as a benchmark for discrimination, its insensitivity to class imbalance necessitated using the F_2-score for final optimization. The F_2 metric ($\beta = 2$) prioritizes recall — penalizing false negatives more than false positives — a clinically grounded decision since missing a high-risk patient poses greater harm than issuing an unnecessary alert. This aligns with best practice in predictive healthcare research (Hu et al., 2021) and guided the selection of the Random Forest as the final model, given its superior recall–specificity balance and minimized missed-readmission risk.

5.2 Comparative Model Interpretation and Clinical Insights

The progression in model selection from Logistic Regression to Decision Tree and finally Random Forest, was an intended movement along a continuum from interpretability to handling complexity in feature space to generalization. There was a contribution of distinct and key analytical strengths from each model selected for the research question, as all of them helped reinforce the clinical validity of important predictors of early (<30 days) diabetic readmission risk.

In the logistic regression approach, number_inpatient, time_in_hospital, number_emergency, and comorbidity score were associated with increased readmission risk, while metformin use and home discharge reduced it. These relationships align with prior clinical evidence: frequent hospital and emergency visits indicate physiological instability (Richstein et al., 2024), while consistent metformin therapy suggests improved glycemic control and stability (Wong et al., 2024). However, due to its linear assumption, Logistic Regression could not capture interaction effects and was moderately sensitive – leading to higher misclassifications in the majority class.

The pruned Decision Tree improved minority-class sensitivity by explicitly modeling non-linear and rule-based patterns. The top predictors included number_inpatient, discharge_disposition_name_Expired, and discharge_disposition_name_Discharged to home, confirming that inpatient frequency and discharge pathway are critical determinants of readmission. Secondary influences — such as comorbidity_score, number_diagnoses, and time_in_hospital — contributed to decision splits but with lower hierarchical importance. From an operational perspective, these results imply that patients with repeated admissions or complex discharge statuses (e.g., transfers to rehabilitation or skilled nursing facilities) face significantly higher early readmission risk. In contrast, patients discharged home or with shorter hospital stays typically demonstrated clinical stability.

This suggests that hospitals could benefit from implementing targeted discharge planning and post-hospitalization monitoring for such patients. Interventions such as follow-up calls, telemedicine check-ins, and diabetes medication adherence programs may help reduce the likelihood of readmission. Furthermore, the association between medication complexity and readmission emphasizes the importance of medication reconciliation and patient education. Simplifying treatment regimens where clinically appropriate and ensuring patients understand their prescriptions could mitigate unplanned returns to the hospital.

While highly interpretable, the single-tree model remained variance-sensitive and somewhat overfitted, necessitating ensemble stabilization through Random Forest integration.

The Random Forest achieved the strongest balance of F2-score, recall, and overall discriminative power It had the lowest total error count (3,838) and the most even distribution between false positives and false negatives, demonstrating effective generalization without significant recall sacrifice. The interpretability of the Random Forest optimized model was enhanced through SHAP analysis with a summary plot reproduced in Figure 6.

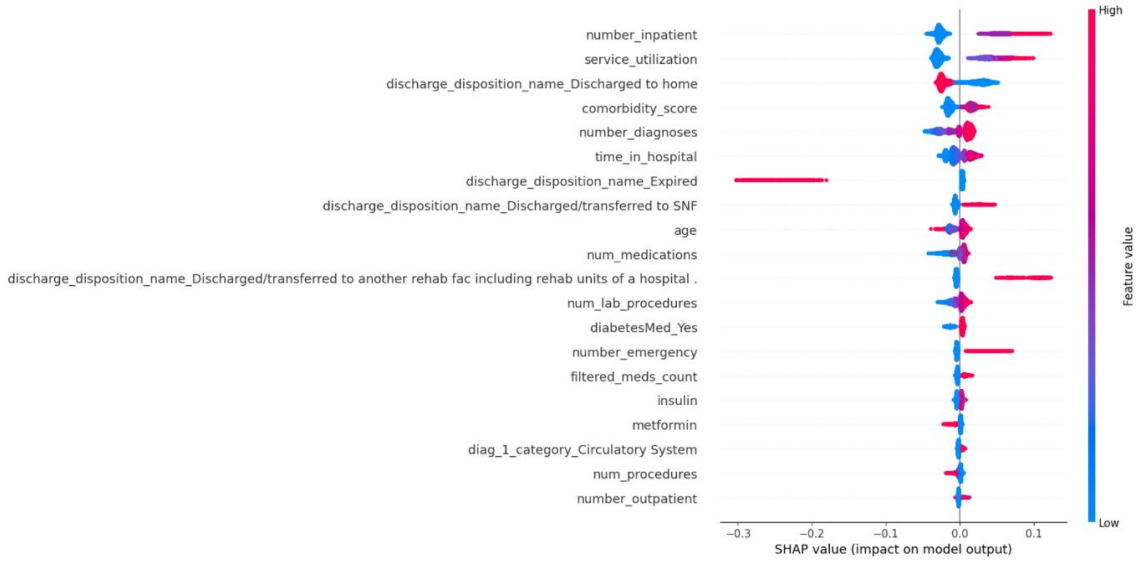


Figure 6. SHAP Summary Plot for Optimized RandomForest Model on Diabetes Early Readmission Risk for First-time Discharged Patients

The SHAP analysis (Figure 6) highlighted stable, interpretable feature attributions. Firstly, number_inpatient and service_utilization exerted the highest positive SHAP impacts, confirming that repeated hospitalizations strongly increase predicted risk. Secondly, comorbidity_score, number_diagnoses, and time_in_hospital had cumulative additive effects on predicted probability. Also, discharge_disposition_name_Discharged to home and metformin presented negative SHAP impacts, protective against readmission. Interestingly, diabetesMed = Yes and insulin presented a heterogeneous SHAP pattern, suggesting that while prescribed medication suggests proper disease management is in place, it may also capture treatment escalation for poorly controlled cases—hence still signalling risk. This interpretability confirmed the clinical consistency of the Random Forest ensemble while offering the most balanced precision–recall trade-off among models. Its recall-focused training strategy (via F_2 optimization) ethically prioritized early identification of high-risk patients — a desirable outcome in clinical risk management, where false negatives equate to potential care failures.

This reinforces the importance of integrated care plans targeting multi-morbid diabetic patients, especially those with cardiovascular or renal complications. This shows that the model outputs may need to be contextualised within patient-specific trajectories rather than interpreted deterministically.

5.2.1 Clinical Recommendation

From a business standpoint, these insights enable hospitals to segment patients into proactive care tiers that inform targeted intervention workflows. High-risk profiles – patients with ≥2 prior inpatient admissions or high comorbidity scores – should automatically trigger referral to a discharge nurse or social worker before release, ensuring continuity of care and medication compliance through multidisciplinary review. Moderate-risk profiles benefit from structured telehealth follow-ups or digital adherence programs within seven days post-discharge, capturing early warning signs and reinforcing recovery. Low-risk discharges may be safely routed to standard outpatient review cycles through automated scheduling systems. Integrating such model-driven triggers into hospital electronic health record (EHR) systems can operationalize predictive insights into actionable workflows, reducing unplanned readmissions and optimizing resource allocation.

