University College London
School of Management

MSIN0097: Predictive Analytics

Individual Assignment

Candidate Number: FSMV3

Due: 22nd of April 2024

Total Word Count: 1957/2000

Individual Assignment

Module Leader Name: Dr. A. P. Moore

Table of Contents

Table of Figures

Table of Figures

1 Introduction

In 2020, diabetes contributed to 7.86 million hospital discharges in the United States, with patients facing increased risks of morbidity and mortality (Center for Disease Control and Prevention, 2023). The disease also led to substantial healthcare costs, accounting for half of the $174 billion spent on diabetes in the U.S., and represented 22% of all hospital inpatient days (American Diabetes Association, 2008). Thirty-day readmission rates, a key healthcare quality indicator, range from 7.7% to 20% among diabetic patients (Jiang *et al.*, 2005; Robbins and Webb, 2006), with up to 55% of these readmissions potentially preventable through better inpatient care and discharge planning (Ashton et al., 1995; Oddone et al., 1996). This study aims to develop machine learning models to to forecast the probability of diabetic patients being readmitted within 30 days post-discharge. The paper discusses dataset insights, the data cleaning process, model implementation, and the evaluation of model performance, concluding with the challenges faced and key findings.

2 Data Insights

2.1 Dataset: Diabetes 130-US Hospitals

The dataset represents the clinical care at 130 US hospitals between 1999 and 2008, detailing clinical care for 101’766 diabetic patients (Clore et al. 2014, details in Appendix I). Each record features hospital admissions, medications, treatments, and lab tests, and tracks early readmission within 30 days of discharge across 47 variables. These features encompass patient demographics, medical history, admission details, hospital stay characteristics (including test results and treatments), and discharge information, such as the destination post-discharge. An extended feature description is available in Appendix II, while a broader categorisation is illustrated in *Figure 1*.

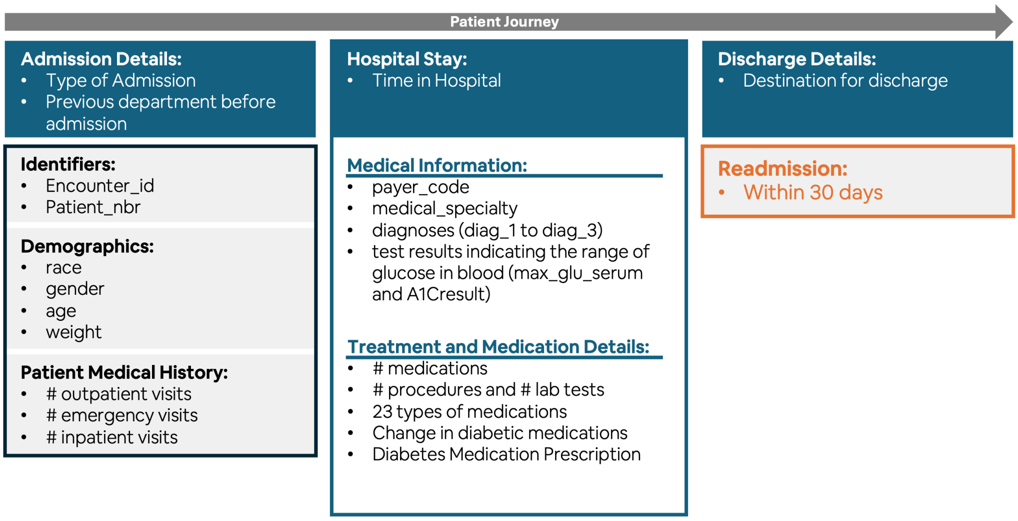


Figure 1 Features present in the dataset (own illustration)

2.2 Explorative Data Analysis

In the initial phase of the exploratory data analysis (EDA), I delved into the profiles of admitted and readmitted patients, focusing on various demographic aspects within each category. As illustrated in

Figure 2, a preliminary review of the dataset reveals a class imbalance typical of medical datasets (Belarouci and Chikh, 2017), with only 9.0% of the entries representing patients readmitted within 30 days. Notably, age seems to be an influential factor in readmissions, particularly for patients aged 65 years. Second, while differences in readmission rates by race suggest that this feature might contribute to the prediction model, it is essential to consider the proportional representation of different races in the patient population. Third, the initial analysis suggests gender may not strongly predict readmission within 30 days, since the distribution between readmitted women and men is balanced.

