NOVA IMS
Information Management School

READY TO BE DISCHARGE: EXAMINING HOSPITAL READMISSIONS

MACHINE LEARNING PROJECT

Group 16

Duarte Barbosa, number: 20230366
Emar Ribas, number: 20230459
Jéssica Vicente, number: 20230744
Peter Falterbaum, number: 20230956
Rita Matias, number: 20230496

December 2023

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

INDEX

Abstract

This study focuses on the analysis of key factors contributing to hospital readmissions, which is a significant factor in medical costs and an indicator of care quality. The goal of the project is to employ machine learning techniques on a dataset that has patient medical records from an US hospital, and evaluate various classification models to predict readmission rates, aiming to identify the most accurate model. Initially, the dataset is prepared through various preprocessing techniques. The approach of the study is twofold: firstly, it develops a binary classification to identify patients' readmission likelihood as 'Yes' or 'No' within a 30-day post-discharge period. Secondly, it creates a multiclass classification to predict patients' readmission more precisely into one of three categories: 'No', 'less than 30 days', and 'greater than 30 days'. As a result, Logistic Regression was found to be the optimal classifier for binary and Random Forest for multiclass classification.

Keywords: Machine Learning, Feature Engineering, Feature Selection, Predictive Model, Logistic Regression, Random Forest

1. Introduction

This study explores the complex factors leading to hospital readmissions, with a specific emphasis on diabetic patients. Our goal is not only to deepen our comprehension of the underlying causes of these readmissions but also to develop sophisticated predictive models to foresee patient readmissions.

Hospital readmissions are a significant factor in the rising costs of healthcare and serve as a key indicator of care quality. In the United States, the annual cost associated with the readmission of diabetic patients is estimated to be $25 billion (Sarthak et al., 2021). As a response, U.S. hospitals have been facing financial penalties for elevated readmission rates since the implementation of the Hospital Readmissions Reduction Program (Hospital Readmissions Reduction Program (HRRP) | CMS, 2020).

Readmission is defined as a patient's return to the hospital within a predetermined period following the initial discharge, often within 30 days. This measure is critical both financially and as a reflection of medical service effectiveness. Through an in-depth analysis of readmission causes, our study seeks to elevate patient care and enhance awareness regarding associated risk factors. Bhuvan et al. (2016) have notably underlined the necessity to differentiate between short-term (within 30 days) and long-term (post-30 days) readmissions for more effective patient care post-discharge. Decreasing readmission rates result therefore not only in cost savings but also in avoiding these penalties.

Prior research, including works by Ostling et al. (2017) and Enomoto et al. (2017), indicates a heightened vulnerability of diabetic patients to hospital readmissions. This trend can be attributed to the high incidence of diabetes, more so within hospitalized populations. A comprehensive exploration of factors influencing these readmission rates is thus crucial, as pinpointing and addressing these factors could diminish readmissions and alleviate their economic burden (Bhuvan et al., 2016).

Existing studies underscore the importance of thorough data preprocessing to achieve robust classifiers, including the management of outliers and imbalanced data, and the identification of high-quality training records versus noisy data. Duggal et al. (2016) focused on a limited data subset, using stringent criteria to qualify patient records for analysis, thereby excluding approximately 83% of the records. The largest reductions were due to the exclusion of "day-care admissions" (about 32%) and omitting records without a diabetes history (around 45%). This approach is expected to yield more reliable outcomes and applicable results for specific patient groups.

Our research seeks the most accurate classifiers for predicting patient readmissions, highlighting the role of data preprocessing in achieving robust classification. The limitations and focus of previous studies, such as those by Duggal et al. (2016), inform our methodology. Our approach is not restricted to a particular demographic; instead, we seek to formulate models that can accurately predict across diverse age groups and medical conditions.

To achieve our objectives, we follow a 2 folded approach: Initially, we will develop classifiers that can predict readmissions on a binary scale (readmission vs. no readmission). Subsequently, we aim to predict readmissions on a multiclass scale, differentiating between no readmission, readmission within 30 days, and readmission after 30 days.

Research by Sharma et al. (2019), employing a similar methodology with common independent variables for readmission prediction, suggests that factors such as time in hospital, discharge

disposition, number of diagnoses, medication count, initial diagnosis, age, gender, and insulin use are significant contributors to their model. We anticipate comparable findings in our study, given the similarities in variables and context.

The structure of this paper is as follows: We begin by exploring the dataset and detailing the data preprocessing steps, setting the stage for the application of various classifiers. We then address our first research objective by discussing the utilization and outcomes of different binary classifiers. This is followed by the development and assessment of various multiclass classifiers. Our aim is to provide insights beneficial not only to healthcare professionals but also to policymakers, thereby enriching the understanding of hospital readmissions, with a particular focus on diabetic patients.

2. Data Exploration and Preprocessing

2.1. Data exploration

This phase explored the raw data to discover initial insights and recognize interesting actionable patterns. In this study, a dataset of 71236 patient encounters from an USA hospital was used. Excluding the encounter ID that acts as index, the dataset has 30 variables, of which 2 are target variables. Among the 28 predictor variables, 18 are categorical and 10 numerical (refer to Annex 1). The categorical features exhibit a wide range of cardinality; for instance, 'change_in_meds_during_hospitalization' and 'prescribed_diabetes_meds' are binary, whereas 'primary_diagnosis', 'secondary_diagnosis', 'additional_diagnosis' and 'medication' contain a diverse array of unique values, with counts of 687, 699, 747 and 303, respectively (Annex 2).

There are no duplicate records, however it is interesting to notice that there are repeated 'patient_id' entries that indicate multiple records for some patients, we will analyze it further in the report.

Further the exploratory analysis, boxplots were used to visualize the distribution of the numerical variables (refer to Annex 3). This revealed that the scales and spreads of the variables are significantly different, emphasizing the necessity of standardization. This will allow the machine learning algorithm to fairly weight each feature without bias towards variables of higher magnitude. Complementarily, histograms were also used to supplement this analysis (refer to Annex 4), having showed that the majority of the considered variables are right-skewed. This fact can lead to model bias as these variables may distort the assumed relationship between the features and the target variable, particularly in models assuming data normality such as Logistic Regression. Additional examination will be done during preprocessing.

Additionally, when examining the distribution of our target variables, they revealed a disproportionate split. In the binary classification scenario, 'Yes' cases account for 11.2%, while 'No' cases dominate at 88.8%. As for the multiclass classification, the distribution is 11.2% for re-admissions within 30 days, 34.9% for those after 30 days, and a majority of 53.9% for 'No' readmissions (Annex 5). This significant imbalance in the dataset highlights the need for careful consideration in the model training process to ensure that the minority class is not overshadowed by the majority. Many machine learning algorithms assume that data is equally distributed across classes, so when this assumption is violated, models tend to be biased towards the majority class, often at the expense of correctly predicting the minority class. We will discuss it further in this report upon the implementation of techniques to handle it.

2.2. Data preprocessing

Given that medical data globally tends to be noisy, inconsistent, and incomplete, it is crucial to first refine this data effectively to make it suitable for predictive analysis. Data preprocessing techniques such as missing value imputation, variable encoding, and data balancing are pivotal in enhancing the precision of the predictive outcomes.

2.2.1. Data Preparation – Transformation & Cleaning

In the initial phase of data preprocessing unidentified data entries, denoted by the symbol '?', were replaced with 'NaN' to facilitate an accurate assessment of missing information within the dataset. This substitution revealed that 13 variables contained missing entries, as detailed in Annex 6.

An informed approach was taken to interpret and handle these missing values based on our understanding of the dataset's features. For the 'payer_code' variable, missing data was interpreted to indicate the absence of associated health insurance coverage. Consequently, 'NaN' values for this variable were replaced with 'No provider'. A similar rationale was applied to the variables 'glucose_test_result' and 'a1c_test_result'. The lack of recorded values for these tests was presumed to mean that the tests were not performed. Therefore, we replaced missing values with 'Not measured' to clearly denote unadministered tests. The same interpretative strategy was consistently employed for the variables 'primary_diagnosis', 'secondary_diagnosis', and 'additional_diagnosis'. We assumed that missing data in these fields suggested no requirement for a diagnosis during the encounters in question. To capture this, missing entries were substituted with 'Not necessary'.