Four random patients (Patient 5, 12, 24, 42) were studied using SHAP waterfall plots as reproduced in Appendix C – Modelling – RandomForest Figures 3-6.

Patient 5 – True Positive [Readmitted] – Predicted Probability Score of 0.714

Patient classification for readmission was strongly influenced by a high number_inpatient (+0.09), service_utilization (+0.06) and elevated comorbidity_score (+0.02), indicating high utilization and complex disease burden. While discharge_disposition_name = Home exerted a small negative contribution (–0.01), it was insufficient to offset cumulative positive risk factors. This correctly identified high-risk patient aligns with clinical expectations for multi-admission diabetic cases. This also illustrated that whilst patient was discharged to home, patient-specific trajectories through other predictors accurately identified the risk of readmission.

Patient 12 – False Positive [Readmitted] – Predicted Probability Score of 0.656

The model overestimated readmission risk primarily due to high number_inpatient (+0.1) and service_utilization (+0.08), comorbidity_score (+0.01), despite stabilizing signals from number_diagnoses (–0.03) and discharge to home (–0.01). This reflects an ethically important trade-off that a false alarm may trigger unnecessary follow-up but this errs on the side of patient safety, consistent with the F_2-optimized recall-oriented approach that was undertaken in fitting the model.

Patient 24 & 42 – True Negative [Readmitted] – Predicted Probability Score of 0.472, 0.319

Both patient 24 and 42, had lower predicted risk that was contributed by dampening effects from discharge_to_home, number_inpatient, number_diagnoses suggesting effective care transition and lower patient risk profile. In the former, minor positive pushes in risk prediction came from number_emergency and num_lab_procedures though these did not surpass the stability observed by discharge_to_home and likely low number_inpatient counts.

5.3 Limitations & Future Work

5.3.1 Data Imbalance and Generalization

Despite class-weight adjustments that were performed, the dataset's inherent imbalance (fewer readmissions) constrained the minority class precision. SMOTE was not implemented primarily due to the strong concerns on synthetic data introduction. Given the low precision, the model may not be able to be generalized to other hospital populations unless retrained with representative cohorts.

5.3.2 Feature Encoding and Electronic Records Dependence

The model depends on structured electronic health record variables, as well as unstructured notes in certain admission and discharge disposition variables. The data lacked potentially important socioeconomic context (other than payer_code which was largely missing) as well as other behavioral determinants (excluding weight which was significantly missing) of health were not part of the dataset. These could serve as significant predictors of readmission risk and could add value in increasing model performance towards the minority class.

5.4 Conclusion

This study developed and compared three predictive models—Logistic Regression, Decision Tree, and Random Forest—to identify early 30-day readmission risk among diabetic patients. Through F_2-optimized training emphasizing recall, the Random Forest achieved the most balanced performance and clinically meaningful sensitivity, reducing missed high-risk cases while maintaining reasonable precision. Consistent predictors such as inpatient frequency, comorbidity burden, and discharge disposition confirmed medical validity. Integrating these predictive insights into hospital workflows can enable proactive, data-driven care coordination—targeting high-risk patients for intervention while optimizing healthcare resource allocation and supporting the broader transition toward explainable and ethical AI in clinical decision-making.

6. References

AAPC. (2025). ICD-9 Codes Lookup - ICD-9 Codes List. Codify by AAPC. Retrieved September 14, 2025, from https://www.aapc.com/codes/icd9-codes-range

Aoyama, K., Pinto, R., Ray, J. G., Hill, A., Scales, D. C., & Fowler, R. A. (2020). Determining associations and estimating effects with regression models in clinical anesthesia. *Anesthesiology*, *133*(3), 500–509. https://doi.org/10.1097/ALN.0000000000003425

Banday, M. Z., Sameer, A. S., & Nissar, S. (2020). Pathophysiology of diabetes: An overview. *Avicenna Journal of Medicine*, *10*(4), 174–188. https://doi.org/10.4103/ajm.ajm_53_20

Bergenstal, R. M., Fahrbach, J. L., Iorga, Ş. R., Fan, Y., & Foster, S. A. (2012). Preadmission glycemic control and changes to diabetes mellitus treatment regimen after hospitalization. Endocrine Practice, 18(3), 371-375.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chong, C. (n.d.). *Predicting hospital readmission of diabetics*. Retrieved September 11, 2025, from https://kaggle.com/code/chongchong33/predicting-hospital-readmission-of-diabetics

Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J.

Collins, J., Abbass, I. M., Harvey, R., Suehs, B., Uribe, C., Bouchard, J., Prewitt, T., DeLuzio, T., & Allen, E. (2017). Predictors of all-cause 30 day readmission among Medicare patients with type 2 diabetes. *Current Medical Research and Opinion*, *33*(8), 1517–1523. https://doi.org/10.1080/03007995.2017.1330258

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. https://doi.org/10.48550/ARXIV.1702.08608

Eby, E., Hardwick, C., Yu, M., Gelwicks, S., Deschamps, K., Xie, J., & George, T. (2015). Predictors of 30 day hospital readmission in patients with type 2 diabetes: A retrospective, case-control, database study. *Current Medical Research and Opinion*, *31*(1), 107–114. https://doi.org/10.1185/03007995.2014.981632

ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650. https://doi.org/10.1111/coin.12410

Fife, D. A., & D'Onofrio, J. (2022). Common, uncommon, and novel applications of random forest in psychological research. *Behavior Research Methods*, *55*(5), 2447–2466. https://doi.org/10.3758/s13428-022-01901-9

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. https://doi.org/10.1002/bimj.201700067

Hu, Y., Chen, R., Gao, H., Lin, H., Wang, J., Wang, X., Liu, J., & Zeng, Y. (2021). Explainable machine learning model for predicting spontaneous bacterial peritonitis in cirrhotic patients with ascites. Scientific Reports, 11, Article 21639. https://doi.org/10.1038/s41598-021-00218-5

Karunakaran, A., Zhao, H., & Rubin, D. J. (2018). Predischarge and postdischarge risk factors for hospital readmission among patients with diabetes. *Medical Care*, *56*(7), 634–642. https://doi.org/10.1097/MLR.0000000000000931

Krause, M., & De Vito, G. (2023). Type 1 and type 2 diabetes mellitus: Commonalities, differences and the importance of exercise and nutrition. *Nutrients*, *15*(19), 4279. https://doi.org/10.3390/nu15194279

Kukde, R. D., Chakraborty, A., Shah, J., Kukde, R., Chakraborty, A., & Shah, J. (2024). A systematic review of recent studies on hospital readmissions of patients with diabetes. *Cureus*, *16*(8). https://doi.org/10.7759/cureus.67513