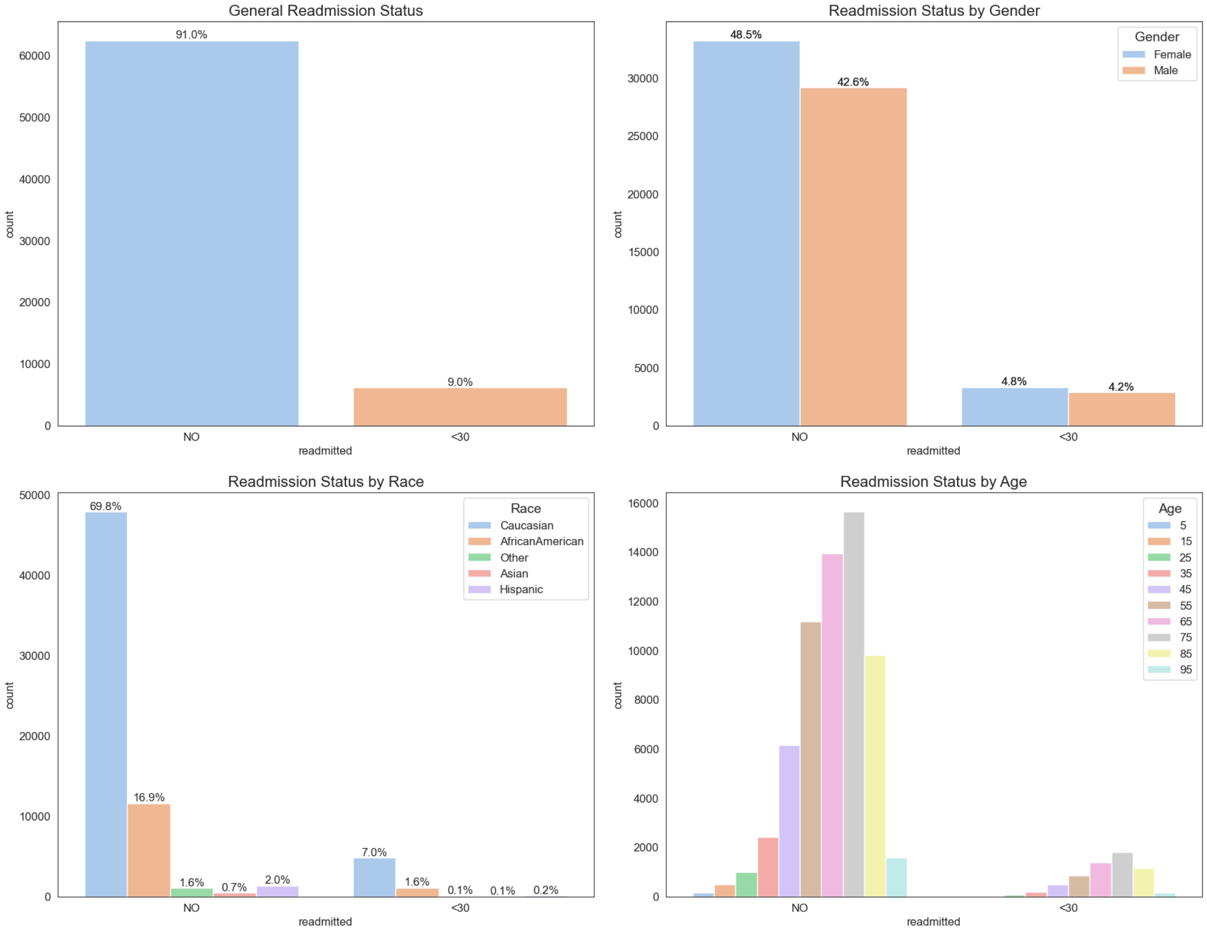


Figure 2 Readmissions by gender, race and, age

The second part of the EDA investigate the length of the hospital stays and how they correlate with age, providing insight into the patient care journey. As *Figure 3* shows, most of the hospitalisations with no readmission are brief, with a peak frequency between 2 to 4 days of stay, followed by a sharp decline in frequency for longer duration. The distribution of hospital stay lengths is more spread out for patients who were readmitted compared to those who were not. The lines are a visual indication of the probability density of hospital stay for the two patient groups.

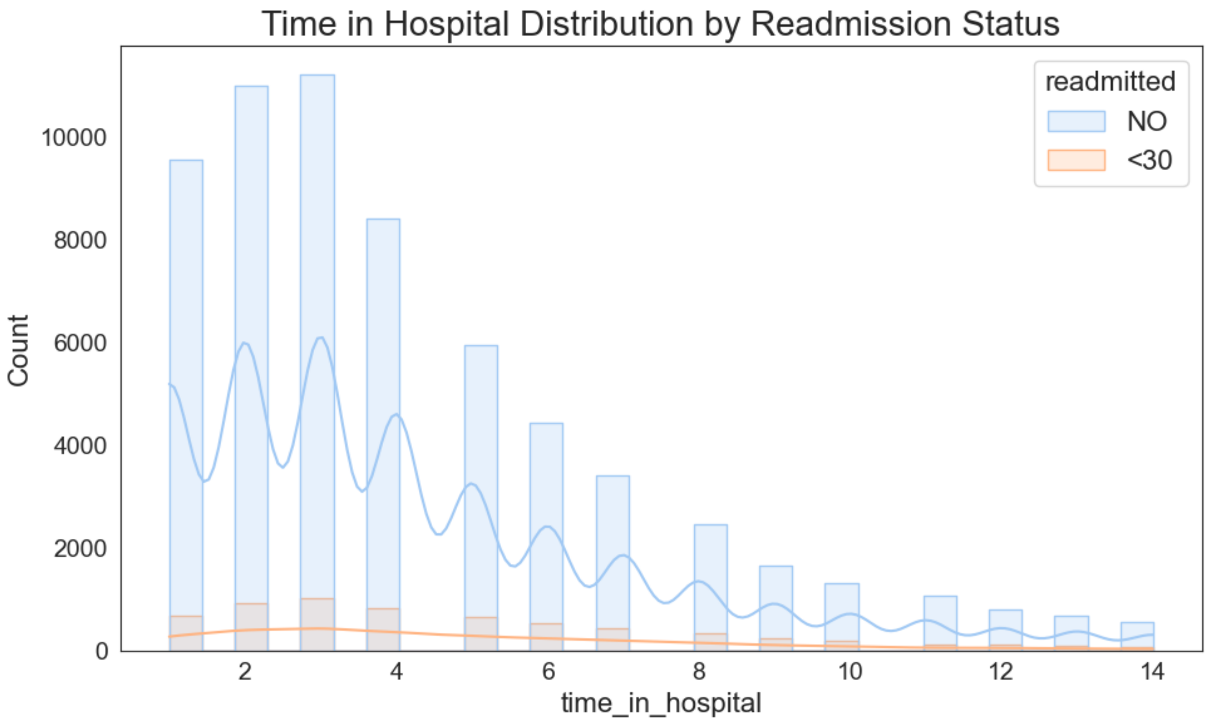


Figure 3 Time in hospital distribution by readmission status

Figure 4 shows that hospital stays tend to lengthen with age. Older age groups show greater variation in stay durations, suggesting more complex health issues. Notably, outliers are more common in these groups, indicating some unusually long stays. Additionally, there's a noticeable difference in average hospital time between readmitted and non-readmitted patients across ages 45, 55, 65, and 95.

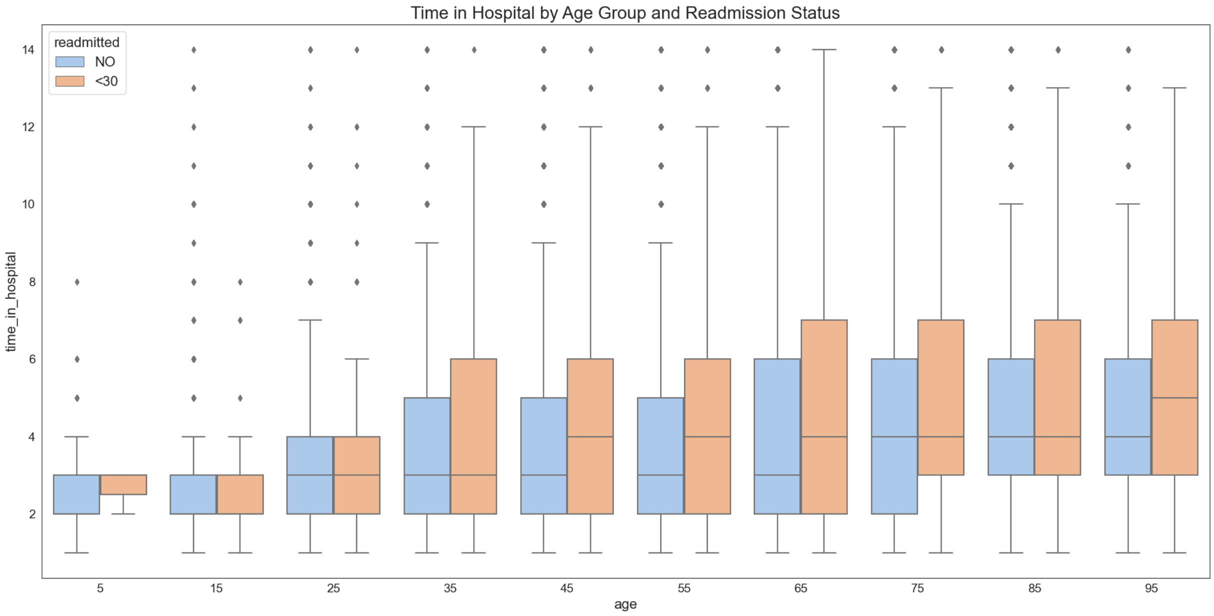


Figure 4 Time in hospital by age group by readmission status

Part three of the EDA, depicted in *Figure 5*, indicates a mild positive correlation between medication count and lab procedures, hinting at more complex health scenarios for patients with more lab tests. Age is somewhat positively linked to longer hospital stays and marginally to lab procedures. The correlations do not strongly predict readmissions, pointing to complex, non-linear variable relationships.