Regarding the 'age' variable, which is categorized into age brackets and recorded as an object type, we transformed the data to reflect the mean age of each bracket. This conversion from categorical to numerical representation allows us to integrate this variable into our predictive models.

The transformation phase finished with the application of the log(x+1) transformation to the numerical variables that showed to be skewed. This transformation helps normalize distributions, stabilize variance, mitigate the influence of outliers and convert multiplicative relationships into additive ones, which are easier to model. The ‘+1’ ensures that the transformation can handle zero values, as the logarithm of zero is undefined.

In the data cleaning process, the objective was to remove records that could potentially distort the analysis. To this end, we excluded the 3 entries where the gender was listed as 'Unknown/Invalid'. Additionally, drawing insights from a comparable study, we also omitted records where the ‘discharge_disposition’ indicated hospice care (removal of 526 entries), as these patients are not typically subject to readmission, which is a known factor in advance.

Following an in-depth examination of the 'patient_id' field, we identified 10 559 unique patient IDs that appeared multiple times within the dataset. The frequency of these repeat encounters varied, with 7 037 patients recorded as having two visits, 2 013 patients with three visits, and at the higher end of the spectrum, one patient attending 33 times (detailed information in Annex 7). Taking this into account, we deliberated on whether to remove duplicate records of patients with more than one hospital visit, electing to retain only the first visit. The upside of doing it is that it enhances the validity of statistical analyses by ensuring each observation is independent. Additionally, machine learning models work better if they depend on independent observations to generalize patterns rather than memorizing specific patient histories, which can lead to overfitting. This approach also reduces computational load since it leads to a dataset with fewer entries. However, it is important to note that we might introduce bias to the study, if these patients differ systematically from those who visited the hospital only once. For instance, patients with chronic conditions might visit the hospital more frequently and excluding them could skew the results and omit patterns pertinent to readmissions. Before making a final decision, we assessed the impact on our model’s performance by contrasting

scenarios that included both single and multiple visits. The conclusion was to not remove duplicate ‘patient_id’ since our models were performing better when considering them.

The data cleaning process finished with the removal of two columns that don’t bring value to the analysis, ‘country’ since it only has one unique value, and ‘weight’ given the high percentage of missing values (around 97%).

2.2.2 Outlier Removal

To remove outliers, we individually analyzed boxplots for each numerical variable and established specific outlier thresholds, allowing for precise and tailored removal. This controlled and nuanced process led to the removal of 0.06% of records (detailed information on it can be found in Annex 8).

The removal of outliers was done before filling missing values, so the latter is based on a more representative dataset, leading to more reliable analysis.

2.2.3 Missing Values Imputation

Upon transformation of the variables as previously described, we assert that the residual missing data genuinely represents instances where information is absent. The specifics of these occurrences are detailed in Annex 9. To handle missing entries, a tailored approach was adopted for each variable, taking into consideration the proportion of missing data and the contextual understanding of the variable in question.

For the variable 'age', which exhibited a relatively minor 5% incidence of missing values, we opted to substitute the absent entries with the mean age, thus preserving the variable's central tendency. In the case of 'race', which had approximately 7% of its values missing, we assigned the label 'Unknown' to all missing entries. This treatment was consistently applied to the 'medical_specialty' variable as well, where a substantial 49% of the data was missing, acknowledging the indeterminate nature of the missing information. Regarding the variables 'admission_type', 'discharge_disposition', and 'admission_source', with missing data rates of about 5%, 3%, and 7% respectively, a proportional imputation method was employed. This technique involved distributing the missing values according to the existing frequency of each category within the respective variable, thereby maintaining the original distributional characteristics of the dataset.

This careful and considered approach to managing missing data underscores our commitment to maintaining the integrity of the dataset and ensuring that the outcomes derived from it are both robust and meaningful.

2.2.4 Feature Engineering

To be able to consider the categorical variables for predictive models, we employed encoding methodologies to transform qualitative information into numerical formats. To do so, we used a combination of four encoding techniques to handle categorical variables effectively.

Binary encoding was applied on ‘gender’ (having 1 for male and 0 for female), ‘payer_code’ (having 0 for cases without insurance and 1 for cases with insurance), ‘prescribed_diabetes_meds’ (having 1 for yes and 0 for no) and, ‘change_in_meds_during_hospitalization’ (having 1 for change and 0 for no change).

Simultaneously, label encoding was applied on both ‘glucose_test_result’ and ‘a1c_test_result’ given the fact that there is some ordinality associated with the categories of these variables. For the first feature we applied a glucose mapping in which entries that did not perform the test or that had normal results were assigned 1, entries with result above 200 were assigned with 2 and entries with result above 300 were assigned with 3, indicating a crescent level of glucose. For the second variable the exact same logic was applied.

To encode the remaining variables with exception of ‘medication’, the approach used was one-hot-encoding, hence each category of each of the variables started being represented as a binary column, indicating the presence or absence of a specific category. To address the issue of features with high cardinality of categories, before encoding them, we consolidated several categories into broader groups (detailed information on the grouping of categories for each feature can be found in Annex 10). This strategy effectively reduced the number of columns generated, mitigating the 'curse of dimensionality’, and preventing the creation of sparse matrices that could complicate the modeling process.

Lastly, frequency encoding was chosen for the ‘medications’ variable to handle its high cardinality of values. Instead of using OHE, which would have created numerous additional columns, frequency encoding replaces each category with its relative frequency within the dataset. This approach preserves the information about the prevalence of each medication category without expanding the feature space.

To finish the feature engineering, we created two variables, the first is named ‘service_utilization’ and it compiles the total number of visits in the year previous to the encounter by summing outpatient, inpatient and emergency visits in previous year. The second is named ‘has_weight’ and accounts if a patient was weighted or not. Although weight itself is not very useful because of the proportion of missing data, knowing if an encounter was weighted might be useful.

It is important to note that during the data preprocessing phase, the test data was processed parallel to training data to ensure consistent format and structure for model application. However, the test dataset did not undergo cleaning or outlier removal to preserve its integrity for real-world model evaluation. Missing values and categorical variables in test data were imputed and encoded based on the patterns learned from training data, preventing data leakage and maintaining the validity of the model's performance assessment.

3. Binary Classification

This chapter delves into the development of a binary classification model, pivotal in predicting patient readmissions in healthcare settings. Our focus is to develop a robust model capable of distinguishing patients at risk of readmission. We detail the intricate process of feature selection, emphasizing its significance in refining our model's predictive accuracy. Following this, we explore the nuances of model training and evaluation, utilizing advanced techniques to ensure reliability and efficiency. The summary of our efforts is represented in a comprehensive analysis, where we interpret the model's performance and its implications in the context of patient readmission. This narrative offers a deep dive into the binary classification's role and impact within healthcare analytics.

Feature Selection

Prior to moving to the model development for our binary classification, it's crucial to conduct a feature selection process post data preprocessing. The advantages of this step are that it allows the algorithm to train quicker, it reduces the complexity of the model and makes it simpler to analyze the results. Moreover, it helps in mitigating overfitting and can potentially improve the accuracy of the model by choosing the most informative variables. Before stepping into the different methods, we proceeded to the normalization of the dataset using MinMax technique.