Kumar, A., Gangwar, R., Zargar, A. A., Kumar, R., & Sharma, A. (2024). Prevalence of diabetes in india: A review of idf diabetes atlas 10th edition. *Current Diabetes Reviews*, *20*(1), e130423215752. https://doi.org/10.2174/1573399819666230413094200

Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J. E., & Makaroff, L. E. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, *128*, 40–50. https://doi.org/10.1016/j.diabres.2017.03.024

Richstein, R. M., Gordon, C., Gozar, M., Ohanesian, L., Fishbein, J., Gottlieb, D. E., Silverman, R. A., & Schulman-Rosenbaum, R. C. (2024). Evaluating Hospital Revisit Risk in Patients Discharged from the Emergency Department with Blood Glucose of 300 mg/dL (16.7 mmol/L) or Greater. Diabetology, 5(7), 656-666. https://doi.org/10.3390/diabetology5070048

Rubin, D. J. (2018). Correction to: Hospital readmission of patients with diabetes. *Current Diabetes Reports*, *18*(4), 21. https://doi.org/10.1007/s11892-018-0989-1

Rubin, D. J., Maliakkal, N., Zhao, H., & Miller, E. E. (2023). Hospital Readmission Risk and Risk Factors of People with a Primary or Secondary Discharge Diagnosis of Diabetes. Journal of Clinical Medicine, 12(4), 1274. https://doi.org/10.3390/jcm12041274

Sah Kanu, D. K., & Khanal, M. (2023). *Implementation of Big Data Analytics on Diabetes 130-US Hospitals for year 1999-2008 for predicting patient readmission*. https://doi.org/10.13140/RG.2.2.18564.30081

Soh, J. G. S., Wong, W. P., Mukhopadhyay, A., Quek, S. C., & Tai, B. C. (2020). Predictors of 30-day unplanned hospital readmission among adult patients with diabetes mellitus: A systematic review with meta-analysis. *BMJ Open Diabetes Research & Care*, *8*(1), e001227. https://doi.org/10.1136/bmjdrc-2020-001227

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, 781670. https://doi.org/10.1155/2014/781670

Wang, S., & Zhu, X. (2022). Predictive modeling of hospital readmission: Challenges and solutions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(5), 2975–2995. https://doi.org/10.1109/TCBB.2021.3089682

Wong, C., Junqueira, E., Poldiak, N. P., Crossley, N., & Jenkins, S. (2024). Influence of Metformin Discontinuation on Readmission Rate in Patients With Acute Heart Failure. Journal of community hospital internal medicine perspectives, 14(4), 12–17. https://doi.org/10.55729/2000-9666.1366

7. Appendix A (Exploratory Data Analysis)

	Feature	Test	Statistic	p-value	Significant	Interpretation
0	race	Chi-square	189.2411	1.18E-36	TRUE	Groups differ
1	gender	Chi-square	42.81208	1.13E-08	TRUE	Groups differ
2	age	Chi-square	234.4436	1.17E-39	TRUE	Groups differ
3	time_in_hospital	Kruskal-Wallis	386.5661	1.14E-84	TRUE	Groups differ
4	payer_code	Chi-square	701.2774	6.24E-127	TRUE	Groups differ
5	medical_specialty	Chi-square	1911.996	2.52E-307	TRUE	Groups differ
6	num_lab_procedures	Kruskal-Wallis	9.888558	0.00712405	TRUE	Groups differ
7	num_procedures	Kruskal-Wallis	519.0069	1.99E-113	TRUE	Groups differ
8	num_medications	Kruskal-Wallis	539.9506	5.64E-118	TRUE	Groups differ
9	number_outpatient	Kruskal-Wallis	1397.752	3.03E-304	TRUE	Groups differ
10	number_emergency	Kruskal-Wallis	2639.159	0	TRUE	Groups differ
11	number_inpatient	Kruskal-Wallis	30032.53	0	TRUE	Groups differ
12	number_diagnoses	Kruskal-Wallis	1594.877	0	TRUE	Groups differ
13	max_glu_serum	Chi-square	19.64572	0.000586582	TRUE	Groups differ
14	A1Cresult	Chi-square	1.422662	0.840246797	FALSE	No difference
15	metformin	Chi-square	193.2456	5.19E-39	TRUE	Groups differ
16	repaglinide	Chi-square	70.79011	2.82E-13	TRUE	Groups differ
17	glimepiride	Chi-square	6.601817	0.359244076	FALSE	No difference
18	glipizide	Chi-square	15.72165	0.015328552	TRUE	Groups differ
19	glyburide	Chi-square	26.73638	0.000162234	TRUE	Groups differ
20	pioglitazone	Chi-square	10.13875	0.118929439	FALSE	No difference
21	rosiglitazone	Chi-square	19.2355	0.003783849	TRUE	Groups differ
22	insulin	Chi-square	972.7084	7.14E-207	TRUE	Groups differ
23	change	Chi-square	208.4684	5.39E-46	TRUE	Groups differ
24	diabetesMed	Chi-square	198.0437	9.89E-44	TRUE	Groups differ
25	discharge_disposition_name	Chi-square	923.2768	2.52E-162	TRUE	Groups differ
26	admission_type_name	Chi-square	442.7289	3.30E-87	TRUE	Groups differ
27	admission_source_name	Chi-square	981.2655	4.61E-187	TRUE	Groups differ
28	diag_1_category	Chi-square	918.6441	1.74E-169	TRUE	Groups differ
29	diag_2_category	Chi-square	520.2487	3.68E-87	TRUE	Groups differ
30	diag_3_category	Chi-square	630.6757	1.01E-109	TRUE	Groups differ

Appendix A - EDA - Table 3. Summary of Test results for feature distribution difference across patient groups

8. Appendix B - Model Implementation Details

8.1 Logistic Regression

8.1.1 Model Overview

Logistic Regression served as the baseline classifier for its interpretability, computational efficiency, and role in establishing a linear benchmark for comparison. It models the log-odds of early readmission as a linear function of predictors, providing coefficient-based interpretability that directly quantifies each variable's directional influence on the target outcome. While limited by its linear assumption and inability to capture complex feature interactions, Logistic Regression offered transparency essential to clinical deployment and model explainability.

8.1.2 Hyperparameters Set-Up

Hyperparameter optimization was performed using GridSearchCV with five-fold stratified cross-validation, maximizing the F_2-score (recall-weighted objective) and benchmarking against ROC–AUC. The tuned configuration yielded: Penalty: L1 (LASSO), C (Inverse Regularization Strength): 0.02, Solver: liblinear, Class Weight: {0:1, 1:6}, max_iter: 1000. The L1 penalty induced sparsity in the coefficient space, effectively performing embedded feature selection by shrinking uninformative weights to zero. Threshold calibration was performed between 0.30 and 0.80, with 0.49 chosen to maximize a combined objective of F_2 (60%) and minimized Type I + Type II errors (40%).