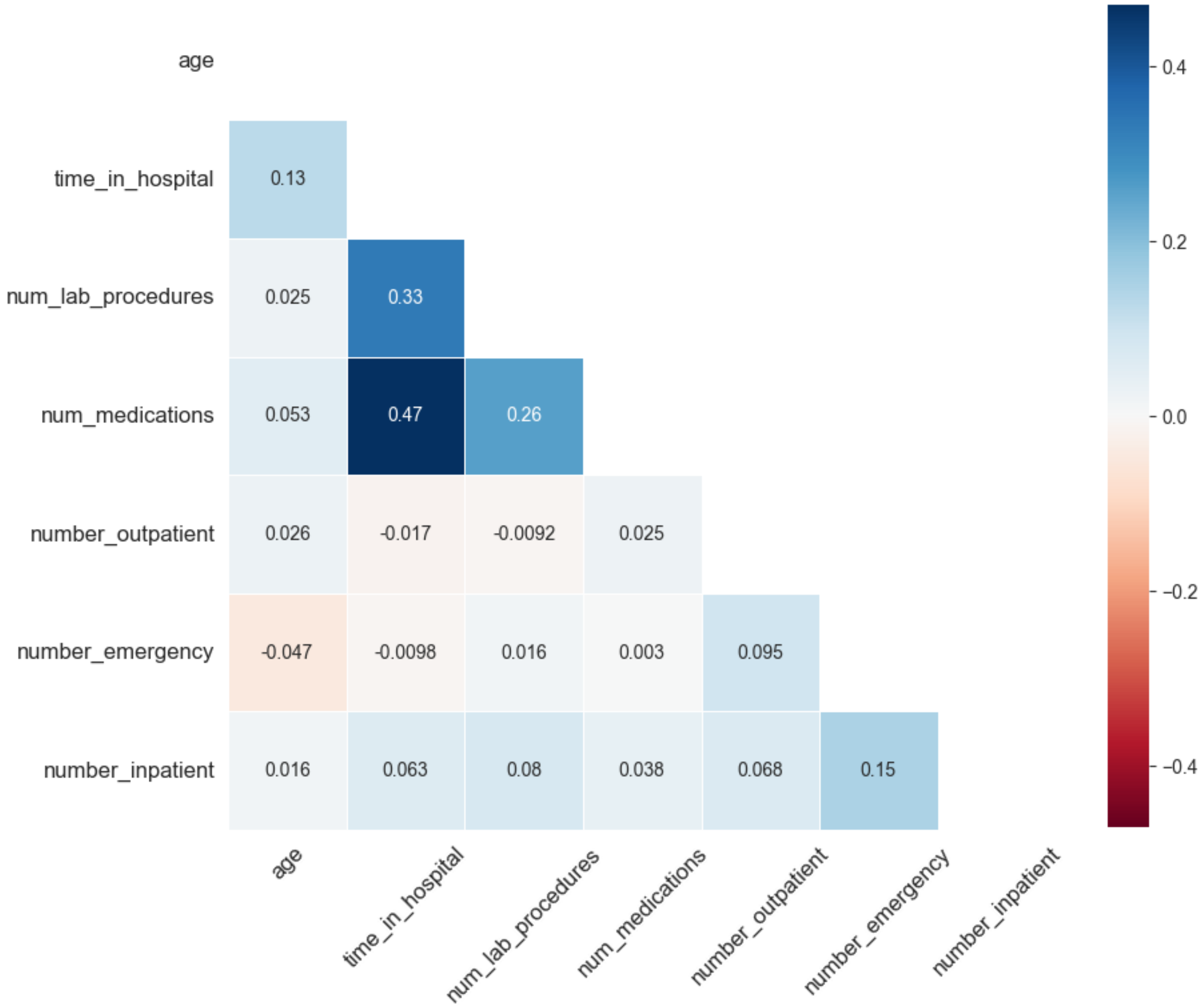


Figure 5 Correlation Matrix

Part four of the EDA examined value distributions in hospital admission variables, as depicted in *Figure 6*'s histograms with a log-scaled y-axis. Most variables are right-skewed, typical in healthcare data where a minority of patients have high event counts, indicative of severe conditions. Some variables, like medication count, vary widely. The log scale also suggests outliers in outpatient, emergency, and inpatient numbers.

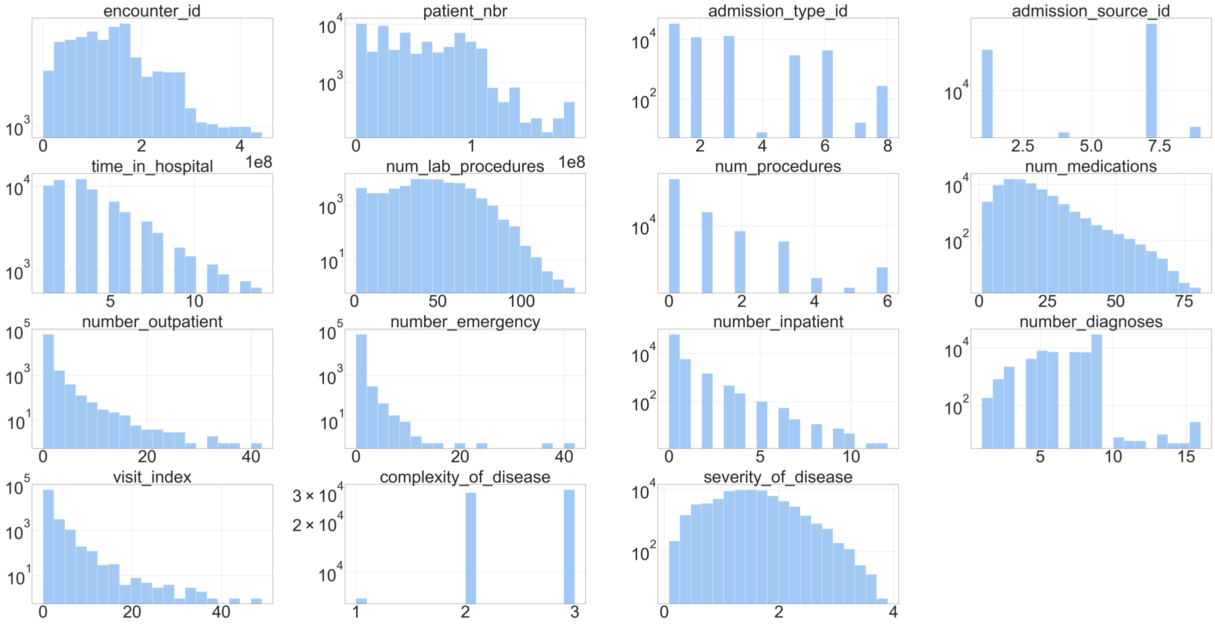


Figure 6 Distribution of continuous variables[1]

Other distributions analysis, as the one for the usage frequency of diabetic-related drugs can be found in Appendix III, since the majority of the distributions shows low frequency of prescription.