To start, we verified the absence of univariate variables by confirming non-zero variance across all features. Univariate variables have constant values across all entries and therefore don't contribute to the predictive power of models. Moving forward, we assessed the correlation among non-binary numerical features to identify multicollinearity and understand how much variables were associated with each other. By analyzing the correlation map (Annex 11), as anticipated, 'service_utilization' showed a relatively high correlation with outpatient, inpatient, and emergency visits in the previous year variables, to handle it the latter variables will ideally not be considered. Subsequently, employing Recursive Feature Elimination with a Logistic Regression base estimator, we identified 22 optimal features (detailed information on these features is available in Annex 12). We also applied Lasso Regression, which deselects features with a zero coefficient, concluding with 44 features to retain and 9 to exclude (for specifics on the retained features, refer to Annex 13). Additionally, the chi-square was used for categorical data and 'patient_id' showed not to be an important predictor, and therefore it will not be considered.

Given the different results RFE and Lasso provide, our approach to come to a conclusion was to combine both, hence variables that are insignificant according to both methods (7 variables – Annex 14) would not be selected. However, we noticed that according to Lasso 'inpatient_visits_previous_year' is a more important predictor than 'service_utilization' (refer to annex 13) and therefore we decided to keep the first instead. This way, we ended up with 43 following variables. Detailed information on the final selected features can be found in Annex 15.

Model Creation and Evaluation

With the preprocessing phase complete, our data is ready for the critical stage of training binary classification models. Our methodology encompasses a diverse array of models. This heterogenous mix includes Logistic Regression, representing linear models; a simple neural network in the form of a Perceptron; Stochastic Gradient Descent for iterative optimization; Decision Tree; and ensemble learners like Random Forest and AdaBoost, each offering unique strengths in handling complex patterns in data.

To optimize these models, we employed grid search, a methodical approach for hyperparameter tuning, ensuring that each model operates at its peak efficiency. For model evaluation we delve into a multi-metric analysis, employing accuracy and recall, with a primary focus on the F1 score. Performance measures to all the models can be found in appendix 18.

Decision Trees tend to overfit unless pruned appropriately, generally leading to high accuracy. Previous research in a similar context of readmission classification, such as the work by Sharma et al. (2019), found Random Forest to be the most effective predictor when prioritizing accuracy. Conversely, AdaBoost operates on the principle of boosting a series of weak learners. Individually, these learners may lack high accuracy, but collectively, they contribute to a stable and robust model. This ensemble technique, therefore, provides a counterbalance to the high variance of individual learners, culminating in consistent performance.

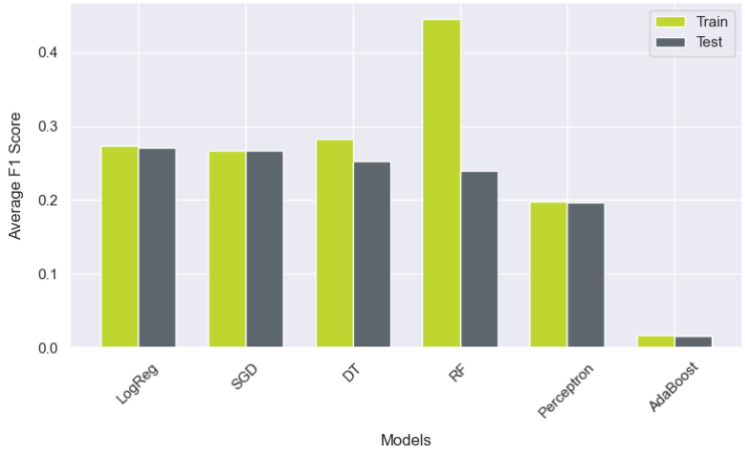


Fig. 1: Average F1 Scores for Train and Test Sets Across Different Models (binary classification).

Figure 1 shows the average F1 Score for the test and training data of each model created. The score represents the average of a 5-fold grid search calculation. With this approach we find as highest performing model Logistic Regression with a F1 score of 0.27.

Using a tuned Random Forest classifier, Figure 2 highlights key features identified as most influential. This aligns with Sharma et al. (2019), where 'time in hospital', 'number of medications', 'number of diagnoses', and 'medication type' are consistently deemed crucial. The overlap in both studies confirms the significance of these variables as robust predictors, crucial to the models' predictive performance.

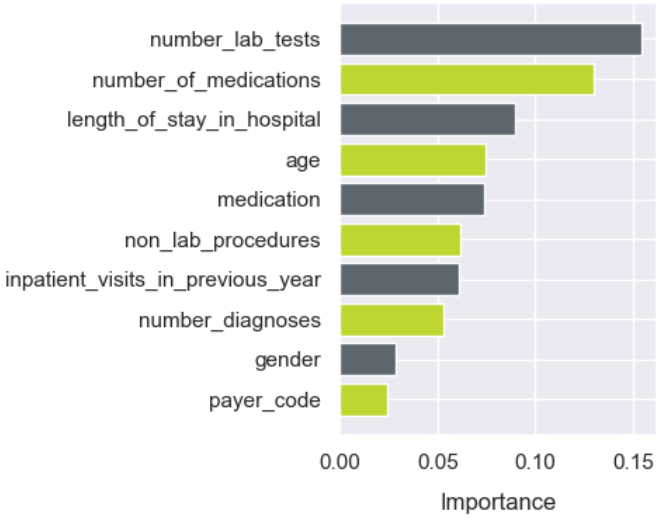


Fig. 2: Most Important Features - Random Forest.

4. Multiclass Classification

This chapter articulates the methodology employed in training and assessing multiclass models which predict the readmission of a patient with the classes 'no readmission', 'readmission within 30 days after initial discharge', and 'readmission after 30 days of the initial discharge'. Compared to the binary target, this approach thereby aims to yield more nuanced conclusions, facilitating a more granular differentiation among patient cases. Such differentiation is pivotal for enhancing healthcare management and reducing hospitalization costs.

This chapter is structured as follows: Initially the exploration of the target variable and necessary preprocessing steps are described. Extending the foundational concepts from previous chapters, this segment delves into a detailed discussion of the selected machine learning algorithms employed for the application and analysis of multiclass classification. The chapter culminates with a comprehensive evaluation of the models' performance, emphasizing their implications for patient readmission prediction.

Data Exploration

We start by checking basic information of our dataset and performing some data visualization of our independent variables and target. Being the initial dataset the same as the one used for the Binary, our data Exploration, Manipulation, and outlier removal are the same as described for Binary Classification.

Feature Selection

Mirroring the feature selection process from binary classification, we adopted similar methods to pinpoint key features for multiclass classification. This strategy of dimensionality reduction not only fastened model computations but also improved data density and minimized noise, thus enhancing the model's predictive accuracy.

We focused on preserving the dataset's integrity, thus we methodically eliminated univariate variables with negligible predictive impact. We also addressed multicollinearity, particularly with the 'service_utilization' feature, to improve data robustness. The integration of Recursive Feature Elimination and Lasso Regression was pivotal in refining our feature set, ensuring the inclusion of only the most pertinent variables.

In conclusion, this process culminated in a synthesis of insights from both RFE and Lasso analyses. We identified 49 features as critical for effective model training (with specifics documented in annex 16). This selection of features lays the foundation for the robustness and reliability of our multiclass classification models.

Model Creation and Evaluation

As in our binary approach, we developed several models on the dataset and tuned the hyperparameters with the application of grid search. In addition to the models used in the binary classification we also trained a Ridge Classifier. With grid search it's always a trade of between giving the biggest possible value range for the grid but consider the implication on the calculation time and the potential use of the additional options of the bigger value range. As for our limited capacities we went with a moderate approach, using multiple values, covering a rather broad then granular range.

In the dataset used for this study, it's evident that the distribution of the target class is imbalanced (Annex 5). Specifically, only a tenth of the instances belong to the 'readmission within 30 days after

initial discharge' class. Addressing this imbalance consequently, we have decided to employ the macro variant of the f1 score for model evaluation. This approach ensures equitable treatment of each class by individually calculating their F1 scores and subsequently averaging these scores across the entirety of the class spectrum. This methodology effectively mitigates the inherent imbalance present in the target variable.