8.1.3 Implementation Details

Continuous variables were standardized using z-score normalization, and class imbalance was handled via weighted learning instead of synthetic oversampling. Post-model fitting, Recursive Feature Elimination (RFE) with cross-validation refined the predictors to 85 active features. The final L1-regularized Logistic Regression (RFE + threshold tuned) achieved an $F_2 = 0.468$, ROC–AUC = 0.708, and Recall = 0.631 on the test set. Key predictors with the largest positive coefficients included: number_inpatient, comorbidity_score, number_emergency, number_diagnoses, age, time_in_hospital. Negative predictors included: metformin, insulin, discharge_to_home, hospice_home, and num_procedures. This model provided strong interpretability and stable generalization but underfit higher-order nonlinear relationships present in the dataset.

8.2 Decision Tree

8.2.1 Model Overview

The Decision Tree (DT) classifier was developed to extend interpretability while capturing nonlinear interactions missed by Logistic Regression. By recursively partitioning the data using the Gini impurity criterion, it generated transparent, rule-based decision logic—ideal for clinical validation. The target imbalance (14.6% readmitted) was handled using class_weight = 'balanced'.

8.2.2 Hyperparameters Set-Up

The model was first grown without constraints (Depth = 60), demonstrating severe overfitting. Cost-Complexity Pruning (CCP) analysis was then performed, evaluating 50 sampled ccp_alpha values, identifying $\alpha \approx 0.00018$ as optimal. Subsequently, GridSearchCV with five-fold cross-validation optimized: max_depth: [5, 7, 9, 10]**,** min_samples_split: [5, 8, 13]**,** min_samples_leaf: [5, 9]**,** criterion: ['gini']**,** ccp_alpha: [0, α/2, α, 2α]**.** The best model was Depth = 9, ccp_alpha = 0.000179, and class_weight = 'balanced', achieving test $F_2 = 0.482$, ROC–AUC = 0.685, and Recall = 0.708.

8.2.3 Implementation Details

The final model implemented was the best_estimator_ identified by the GridSearchCV. This pruned Decision Tree resulted in a final depth of 9 with 27 leaf nodes. Feature importance scores (Gini importance) indicated that number_inpatient, discharge_disposition_name_Expired, and discharge_disposition_name_Discharged to home

were the most influential predictors. Clinically, this highlights prior hospital utilization and specific discharge circumstances (like death or discharge home) as key factors learned by the model for predicting readmission risk.

8.3 Random Forest

The Random Forest (RF) ensemble was implemented to overcome single-tree variance through bootstrap aggregation (bagging), improving generalization and stability. Each tree trained on random subsets of data and features, producing decorrelated predictions aggregated via majority voting. This design balanced bias–variance trade-offs while enabling feature interaction discovery, making it highly effective for heterogeneous healthcare data.

8.3.1 Hyperparameter Optimization

Optimization was conducted using HalvingRandomSearchCV with five-fold stratified CV and F_2-score as the primary objective. The parameter space included: {n_estimators: 320–400, max_depth: 6–10max_features: (0.15–0.4 fraction), min_samples_split: 8–13, min_samples_leaf: 5–9, class_weight: [{0:1, 1:4}, {0:1, 1:5}, {0:1, 1:6}] The best configuration achieved: n_estimators = 400, max_depth = 9, max_features ≈ 0.27, min_samples_split = 5, min_samples_leaf = 9, and class_weight = {0:1, 1:6}. This configuration yielded $F_2 = 0.470$, ROC–AUC = 0.708, and Recall = 0.605 on the test set, representing the best balance between recall sensitivity and false-positive control.

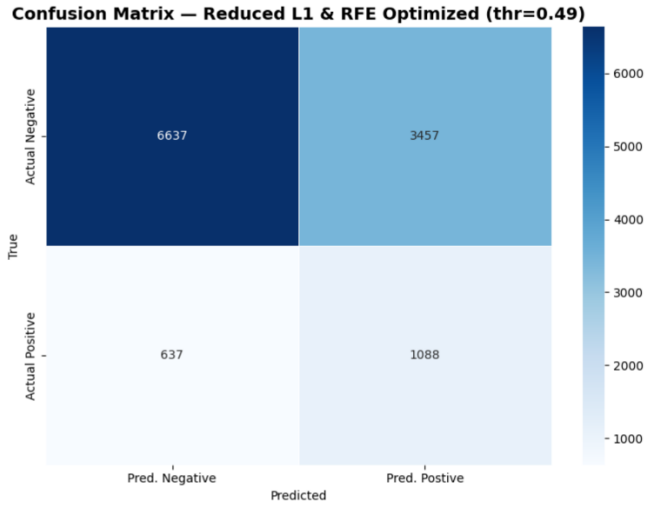
8.3.2. Implementation and Threshold Calibration

Following feature selection using Recursive Feature Elimination (RFE), dimensionality was reduced from 119 to 91 predictors. A threshold sweep (0.30–0.80) was then conducted to balance recall and total error cost, identifying 0.51 as the optimal threshold—maximizing recall-weighted sensitivity while maintaining operational precision.

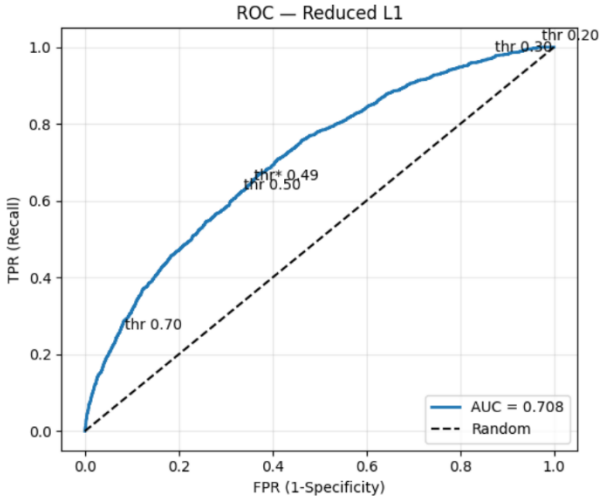
The final Random Forest exhibited lowest total misclassification (3,838), confirming strong generalization and minimal overfitting. Key features reinforced those from earlier models: Positive drivers: number_inpatient, comorbidity_score, number_diagnoses, service_utilization. Negative drivers: discharge_to_home, metformin, insulin. SHAP interpretability analysis (Appendix C, Figure 1) validated these patterns, revealing additive and clinically consistent risk gradients across patient subgroups.