2.3 Data Cleaning and Pre-Processing

During the data cleaning and pre-processing stages, the dataset underwent a systematic refinement (see Appendix IV: Data Cleaning Steps). When checking for missing values, variables with substantial missingness, such as 'weight', were excluded, while for other attributes like 'A1Cresult', non-measurement instances were retained with semantic adjustment to clarify the future modelling. Secondly, when checking for data independence, which is critical for logistic regression, patient uniqueness was ensured by eliminating duplicate records. Thirdly, the number of categories has been consolidated for diverse columns to obviate the curse of dimensionality. Finally, all categorical variables have been hot-encoded, and target variable has been converted to binary labels. The number of observations in the dataset after cleaning and one hot encoding amounts at 68'577, with 197 features (excluding target variable). The chosen test train split is 30%.

[1] Horizontal axis represent their value, vertical axis represents count of observations present in that value bucket.

3 Models

3.1 Model Objectives & Performance Metrics

Models are evaluated using precision, recall, and the harmonic F1 score on the test set. Precision is vital in healthcare, ensuring that predictions of readmission are likely correct and prevent unwarranted treatments. Recall helps identify patients needing further care, crucial for preventing unsafe discharges. The F1 Score balances precision and recall, optimizing hospital resource use and maximizing patient safety. Accuracy is not used as a metric here because it can be misleading in imbalanced datasets where the proportion of actual cases is skewed.

3.2 Methodological Approach

Following section highlights how models have been selected and how different strategies for model performance optimisation have been implemented. Three different types of predictive models have been evaluated (see Discussion of the results): logistic regression(LR), random forest(RF), and neural networks (NN). While logistic regression is chosen for its simplicity and interpretability, which in the domain of medical data is a paramount, random forest is selected for its robustness to overfitting and ability to handle unbalanced data effectively. The Neural networks selected are multi-layer perceptron. They have also been included due to their capacity to model complex nonlinear relationships, since the correlations found in EDA indicated a potential non-linear relationship between the data.

To handle the bias in the data two different approaches have been evaluated. The synthetic minority over-sampling technique (SMOTE), which helps to tackle the class imbalance in the dataset. The underrepresented class of interest is the readmission, SMOTE artificially augment this class by creating synthetic samples (Maklin, 2022). Another approach to compensate class imbalances is the manual definition of class weights. With this approach the model is more sensitive to the minority class and thus, reducing bias in predictions (Brownlee, 2020).

To improve the models, feature scaling, selection and engineering have been performed. With rescaling through the function StandardScaler it was possible rescales the features to have a mean of zero and a standard deviation of one, which should help to ensure that all features contribute equally to the model (Brownlee, 2020b). Furthermore, feature selection was performed by investigating variable importance from the logistics regression and random forest models. This systematic way was imperative since domain expertise was missing to understand interplay of medical data. For feature engineering, three intuitive new measures were derived to enhance model performance, as shown in *Table 1.*

Feature	**Description**
Visit Index	Aggregate measure of healthcare utilisation calculated the sum of visits of a patient in a medical facility within the year prior the hospitalisation. $visit_{index} = \#outpatient + \#emergency + \#inpatient$
Disease Complexity	Measure quantified by the distinct categories of primary, secondary, and tertiary diagnoses. More complex cases are characterised by diverse diseases diagnoses. The intuition behind this feature is the more diagnoses, the more fragile the health status of the patient, increasing potentially the odds of being readmitted. $complexity_{disease} = disease_1 + disease_2 + disease_3$
Disease Severity	A composite index reflecting the overall disease burden, by incorporating various clinical interventions and outcomes. $severity_{disease} = \frac{\#medications + \#lab\ procedures + \#diagnoses}{total_{severity\ values}}$

Table 1 Manually created features

4 Discussion of the Results

As part of the project, eight models have been trained: four with logistic regression, two with random forest, and four neural networks.

4.1 Linear Models

Preliminary models using logistic regression and random forests explored class imbalance solutions and feature preprocessing. These models provided insights into threshold selection and feature standardization but did not significantly enhance prediction of hospital readmissions. The logistic regression models showed variations in performance based on the techniques applied for handling class imbalance and feature preprocessing, while random forest models offered insights into the impact of class weighting and threshold optimization on model accuracy. Detailed results, including performance metrics and feature importance analyses, are documented in Appendix V for logistic regression (LR) and Appendix VI for random forests (RF).

4.2 Neural Networks

The multilayer perceptron models(NN) in the project demonstrated varying levels of performance with increasing complexity and feature selection techniques, see *Table 1.*

	NN1	NN2	NN3	NN4
Architecture[2] - Layers	3 layers (64, 32, 1)	3 layers (32, 16, 1)	3 layers (32, 16, 1)	4 layers (384, 416, 32, 32)
Architecture – Dropout	None	Applied	Applied	Applied
Random Search	None	None	None	Applied
Number of parameters[3]	14’785	6’881	-	189’569
Ratio Data[4]:Parameters	3.25	6.97	-	0.25
Class Imbalances Techniques	None	Class Weights[5]	Class Weights SMOTE	Class Weights SMOTE
Features Selection	All	All	Top 20 RF features	Top 20 RF features
Feature Standardisation	Applied	Applied	Applied	Applied
Precision	14.32%	8.97%	16.18%	12.50%
Recall	7.78%	**99.28%**	23.73%	21.75%
F1 Score	10.80%	16.45%	**19.24%**	11.29%

Table 2 Neural Network models performance

The initial model, NN1, established a baseline with a lower data point-to-parameter ratio than recommended[6], potentially affecting performance. To improve this, NN2 halved the number of parameters and implemented manual class weights to prioritize the minority class, which led to a near 6% increase in the F1 score. NN2 exhibited a high recall rate of 99.28%, indicating effectiveness in identifying readmissions. However, NN2 has a low precision, meaning that it is catching a significant number of false positive (incorrectly predicting readmission when it is not necessary). This could lead to unnecessary treatments and increased healthcare costs.

Model NN3, which incorporated SMOTE and the top 20 features identified by RF4, was designed to achieve a balance between precision and recall. Reducing the number of features aimed to mitigate the noise and enhance the model's predictive strength. It worth noting that all the created features were included in the list, indicating that they capture relationships not evident in the individual classes alone. Adjusted class weights, based on the dataset's class distribution, contributed to a more balanced

[2] RELU Activation for hidden layers. Sigmoid for output layer

[3] Number of parameters is calculated by the sum of parameters for each layer: (Inputs x Neurons)+ Biases.

[4] The training set contains 48'003 data points.