Compared to other implementations of the f1 score, the macro approach attributes uniform significance to each class. This approach aligns seamlessly with our study objectives, as it is crucial to accord equal importance to all three classes to comprehensively understand the outcomes. Both – patients who are readmitted within and those after 30 days after initial discharge – need thorough examination. The former one due to the urgency and the potential for early intervention, while the latter merits analysis to detect possibly chronic underlying causes. Equally critical are the instances where readmission does not occur, as these cases provide a foundational comparison to identify effective treatments and interventions leading to positive patient outcomes.

The top-performing models identified in our analysis are the Random Forest, which demonstrates a propensity for overfitting on the training data, along with Logistic Regression and a straightforward Decision Tree. Detailed metrics for each of these models are presented in Figure 3. Using the tuned instance of a random forest classifier we can assess the most important features which are available in annex 17; performance measures to all the models can be found in appendix 19. Compared to the results of the binary classification we can see that not only the order of the top ten variables changed, but also other features are at the top. The feature "average pulse (in bpm)" is not even considered as significant for the binary classification and was not considered to train the models. But it is the most important separator for the multiclass implementation of random forest.

As shown in the chapter about the binary classification we can see that also predictor's data is quite imbalanced. In this sense, models which are computed with the consideration of imbalanced data perform a lot better than models without transformation. The best classifier for our multiclass target is random forest with a F1 score of 0.42.

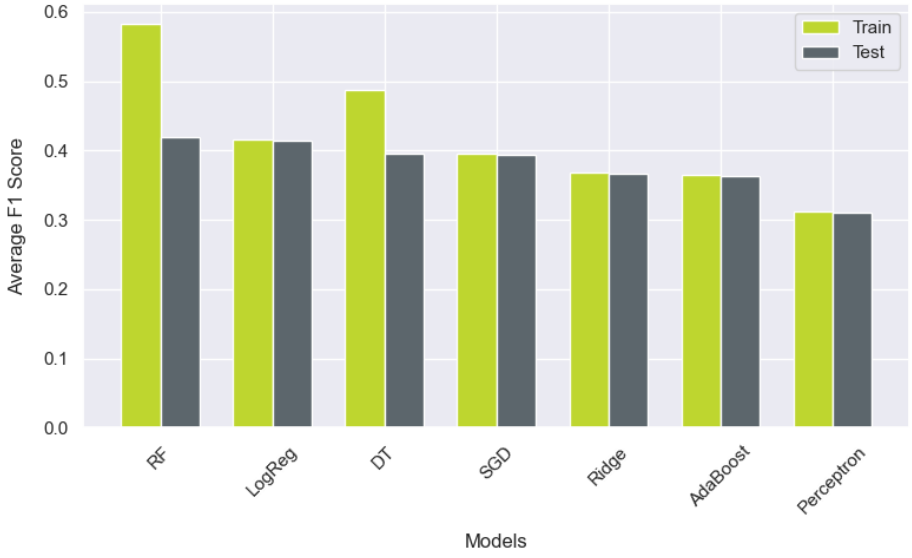


Fig. 3: Average F1 Scores for Train and Test Sets Across Different Models (multi class classification).

5. Conclusion

The study focuses on understanding factors contributing to hospital readmissions, as hospital readmissions represent a significant cost factor in healthcare. The study aims to develop predictive models using machine learning techniques.

We created several machine learning models that can analyze the key factors contributing to hospital readmissions to identify the most accurate model for predicting them.

We chose Logistic Regression as the binary classifier for predicting hospital readmissions in patients, supported by its superior performance measured by the F1 score, which balances precision and recall. The interpretative nature of Logistic Regression is crucial for healthcare professionals and decision-makers, providing clear insights into risk factors. This choice aligns with the central objective of the study: not only accurately predicting readmissions but also increasing awareness and contributing to effective intervention strategies in healthcare management.
For the multi class classification, our study identified the Random Forest model as the superior predictor, particularly when emphasizing the macro F1 score. Employing this model illuminates the significance of chosen features, directly feeding into the practical application of reducing hospital readmission rates. A key point of investigation centers around the feature “average pulse (in bpm)”, identified as a pivotal predictor. It invites further exploration to discern if a patient's pulse rate transcends its role as a mere statistical indicator in machine learning models and delves into underlying causes that could be actionable. Such insights might unveil strategies not only for predictive accuracy but also for tangible interventions, potentially leading to a decrease in readmission rates. This holistic approach, merging data-driven insights with clinical applications, is pivotal in enhancing patient outcomes and operational efficiencies.

During the project, we faced some limitations in computational power. We couldn't perform Recursive Feature Elimination (RFE) with the base estimator Support Vector Classifier due to time restrictions. Additionally, we did not consider primary diagnosis, secondary diagnosis, and additional diagnoses due to a lack of knowledge on how to handle this information. Imbalance of the data and especially in the target variables could also be treated already in the preprocessing by methods like SMOTE. This could lead to other predictors and might result in different predictive measures. Consequently, we believe there is room for improvement through more detailed data research.

In optimizing our patient readmission prediction models, it’s an option to assess the effects of retaining duplicate patient records and explore additional features for enhanced performance. Dynamic evaluation strategies could be implemented to adapt to evolving healthcare dynamics, and external validation across diverse settings will validate the models' robustness. This comprehensive approach aims to refine accuracy and applicability, contributing to advancements in patient care and resource optimization.

The final model can be used by healthcare organizations to identify patients at high risk for readmission and take proactive measures to prevent it, aligning with efforts to provide high-quality care and maintain financial viability through advancements in healthcare management, cost reduction, and improvement in patient care.

References

Bhuvan, M. S., Kumar, A., Zafar, A., & Kishore, V. (2016). Identifying Diabetic Patients with High Risk of Readmission. https://arxiv.org/abs/1602.04257v1

Duggal, R., Shukla, S., Chandra, S., Shukla, B., & Khatri, S. K. (2016). Predictive risk modelling for early hospital readmission of patients with diabetes in India. International Journal of Diabetes in Developing Countries, 36(4), 519–528. https://doi.org/10.1007/S13410-016-0511-8/TABLES/3

Enomoto, L. M., Shrestha, D. P., Rosenthal, M. B., Hollenbeak, C. S., & Gabbay, R. A. (2017). Risk factors associated with 30-day readmission and length of stay in patients with type 2 diabetes. Journal of Diabetes and Its Complications, 31(1), 122–127. https://doi.org/10.1016/j.jdiacomp.2016.10.021

Hospital Readmissions Reduction Program (HRRP) | CMS. (n.d.). Retrieved December 15, 2023, from https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp

Ostling, S., Wyckoff, J., Ciarkowski, S. L., Pai, C.-W., Choe, H. M., Bahl, V., & Gianchandani, R. (2017). The relationship between diabetes mellitus and 30-day readmission rates. Clinical Diabetes and Endocrinology 2017 3:1, 3(1), 1–8. https://doi.org/10.1186/S40842-016-0040-X

Sarthak, Shukla, S., & Prakash Tripathi, S. (2021). Embpred30: Assessing 30-days readmission for diabetic patients using categorical embeddings. Advances in Intelligent Systems and Computing, 1168, 81–90. https://doi.org/10.1007/978-981-15-5345-5_7

Sharma, A., Agrawal, P., Madaan, V., & Goyal, S. (2019). Prediction on diabetes patient's hospital readmission rates. Proceedings of the Third International Conference on Advanced Informatics for Computing Research, 1–5. https://doi.org/10.1145/3339311.3339349