9. Appendix C – Results

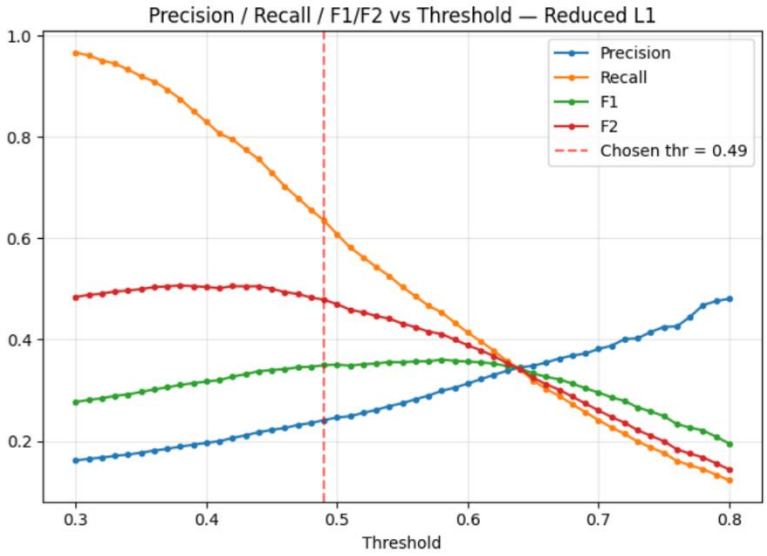
9.1 Logistic Regression



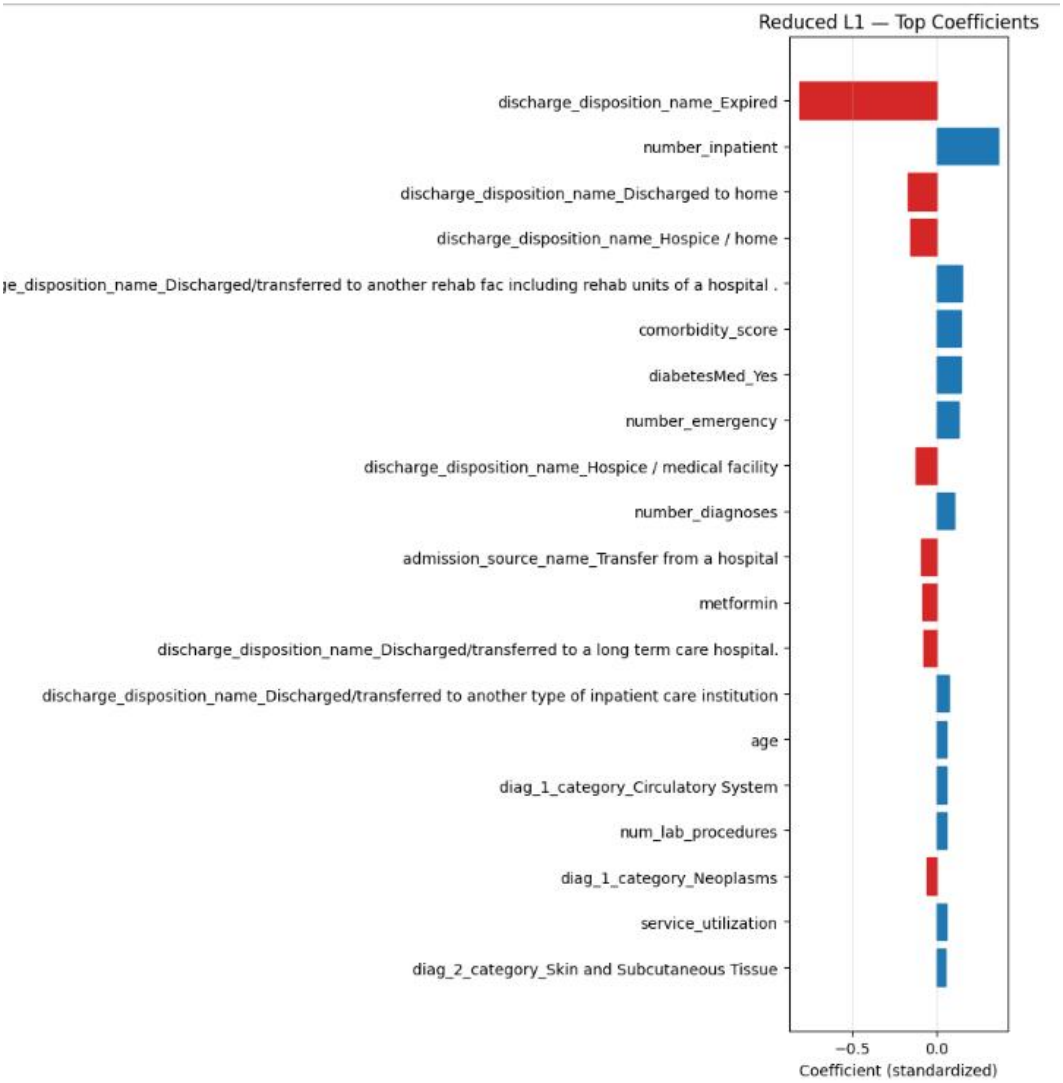
Appendix C - Modelling - Logistic Regression Figure 1. Confusion matrix for Reduced L1 (LASSO) Logistic Regression



Appendix C - Modelling – Logistic Regression Figure 2. ROC Curve for Reduced L1 (LASSO) Logistic Regression

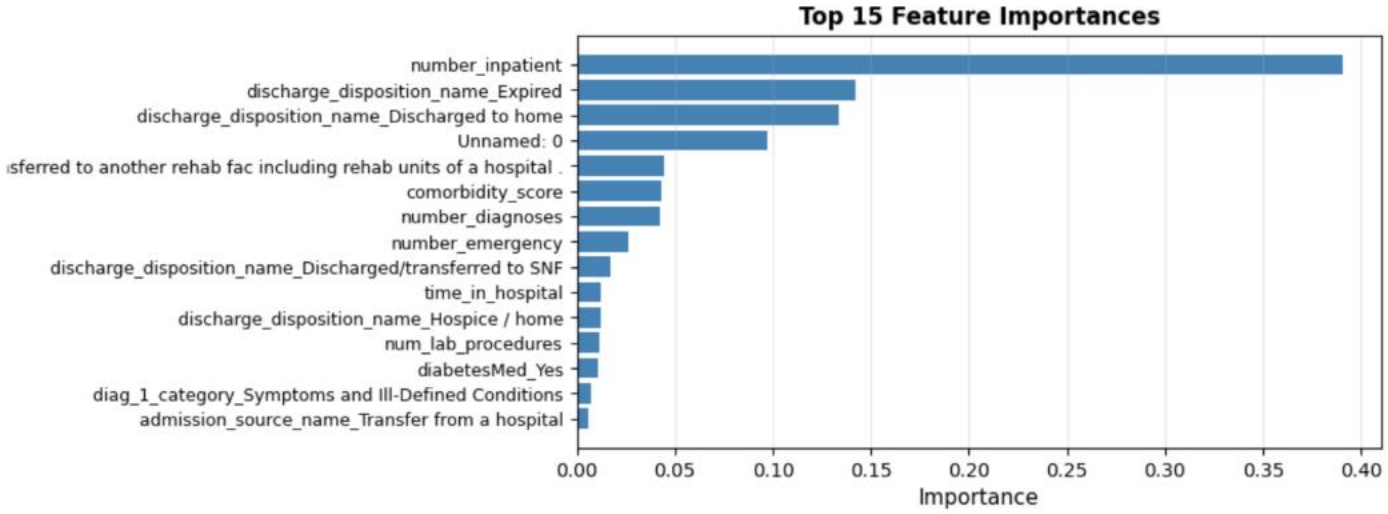


Appendix C - Modelling – Logistic Regression Figure 4. Precision/Recall/F1 vs Threshold for Reduced L1 (LASSO) Logistic Regression