[5] See Appendix: Class Weights

[6] The recommend principle is to have about ten times as many data points as the total number of internal parameters (Moore, 2024).

performance, with an approximate 3% increase in F1 score over previous models, indicating better generalization. This balance is evidenced by the performance curves for NN2 and NN3 at various thresholds, as shown in *Figure 7* and *Figure 8*. NN3's balanced approach suggests improved ability to distinguish between classes, making it suitable for scenarios where both false positives and false negatives carry significant consequences.

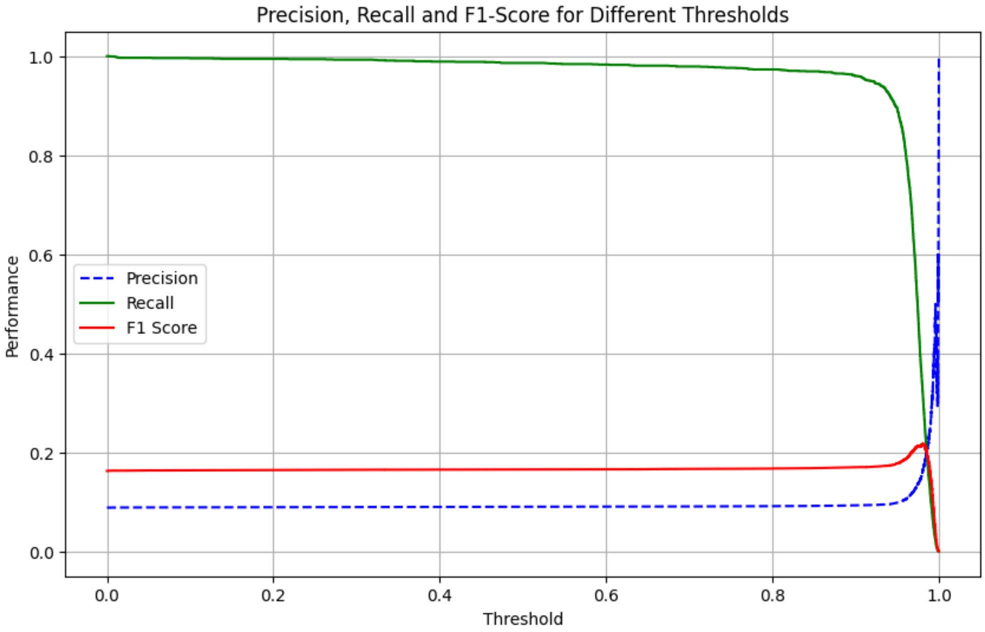


Figure 7 Model NN2 performance metrics at different thresholds

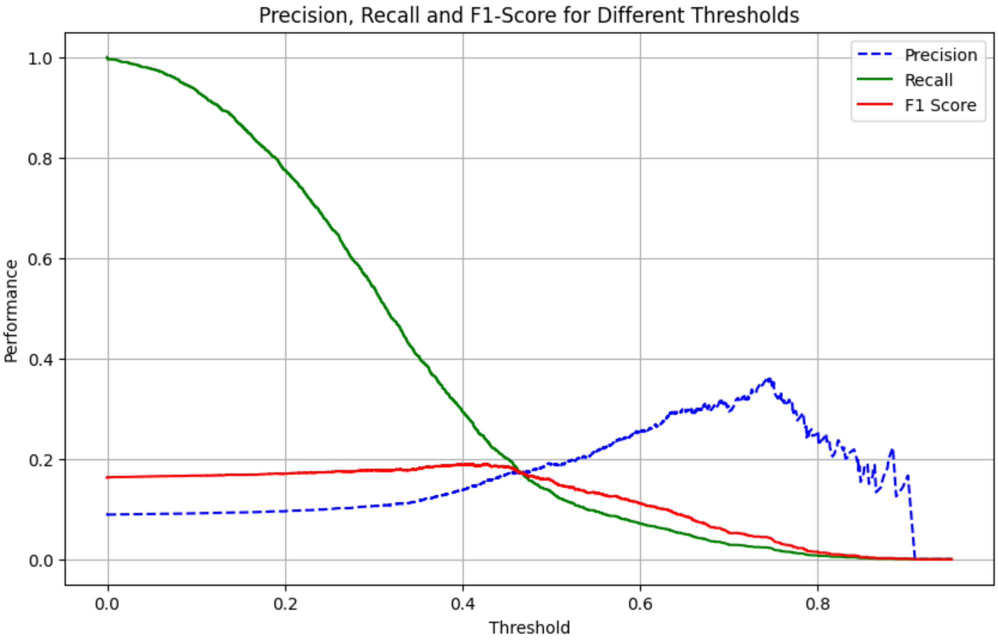


Figure 8 Model NN3 performance metrics at different thresholds

Model NN4's architecture, with 189'569 parameters, was designed to fine-tune predictions of patient readmissions. The tuning involved adjusting the number of layers and rates of dropout and learning to optimize key metrics—precision, recall, and the area under the precision-recall curve. Class imbalances were addressed by applying differential weights to classes, a method intended to enhance the model's sensitivity to less frequent but critical events such as patient readmissions.

The complexity of NN4's design, intended to capture subtle patterns in the data, comes with the risk of overfitting. As *Figure 9* shows, this seems to be the case with NN4, as indicated by the consistently improving training metrics over epochs—suggesting the model is effectively learning from the training dataset.

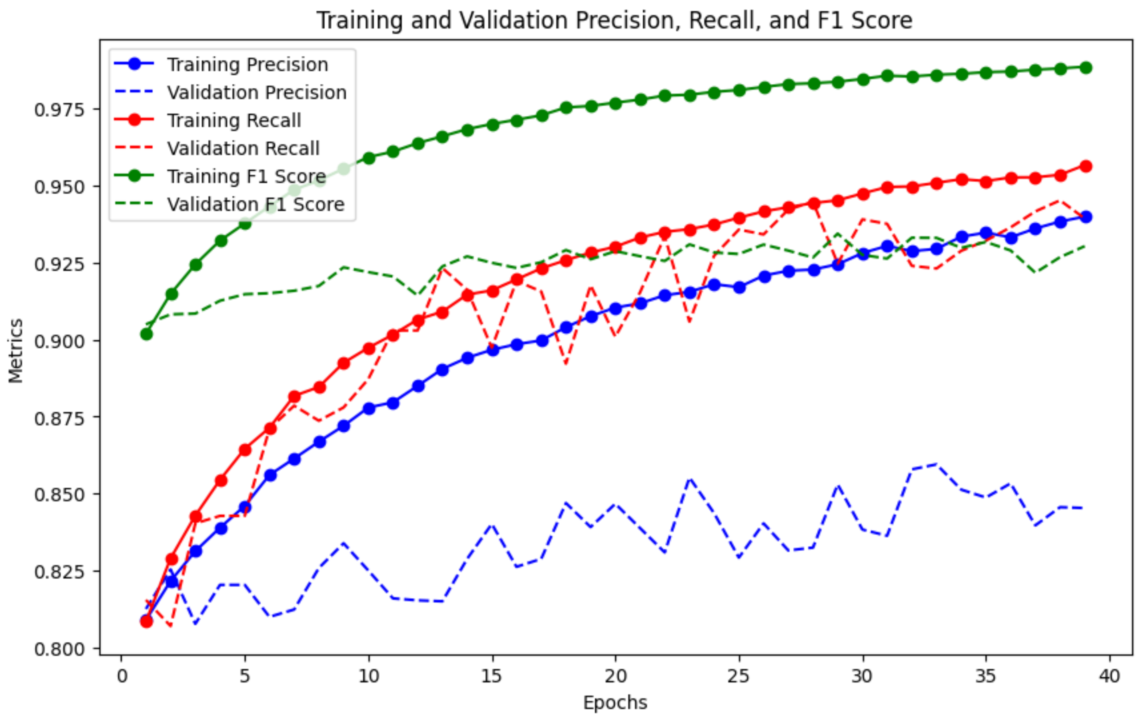


Figure 9 Training process for NN4

The test set results confirm this supposition, with a low F1 score (11.29% vs 19% of NN3). A high recall (21%) with a low precision (12%) means the model is good at identifying patients at risk of readmission but at the cost of also incorrectly identifying many who are not at risk.