Appendix

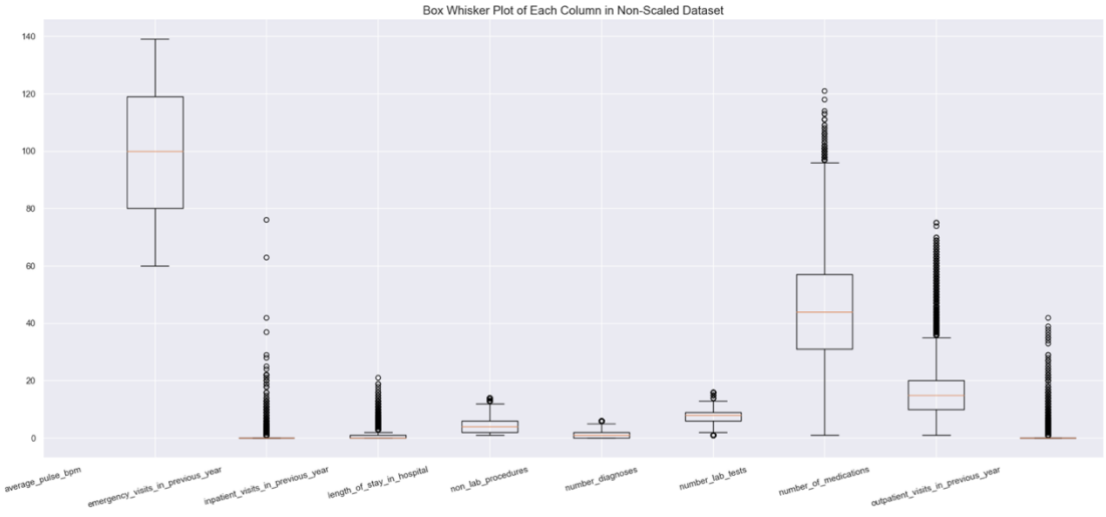
Annex 1 – Table that compiles all the variables of the initial dataset, their type, and their classification according to the information they provide.

Variables Categories	Variables Names	Details
Identifiers	encounter_id, patient_id	ID
Demographic	country, race and gender	Categorical Nominal
	age, weight	Categorical Ordinal (classes)
Health Insurance	payer_code	Categorical Nominal
Past healthcare utilization	outpatient_visits_in_previous_year	Numerical
	emergency_visits_in_previous_year	
	inpatient_visits_in_previous_year	
Patient	admission_type, admission_source, discharge_disposition, medical_specialty	Categorical Nominal
	average_pulse_bpm, number_lab_tests, non_lab_procedures, number_of_medications, number_diagnoses, length_of_stay_in_hospital	Numerical
	primary_diagnosis, secondary_diagnosis, additional_diagnosis	Categorical (ICD9 codes)
Clinical Tests Results	glucose_test_result, a1c_test_result	Categorical Ordinal
Medication details	medication	Categorical nominal (lists of medications)
	prescribed_diabetes_meds, change_in_meds_during_hospitalization	Categorical Nominal (binary)

Annex 2 – Number of unique values of each variable.

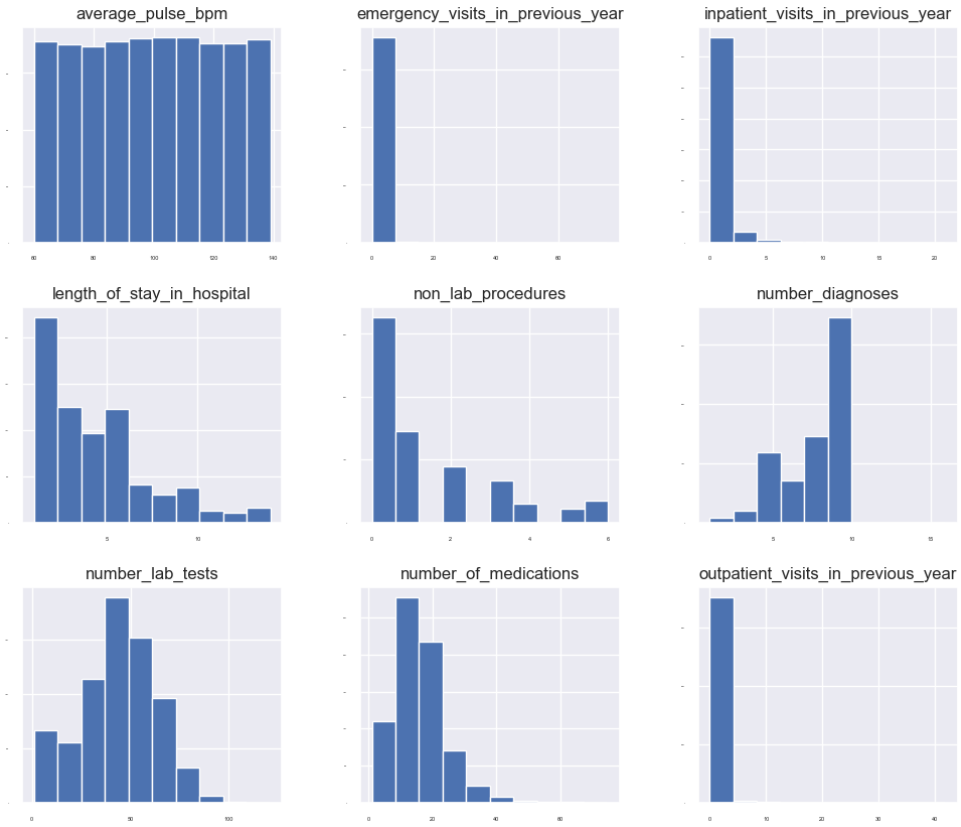
Variables	# Categories
encounter_id	71236
country	1
patient_id	53985
race	6
gender	3
age	10
weight	10
payer_code	18
outpatient_visits_in_previous_year	38
emergency_visits_in_previous_year	30
inpatient_visits_in_previous_year	21
admission_type	7
medical_specialty	69
average_pulse_bpm	80
discharge_disposition	25
admission_source	16
length_of_stay_in_hospital	14
number_lab_tests	114
non_lab_procedures	7
number_of_medications	72
primary_diagnosis	687
secondary_diagnosis	699
additional_diagnosis	747
number_diagnoses	16
glucose_test_result	3
a1c_test_result	3
change_in_meds_during_hospitalization	2
prescribed_diabetes_meds	2
medication	303
readmitted_binary	2
readmitted_multiclass	3

Annex 3 – Boxplots of each numerical variable in non-scaled environment. Here it is possible to observe the different spreads of the considered variables, which emphasizes the necessity to standardize these variables.

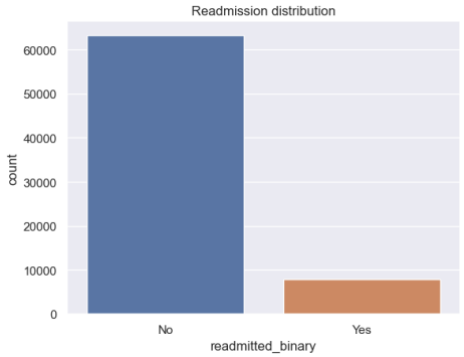


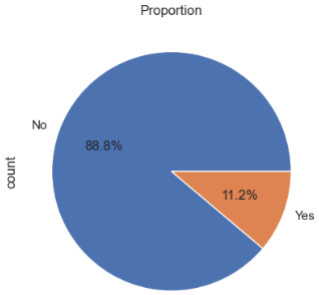
Annex 4 – Histograms of each numerical variable. Here it is possible to visualize the different distribution of values for each considered variable.