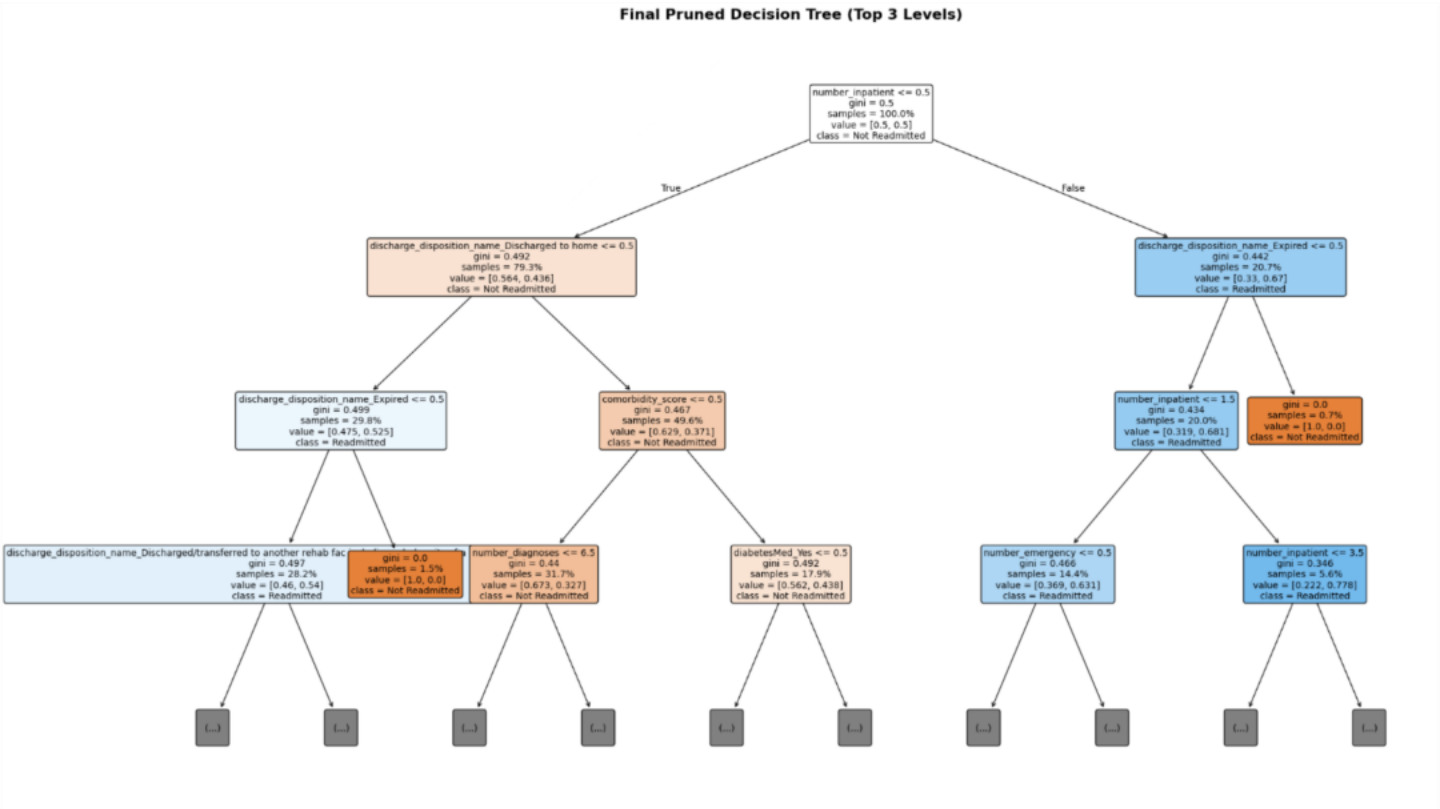


Appendix C - Modelling – Logistic Regression Figure 5. Top coefficients from Reduced L1 (LASSO) Logistic Regression

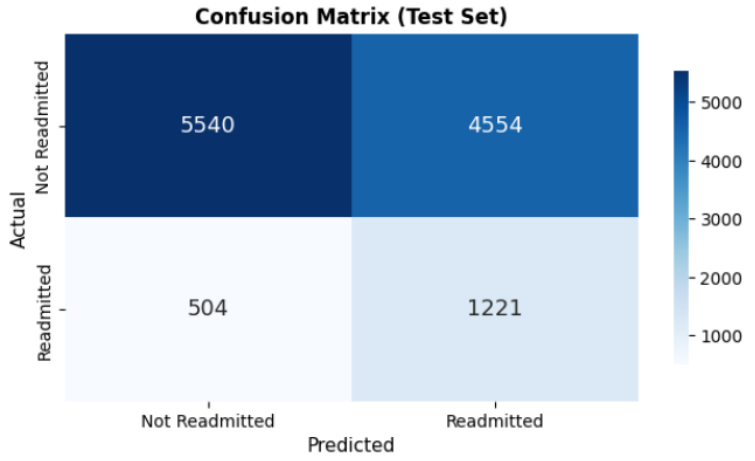
9.2 Decision Tree



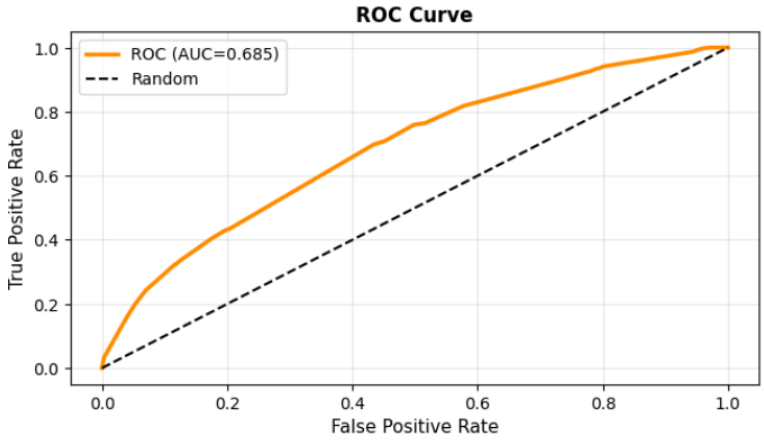
Appendix C - Modelling – Decision Tree Figure 1. Top 15 Feature Importance for Decision Tree Model



Appendix C - Modelling – Decision Tree Figure 2 Final Pruned Decision Tree (Top 3 Levels)