5 Conclusion

This project assessed various models for predicting hospital readmissions, providing valuable insights. Utilizing techniques like SMOTE and class weights improved predictability. Feature standardization and engineering further enhanced performance. The NN3 model struck an optimal balance between precision and recall, vital in healthcare. However, NN4 likely overfitted, as training data suggested. Future research should examine misclassification costs to refine decision thresholds for cost-effective preventive care. Enhancing model generalizability remains crucial for optimizing patient care and resource allocation.

References

Repository: https://github.com/acerutti/ucl-ML-hospital-readmissions

American Diabetes Association (2008) 'Economic costs of diabetes in the U.S. in 2007', Diabetes Care, 31, pp. 596–615.

Ashton, C.M., Kuykendall, D.H., Johnson, M.L., Wray, N.P. and Wu, L. (1995) 'The Association between the Quality of Inpatient Care and Early Readmission', Annals of Internal Medicine, 122(6), p. 415. doi:10.7326/0003-4819-122-6-199503150-00003.

Axon, R.N. and Williams, M.V. (2011) 'Hospital Readmission as an Accountability Measure', JAMA, 305(5), pp. 504–505. doi:10.1001/jama.2011.72.

Belarouci, S. and Chikh, M.A. (2017) 'Medical imbalanced data classification', Advances in Science, Technology and Engineering Systems Journal, 2(3), pp. 116–124. doi:10.25046/aj020316.

Brownlee, J. (2020a) '8 Tactics to combat imbalanced classes in your machine learning dataset', Machine Learning Mastery. Available at: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ (Accessed: 16 April 2024).

Brownlee, J. (2020b) 'How to use StandardScaler and MinMaxScaler Transforms in Python', Machine Learning Mastery. Available at: https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/ (Accessed: 16 April 2024).

Center for Disease Control and Prevention (2023) National Diabetes Statistics Report, National Diabetes Statistics Report Website. Available at: https://www.cdc.gov/diabetes/data/statistics-report/index.html (Accessed: 16 April 2024).

Clore, J., Cios, K., DeShazo, J. and Strack, B. (2014) Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository. Available at: https://doi.org/10.24432/C5230J (Accessed: 15 February 2024).

Dhatariya, K., Mustafa, O.G. and Rayman, G. (2020) 'Safe care for people with diabetes in hospital', Clinical Medicine, 20(1), pp. 21–27. doi:10.7861/clinmed.2019-0255.

Jiang, H., Stryer, D., Friedman, B. and Andrews, R. (2005) 'Racial/Ethnic disparities in potentially preventable readmissions: the case of diabetes', American Journal of Public Health, 95(9), pp. 1561–1567. doi:10.2105/ajph.2004.044222.

Maklin, C. (2022b) 'Synthetic Minority Over-sampling TEChnique (SMOTE) - Cory Maklin - Medium,' Medium, 21 May. https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c (Accessed: 18 April 2024).

Oddone, E.Z., Weinberger, M., Horner, M., Mengel, C., Goldstein, F., Ginier, P., Smith, D., Huey, J., Farber, N.J., Asch, D.A. and Loo, L. (1996) 'Classifying general medicine readmissions', Journal of General Internal Medicine, 11(10), pp. 597–607. doi:10.1007/bf02599027.

Robbins, J.M. and Webb, D.A. (2006) 'Diagnosing diabetes and preventing rehospitalisations', Medical Care, 44(3), pp. 292–296. doi:10.1097/01.mlr.0000199639.20342.87.

Appendix

Appendix I: Criteria for Data Extraction

The data present in the dataset used for this project was extracted from a US database of 130 hospitals and integrated delivery networks. Information was retrieved from the database if it satisfied the following criteria:

(1) It is an inpatient encounter (a hospital admission).

(2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered into the system as a diagnosis.

(3) The length of stay was at least 1 day and at most 14 days.

(4) Laboratory tests were performed during the encounter.

(5) Medications were administered during the encounter

Appendix II: Dataset Detailed Description

Variable Name	Type	Description
encounter_id		Unique identifier of an encounter
patient_nbr		Unique identifier of a patient
race	Categorical	Values: Caucasian, Asian, African American, Hispanic, and other
gender	Categorical	Values: male, female, and unknown/invalid
age	Categorical	Grouped in 10-year intervals: [0, 10), [10, 20),..., [90, 100)
weight	Categorical	Weight in pounds.
admission_type_id	Categorical	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
discharge_disposition_id	Categorical	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
admission_source_id	Categorical	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
time_in_hospital	Integer	Integer number of days between admission and discharge
payer_code	Categorical	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay
medical_specialty	Categorical	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
num_lab_procedures	Integer	Number of lab tests performed during the encounter
num_procedures	Integer	Number of procedures (other than lab tests) performed during the encounter
num_medications	Integer	Number of distinct generic names administered during the encounter
number_outpatient	Integer	Number of outpatient visits of the patient in the year preceding the encounter
number_emergency	Integer	Number of emergency visits of the patient in the year preceding the encounter
number_inpatient	Integer	Number of inpatient visits of the patient in the year preceding the encounter
diag_1	Categorical	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
diag_2	Categorical	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
diag_3	Categorical	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values