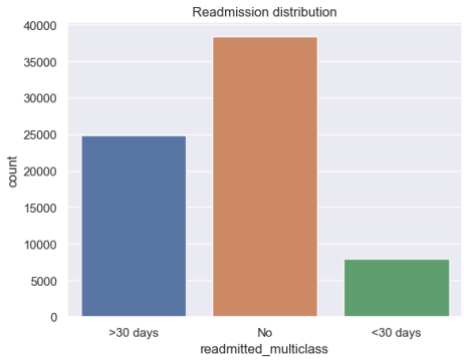
- **Average Pulse BPM**: This histogram shows a fairly uniform distribution, suggesting that the average pulse rate of patients doesn't vary widely.
- **Inpatient, Outpatient and Emergency Visits in Previous Year**: The distribution of these three variables is heavily right-skewed, indicating that most patients had few or no visits of the correspondent type in the previous year.
- **Length of Stay in Hospital**: The histogram is right-skewed with a long tail, suggesting that most patients had a short stay in the hospital.
- **Non-Lab Procedures**: The data is right-skewed, indicating that most patients underwent few or no non-lab procedures.
- **Number of Diagnoses**: The distribution shows a bimodal pattern, with peaks around 5 and 9 diagnoses. This could indicate common patterns in the number of diagnoses patients receive, where one set has a relatively simple health profile with few diagnoses and the other has a more complex health profile with multiple diagnosis.
- **Number of Lab Tests**: The histogram shows a right-skewed distribution with most patients having few lab tests. The peak at the lower end suggests that a standard set of lab tests is common for most patients.
- **Number of Medications**: This distribution is somewhat bell-shaped but skewed to the right, indicating that while there is a common range of medications prescribed, there are patients who are prescribed a larger number of medications.

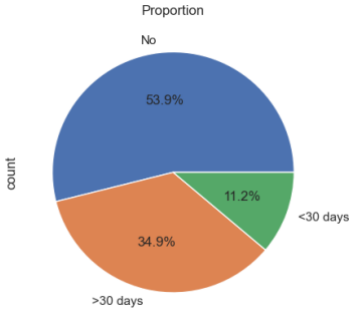


Annex 5 – Readmission distribution for both binary classification and multiclass classification.









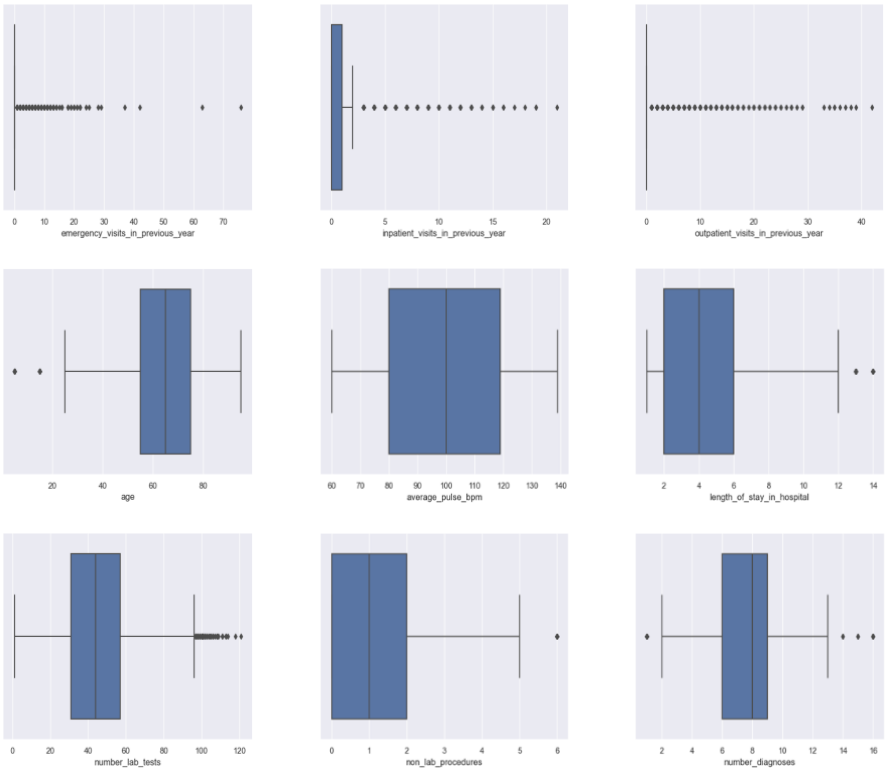
Annex 6 – Percentage of missing values of each of the variables. Variables that are not present in the below table don't have missing data.

Variables	% of Missing Values
race	7.117%
age	4.993%
weight	96.847%
payer_code	39.588%
admission_type	5.202%
medical_specialty	49.023%
discharge_disposition	3.636%
admission_source	6.623%
primary_diagnosis	0.022%
secondary_diagnosis	0.368%
additional_diagnosis	1.415%
glucose_test_result	94.823%
a1c_test_result	83.273%

Annex 7 – The below table details the frequency of patient visits to the hospital.

Nr. of Visits per patient	Nr. of patients
1	43426
2	7037
3	2013
4	809
5	340
6	158
7	79
8	45
9	20
10	17
11	14
12	8
13	3
14	6
15	6
17	1
18	1
19	1
33	1

Annex 8 – Individual boxplot of each numerical variable to analyze which outliers would be removed. To do so, a nuanced approach was taken.



Annex 9 – Percentage of actual missing values (absent information of each of the variables).

Variables	% of Missing Values
race	7.117%
age	4.993%
weight	96.847%
admission_type	5.202%
medical_specialty	49.023%
discharge_disposition	3.636%
admission_source	6.623%

Annex 10 – Explanation on the grouping of categories for the different variables:

The grouping of categories for the **'admission_type'** variable was made based on the goal to group the less frequent categories, hence 'Trauma Center', 'Newborn', 'Not Mapped', 'Not Available' were grouped together into 'Other'.

The grouping of categories for the **'admission_source'** variable was made based on the goal to group the less frequent categories, hence ' Sick Baby', 'Normal Delivery', ' Extramural Birth', 'Not Available', ' Transfer from Ambulatory Surgery Center', ' Transfer from critial access hospital', ' Transfer from hospital inpt/same fac reslt in a sep claim', ' Court/Law Enforcement' were grouped together into 'Other'.

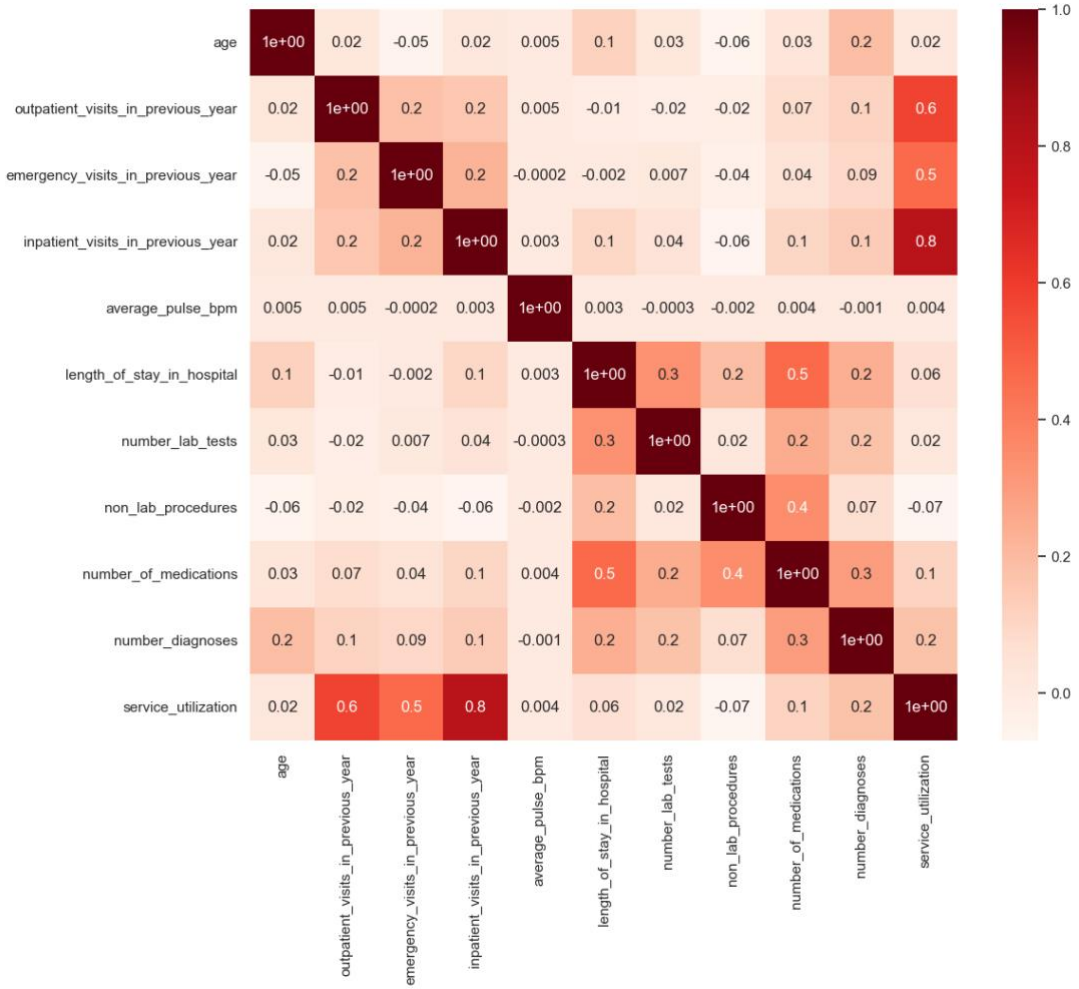
The grouping of categories for the **'medical_specialty'** variable was made based on fitting each value into generalized categories, aiming at reducing the granularity of medical specialties.