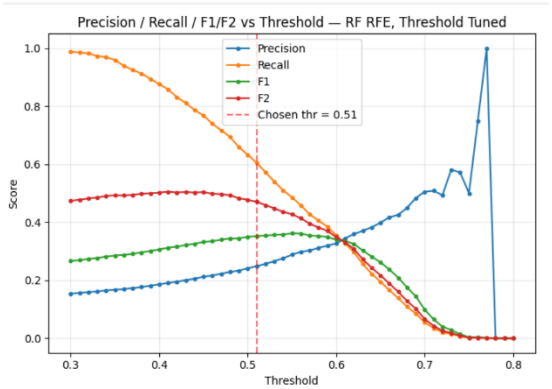


Appendix C - Modelling – Decision Tree Figure 6. Confusion Matric for Decision Tree Model

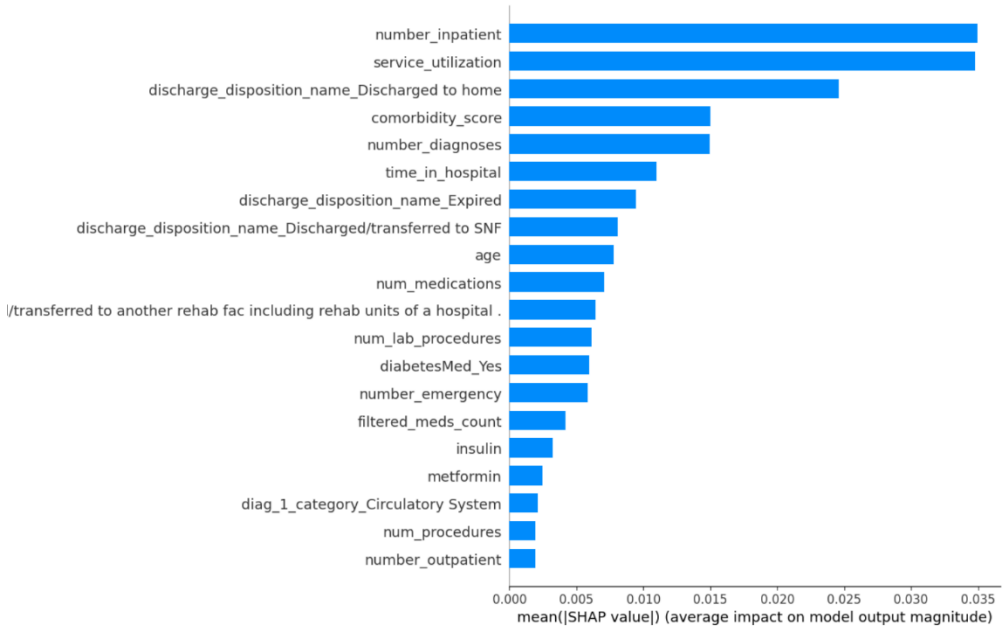


Appendix C - Modelling – Decision Tree Figure 7. ROC Curve for Decision Tree Model

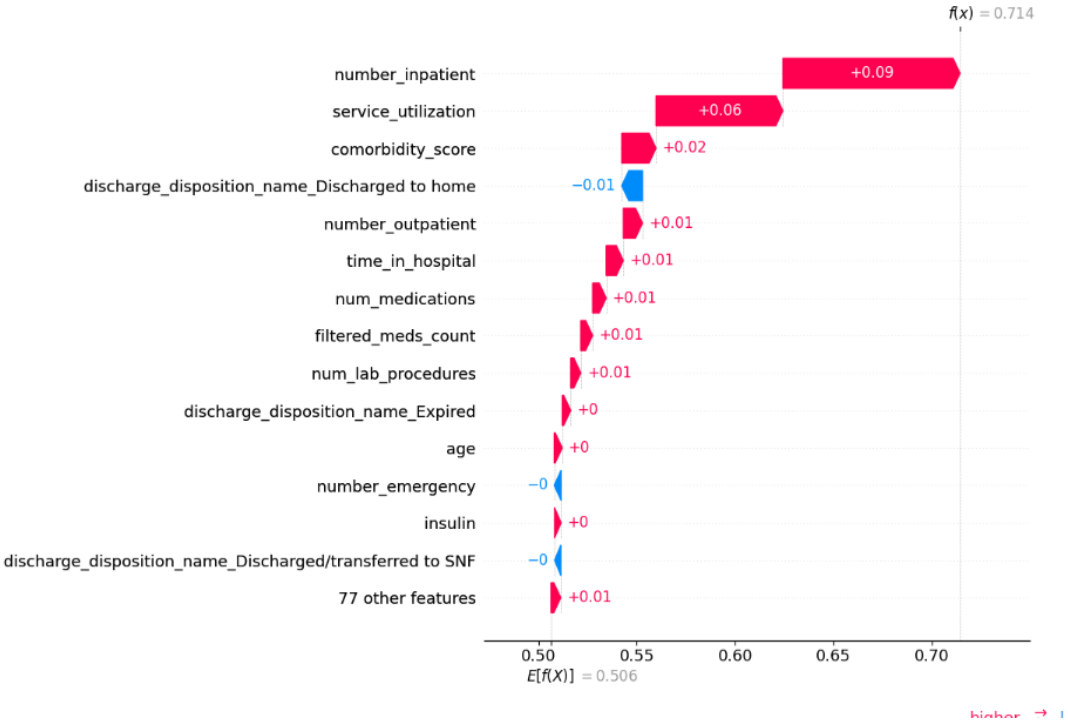
9.3 Random Forest



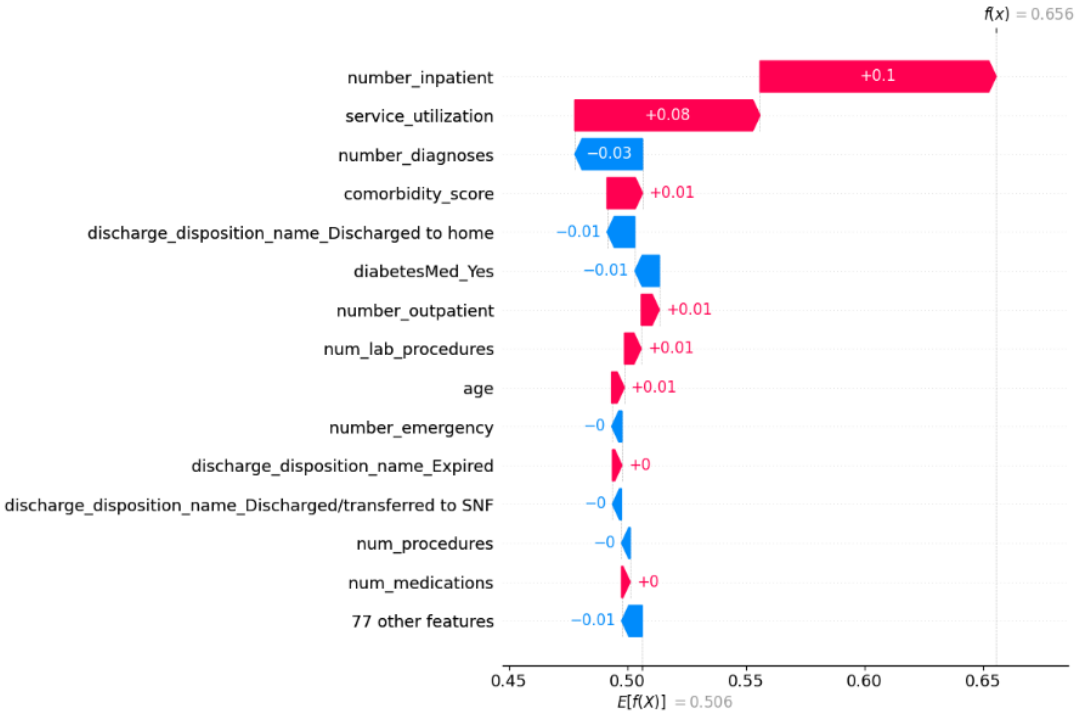
Appendix C - Modelling - RF Figure 5. Precision/Recall/F1 vs Threshold for Optimized RF RFE Eliminated and Threshold Tuned Model