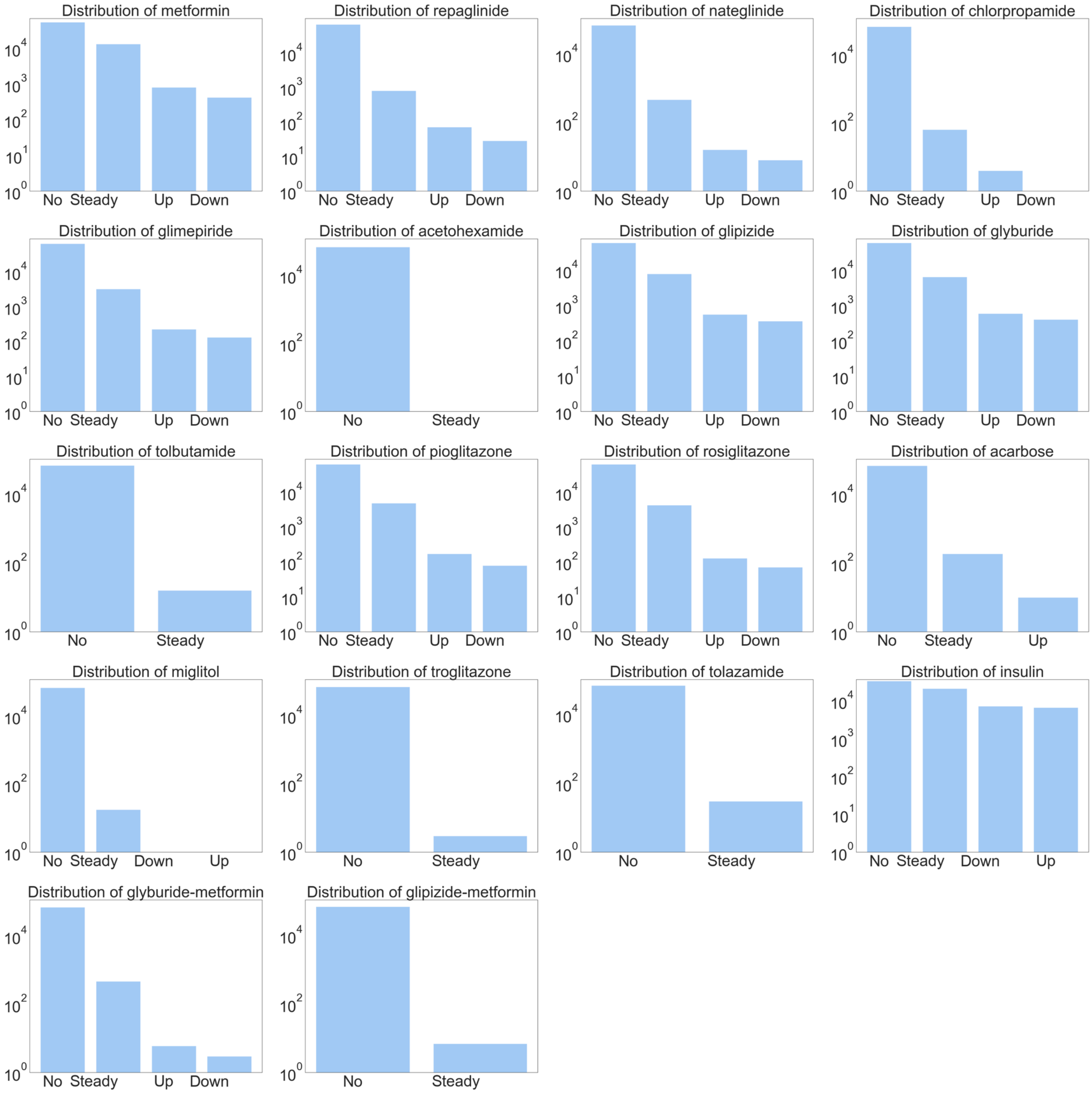
number_diagnoses	Integer	Number of diagnoses entered to the system
max_glu_serum	Categorical	Indicates the range of the result or if the test was not taken. Values: >200, >300, normal, and none if not measured
A1Cresult	Categorical	Indicates the range of the result or if the test was not taken. Values: >8 if the result was greater than 8%, >7 if the result was greater than 7% but less than 8%, normal if the result was less than 7%, and none if not measured.
metformin	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
repaglinide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
nateglinide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
chlorpropamide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
glimepiride	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
acetohexamide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
glipizide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
glyburide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
tolbutamide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased

		during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
pioglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
rosiglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
acarbose	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
miglitol	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
troglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
tolazamide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
examide	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
citoglipton	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
insulin	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
glyburide-metformin	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed

glipizide-metformin	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
glimepiride-pioglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
metformin-rosiglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
metformin-pioglitazone	Categorical	The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed
change	Categorical	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: change and no change
diabetesMed	Categorical	Indicates if there was any diabetic medication prescribed. Values: yes and no
readmitted	Categorical	Days to inpatient readmission. Values: <30 if the patient was readmitted in less than 30 days, >30 if the patient was readmitted in more than 30 days, and No for no record of readmission.

Appendix III: Distribution of Diabetic-Related Drugs

The figure below presents the usage frequency distribution of diabetic-related drugs on a logarithmic scale. The categories generally encompass "steady" for regular use, "no" for non-administration, and indicators for dosage adjustments—"up" for an increase and "down" for a decrease. The distributions reveal that some drugs exhibit low frequencies, as evidenced by minimal heights on the log scale, suggesting they may be prescribed for more complex conditions. For the modelling part and feature selection part it is important to observe if higher usage frequency in a particular drug might be more relevant as feature since they might reflect a common treatment pattern. Furthermore, drugs with frequent dosage changes might indicate unstable health conditions or need for additional medication management, which might be associated with readmission risk.



Appendix IV: Data Cleaning Steps

1. Recoding Readmission Variable: The 'readmitted' column was recoded to address the study's scope. Entries indicating readmission after 30 days ('>30') were reclassified to 'NO' to focus on early readmissions which are pertinent to the study’s objectives.
2. Removing Columns with Excessive Missing Values: The 'weight' column, which contained the highest proportion of missing values, was removed from the dataset to prevent potential bias and data sparsity issues.
3. Handling Not Measured Indicators: Columns 'A1Cresult' and 'max_glu_serum' were preserved despite having 'NaN' entries indicating tests were not performed. To prevent misinterpretation by the model, 'NaN' values were maintained, and columns were renamed accordingly.
4. Assessing Medical Specialty: Before deciding on its removal, the 'medical specialty' column underwent a detailed review due to its potential impact on readmission rates.
5. Excluding Irrelevant Columns: The 'payer_code' column was dropped, having been deemed non-informative for predicting readmission.
6. Maintaining Informative Columns with Few Missing Values: The 'admission_source_description' and 'discharge_disposition_description' columns were retained in their original form as their missing values were under 7% and 4%, respectively.
7. Managing Missing Racial Data: Entries with missing 'race' data, accounting for 2.23% of the dataset, were omitted to avoid biased imputation.
8. Imputing Diagnostic Codes: For 'diag_1', 'diag_2', and 'diag_3' columns, missing values were filled with the most common diagnoses present within each respective column.
9. Gender Data Correction: Rows with 'Unknown/Invalid' gender data were identified and excluded to preserve data accuracy.
10. Addressing Duplicate Records: The dataset was scrutinized for duplicate entries using the 'patient_nbr' identifier. Only the first occurrence of each patient's record was kept to ensure the independence of observations.
11. Eliminating Biased Data: To avoid bias in the model, patients who were documented to have died post-ICU were excluded from the dataset, as their readmission was impossible.
12. Transforming Age Data: The age data, originally presented in ranges, was converted to numerical values to reflect its ordinal nature.
13. Consolidating Discharge Information: A new mapping strategy was employed to consolidate the 'discharge_disposition_id' into fewer, more general categories. Special cases were handled with a lambda function for accuracy.
14. Reducing Medical Specialty Categories: The number of categories in the 'medical specialty' column was significantly reduced from 71 to 11, followed by the removal of the original column with granular data.