- **Pediatrics**: This group includes specialties focused on children's health, such as 'Pediatrics', 'Neonatology', and care for 'Newborns'.
- **Surgery**: It captures all surgical-related specialties, including general 'Surgery' and more specialized fields like 'Orthopedics-Reconstructive' and 'Podiatry'.
- **General Practice**: This category includes primary care fields like 'Family/General Practice', 'Internal Medicine', and 'Hospitalist', which deal with a broad spectrum of diseases and health issues, and 'PhysicianNotFound' for unspecified practices.
- **Cardiology**: Encompasses all heart-related specialties, including general 'Cardiology' and surgical subspecialties like 'Surgery-Cardiovascular' and 'Surgery-Cardiovascular/Thoracic', as well as 'Cardiology-Pediatric'.
- **Psychiatry/Psychology**: This combines mental health specialties, including 'Psychiatry', its child/adolescent branch, and 'Psychology'.
- **Oncology/Hematology**: Groups together specialties dealing with cancer ('Oncology') and blood disorders ('Hematology'), including gynecologic oncology.
- **Neurology**: This category includes 'Neurology' and its surgical counterpart 'Surgery-Neuro', as well as pediatric neurology, covering disorders of the nervous system.
- **Gastroenterology**: Stands alone as its own category due to the specialized nature of this field, which focuses on the digestive system.
- **Nephrology**: Also stands alone, specializing in kidney functions and diseases.
- **Orthopedics**: Isolated as a separate category, it deals with the musculoskeletal system.
- **Other Specialties**: Any specialty not covered by the above categories is classified here to include all other medical fields.

The grouping of categories for the **'discharge_disposition'** variable was made based on the level of care they require after leaving the hospital. This approach simplifies the variety of individual discharge statuses into broader and more manageable groups.

- **Discharged to home**: This is assigned when the disposition exactly matches 'Discharged to home', indicating the patient was released to their home without further immediate institutional care.
- **Discharged to Special Nursing Facility**: This category is for dispositions containing 'Discharged/transferred to SNF' or mentioning 'ICF', implying that the patient requires ongoing care but at a less intensive level than a hospital.
- **Discharged to home with health service**: Assigned when the disposition includes the phrase 'home with home health service', suggesting the patient needs additional medical support services at home.
- **Expired**: Used when the disposition includes the term 'Expired', indicating that the patient passed away.
- **Transferred to another hospital/institution**: Chosen for dispositions containing 'hospital' or 'care institution', which means the patient was transferred to another medical facility for further care or treatment.
- **Left Against Medical Advice**: Selected when 'AMA' is part of the disposition, indicating the patient left the hospital on their own decision.
- **Transferred to Long Term Care**: This is for dispositions that include 'long term care hospital' or 'Medicare approved swing bed', indicating the patient is moving to a facility that provides long-term medical care.
- **Transferred to Psychiatric Facility**: Used when the disposition states 'psychiatric hospital', signifying the patient has been moved to a facility specializing in psychiatric care.
- **Other**: This is a catch-all category for any dispositions that do not explicitly fit the descriptions of the above categories.

Annex 11 – Correlation map between non-binary numerical variables. Overall there are no significant correlations between the variables, with exception of 'service_utilization' that shows relatively high correlations with the variables used to create it, especially 'inpatient_visits_in_previous_year'.

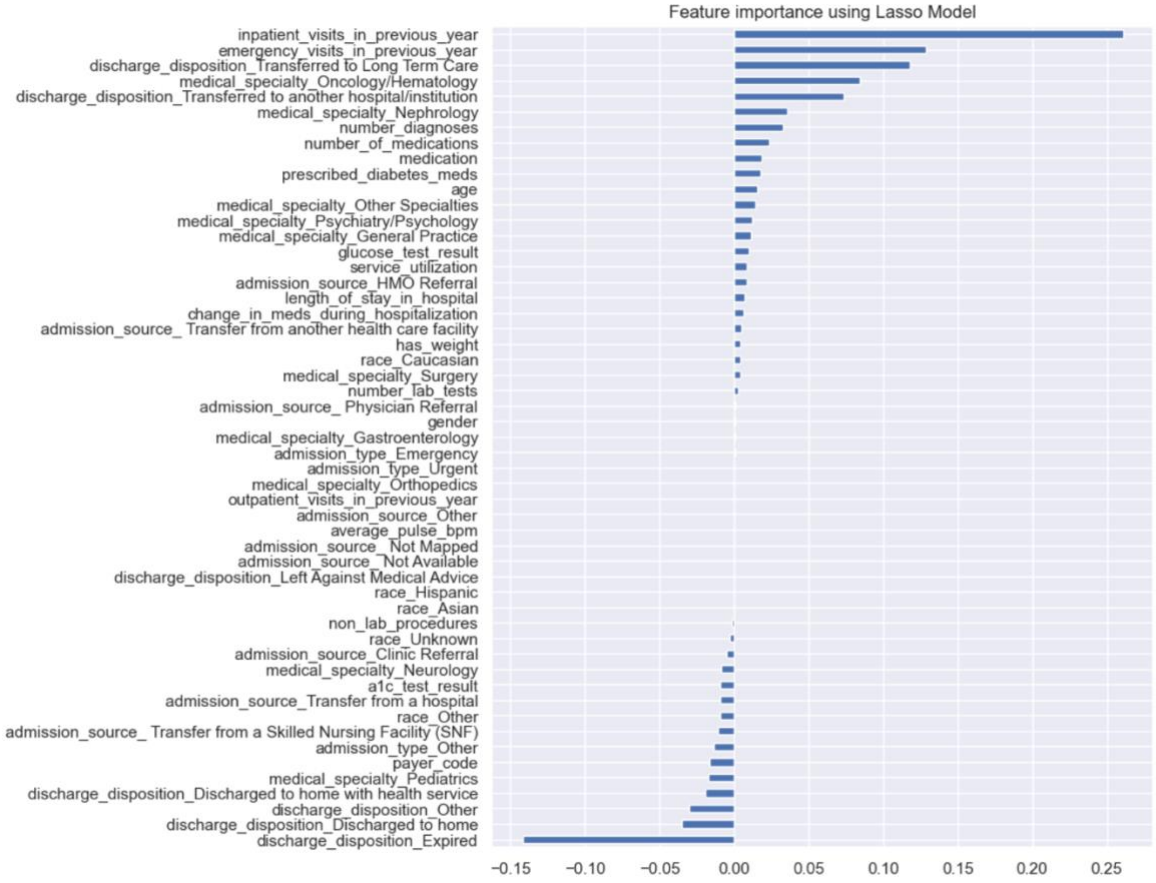


Annex 12 – Features selected using RFE with base estimator Logistic Regression.