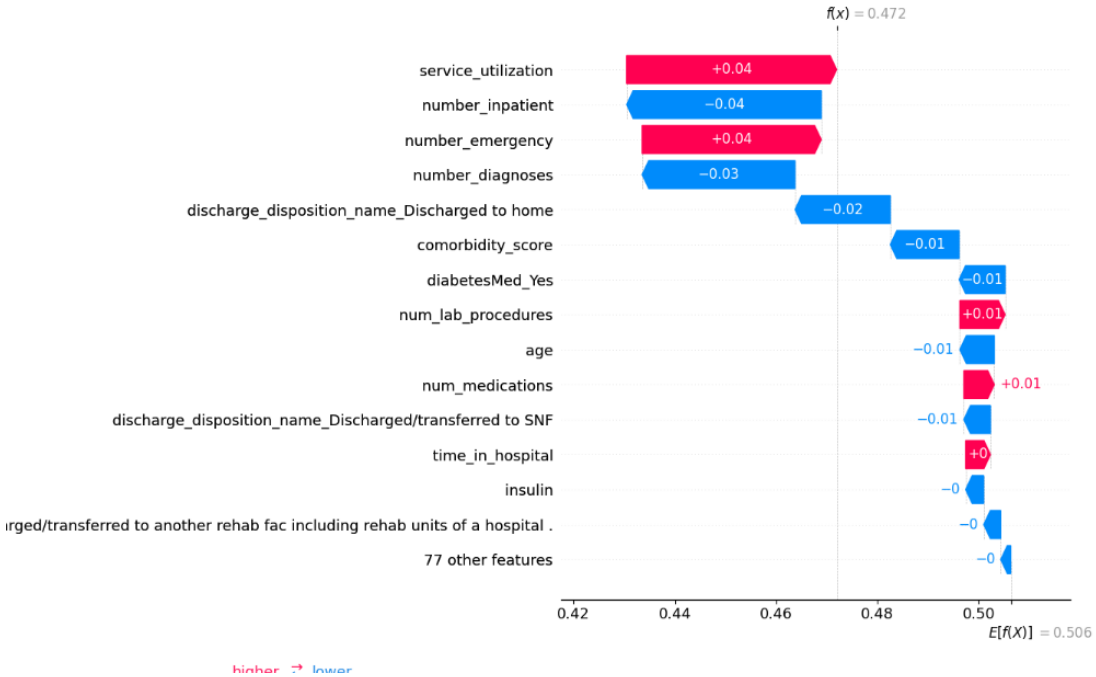
Appendix C - Modelling - RF Figure 6. SHAP Feature Importance Plot for Best RF Model



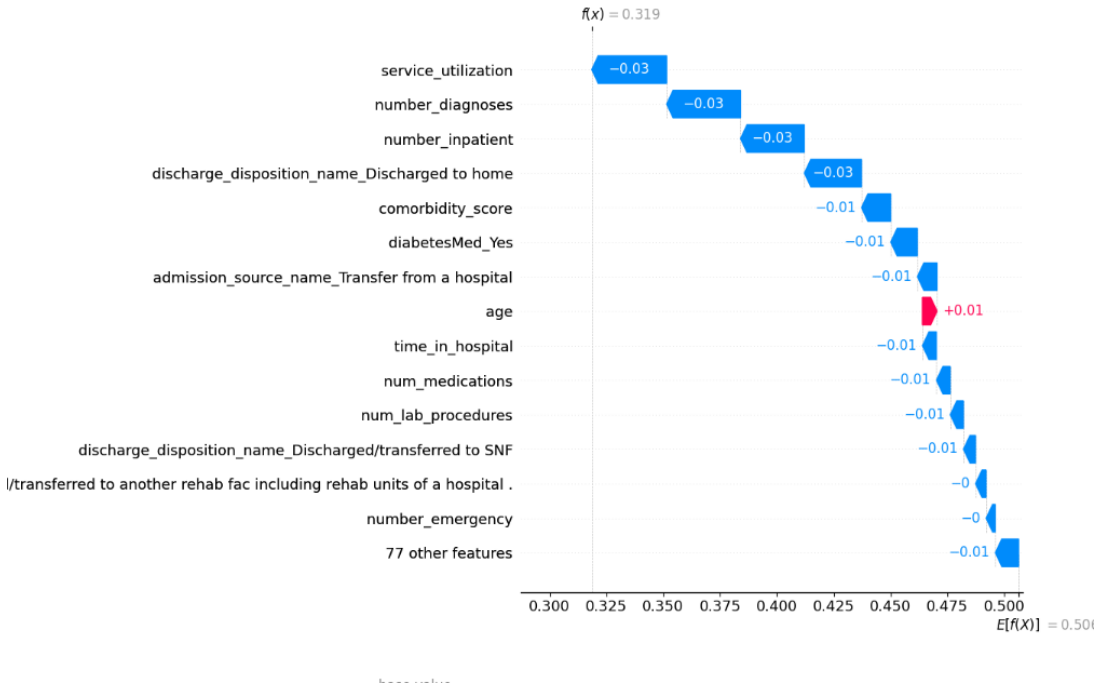
Appendix C - Modelling - RF Figure 7. Random Patient Index 5 SHAP Waterfall Plot for Visualization of Model Decision Pathway - True Positive Case



Appendix C - Modelling - RF Figure 8. Random Patient Index 12 - SHAP Waferfall Plot for Visualization of Model Decision Pathway - False Positive Case



Appendix C - Modelling - RF Figure 9. Random Patient Index 24 - SHAP Waterfall Plot for Visualization of Model Decision Pathway - True Negative Case



Appendix C - Modelling - RF Figure 10. Random Patient Index 42 - SHAP Waterfall Plot for Visualization of Model Decision Pathway - True Negative Case