15. Categorizing Disease Codes: To prevent the issue of high dimensionality from one-hot encoding, ICD-11 disease codes were grouped into fewer categories, aligning with the World Health Organization's classifications.

Appendix V: Logistic Regression Models

The logistic regression models as shown in *Table 3* differentiate from how they handle class imbalance and from feature preprocessing. The goal was to explore best practices for addressing class imbalance and selecting adequate threshold to correctly identify patients needing readmission without including too many patients unnecessarily.

	L1	L2	L3	L4
Class Imbalances Techniques	None	None	Manual Class Weights	SMOTE
Threshold Search for Max F1	None	None	Applied	Applied
Feature Standardisation	None	Applied	Applied	Applied
Precision	13.49%	13.14%	17.85%	37.50%
Recall	50.03%	51.18%	29.48%	0.16%
F1 Score	21.25%	21.25%	**22.24%**	0.33%

Table 3 Logistic regression models performance

From this step we observe that selecting adequate threshold increase the balance between precision and recall. Furthermore, we deduce that standardisation with (StandardScaler function) on the values did not lead to clear improvement in the model’s performance. Moreover, the use of SMOTE doesn't seem to have drastically improved the model's ability to predict readmissions. A potential reason for this could be that while SMOTE has created more samples for the minority class, these samples have introduced noise or these were not representative of the true underlying distribution.

Choosing the right threshold for predicting hospital readmissions is crucial, striking a balance between the need for high recall to safeguard patient health and the need to limit false positives to avoid overburdening healthcare systems. Optimal thresholding ensures accurate identification of patients at risk of readmission without incurring unnecessary costs or interventions. *Figure 10* visually demonstrates this delicate trade-off, emphasizing the importance of a balanced approach, as illustrated in *Figure 10*, to enhance model performance and healthcare outcomes.

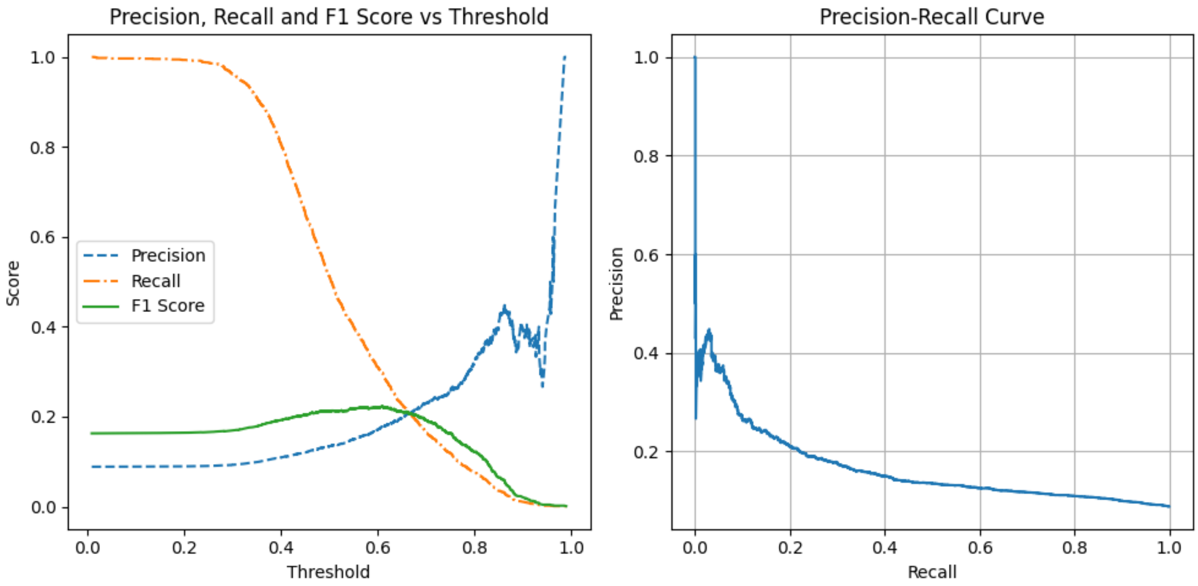


Figure 10 Model LR2 performance metrics at different threshold.

Appendix VI: Random Forest Models

In a second modelling step, we analysed feature importance through different random forest models. *Table 4*, presents the results of the models trained.

	RF1	RF2
Class Imbalances Techniques	SMOTE	None
Threshold Search for Max F1	Applied	Applied
Feature Standardisation	None	Applied
Precision	25.75%	20.00%
Recall	3.78%	21.00%
F1 Score	6.84%	**20.00%**

Table 4 Random Forest models performance

The RF2 after the Gridsearch provides a more balanced performance between precision and recall, which is desirable based on your objectives. RF2 may have benefitted from a more complex decision boundary and a Grid Search to optimize hyperparameters. Class_weight has been set manually in L3, while Random Forest used Grid Search, potentially leading to a better fine-tuning. While for optimise the F1 score in L3 model the optimal threshold was 0.61, for RF2 after the Gridsearch fine tuning, the optimal threshold was 0.16, see figure below. This highlights again that the optimal threshold is model-depended.

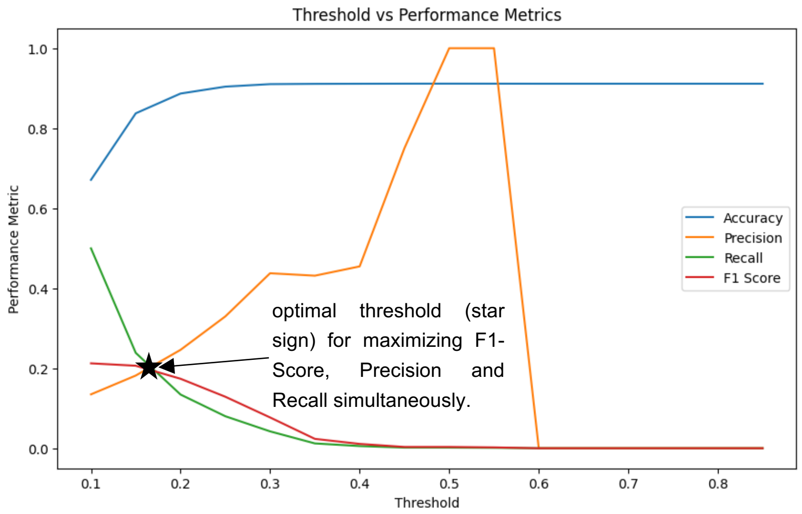


Figure 11 Model RF2 performance metrics at different thresholds

Appendix VI: Class Weights for Models NN2 & NN3

NN2:

$$Not\ readmitted\ Patients: \qquad 1$$

$$Readmitted\ Patients: \qquad 300$$

NN3:

$$Not\ readmitted\ Patients: \qquad 1$$

$$Readmitted\ Patients: \qquad \frac{len(y_{train}) - \sum y_{train}}{\sum y_{train}} \approx 10$$