gender	False
age	True
payer_code	False
outpatient_visits_in_previous_year	True
emergency_visits_in_previous_year	True
inpatient_visits_in_previous_year	True
average_pulse_bpm	False
length_of_stay_in_hospital	False
number_lab_tests	False
non_lab_procedures	False
number_of_medications	True
number_diagnoses	True
glucose_test_result	False
a1c_test_result	False
change_in_meds_during_hospitalization	False
prescribed_diabetes_meds	True
medication	True
race_Asian	False
race_Caucasian	False
race_Hispanic	False
race_Other	True
race_Unknown	False
admission_type_Emergency	False
admission_type_Other	False
admission_type_Urgent	False
medical_specialty_Gastroenterology	False
medical_specialty_General Practice	False
medical_specialty_Nephrology	True
medical_specialty_Neurology	True
medical_specialty_Oncology/Hematology	True
medical_specialty_Orthopedics	False
medical_specialty_Other Specialties	False
medical_specialty_Pediatrics	True
medical_specialty_Psychiatry/Psychology	False
medical_specialty_Surgery	False
discharge_disposition_Discharged to home	True
discharge_disposition_Discharged to home with health service	True
discharge_disposition_Expired	True
discharge_disposition_Left Against Medical Advice	False
discharge_disposition_Other	True
discharge_disposition_Transferred to Long Term Care	True
discharge_disposition_Transferred to another hospital/institution	True
admission_source_ Not Available	False
admission_source_ Not Mapped	True
admission_source_ Physician Referral	False
admission_source_ Transfer from a Skilled Nursing Facility (SNF)	False
admission_source_ Transfer from another health care facility	False
admission_source_Clinic Referral	False
admission_source_HMO Referral	True
admission_source_Other	False
admission_source_Transfer from a hospital	False
service_utilization	True
has_weight	False

Annex 13 – Features selected using Lasso. This method eliminates variables that have coefficient equal to zero.

```
gender                                                          0.001039
age                                                             0.016039
payer_code                                                     -0.016733
outpatient_visits_in_previous_year                             -0.000000
emergency_visits_in_previous_year                               0.128670
inpatient_visits_in_previous_year                               0.261084
average_pulse_bpm                                              -0.000000
length_of_stay_in_hospital                                      0.007037
number_lab_tests                                                0.002539
non_lab_procedures                                             -0.001764
number_of_medications                                           0.023745
number_diagnoses                                                0.033191
glucose_test_result                                             0.009872
a1c_test_result                                                -0.009122
change_in_meds_during_hospitalization                           0.006082
prescribed_diabetes_meds                                        0.017536
medication                                                      0.018594
race_Asian                                                      0.000000
race_Caucasian                                                  0.004334
race_Hispanic                                                   0.000000
race_Other                                                     -0.009410
race_Unknown                                                   -0.003200
admission_type_Emergency                                        0.000735
admission_type_Other                                           -0.013264
admission_type_Urgent                                           0.000280
medical_specialty_Gastroenterology                              0.000740
medical_specialty_General Practice                              0.011785
medical_specialty_Nephrology                                    0.035670
medical_specialty_Neurology                                    -0.008729
medical_specialty_Oncology/Hematology                           0.084467
medical_specialty_Orthopedics                                   0.000000
medical_specialty_Other Specialties                             0.014219
medical_specialty_Pediatrics                                   -0.017510
medical_specialty_Psychiatry/Psychology                         0.012188
medical_specialty_Surgery                                       0.004062
discharge_disposition_Discharged to home                       -0.034749
discharge_disposition_Discharged to home with health service   -0.018966
discharge_disposition_Expired                                  -0.141934
discharge_disposition_Left Against Medical Advice               0.000000
discharge_disposition_Other                                    -0.030258
discharge_disposition_Transferred to Long Term Care             0.118052
discharge_disposition_Transferred to another hospital/institution  0.073447
admission_source_ Not Available                                 0.000000
admission_source_ Not Mapped                                    0.000000
admission_source_ Physician Referral                            0.001067
admission_source_ Transfer from a Skilled Nursing Facility (SNF) -0.010713
admission_source_ Transfer from another health care facility    0.004820
admission_source_Clinic Referral                               -0.005252
admission_source_HMO Referral                                   0.008250
admission_source_Other                                         -0.000000
admission_source_Transfer from a hospital                      -0.009350
service_utilization                                             0.008301
has_weight                                                      0.004451
dtype: float64
Lasso picked 44 variables and eliminated the other 9 variables
```



Annex 14 – Insignificant features according to both RFE and Lasso:

- 'discharge_disposition_Left Against Medical Advice'
- 'medical_specialty_Orthopedics'
- 'race_Hispanic'
- 'admission_source_Other'
- 'average_pulse_bpm'
- 'race_Asian'
- 'admission_source_ Not Available'

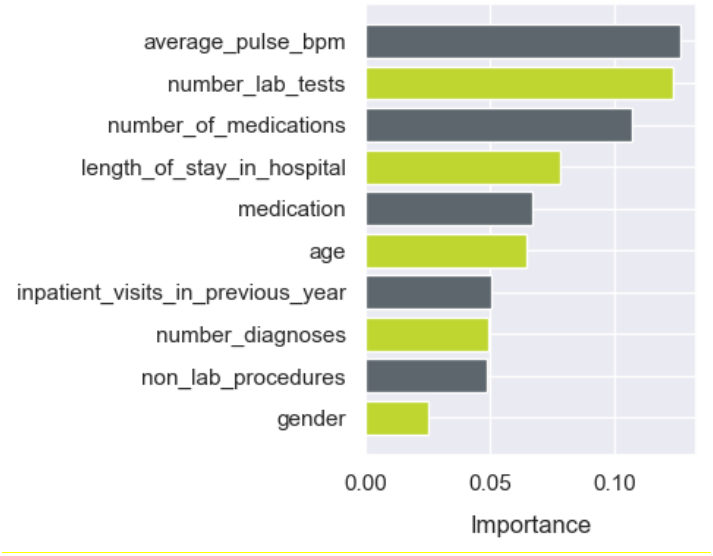
Annex 15 – Final selected features (binary classification):

- 'gender'
- 'age'
- 'payer_code'
- 'inpatient_visits_in_previous_year'
- 'length_of_stay_in_hospital'
- 'number_lab_tests'
- 'non_lab_procedures'
- 'number_of_medications'
- 'number_diagnoses'
- 'glucose_test_result',
- 'a1c_test_result'
- 'change_in_meds_during_hospitalization'
- 'prescribed_diabetes_meds'
- 'medication'
- 'race_Caucasian'
- 'race_Other'
- 'race_Unknown'
- 'admission_type_Emergency'
- 'admission_type_Other'
- 'admission_type_Urgent'
- 'medical_specialty_Gastroenterology'
- 'medical_specialty_General Practice'
- 'medical_specialty_Nephrology'
- 'medical_specialty_Neurology'
- 'medical_specialty_Oncology/Hematology'
- 'medical_specialty_Other Specialties'
- 'medical_specialty_Pediatrics'
- 'medical_specialty_Psychiatry/Psychology'
- 'medical_specialty_Surgery'
- 'discharge_disposition_Discharged to home'
- 'discharge_disposition_Discharged to home with health service'
- 'discharge_disposition_Expired'
- 'discharge_disposition_Other'
- 'discharge_disposition_Transferred to Long Term Care'
- 'discharge_disposition_Transferred to another hospital/institution',
- 'admission_source_ Not Mapped'
- 'admission_source_ Physician Referral'
- 'admission_source_ Transfer from a Skilled Nursing Facility (SNF)'
- 'admission_source_ Transfer from another health care facility'
- 'admission_source_Clinic Referral'
- 'admission_source_HMO Referral',
- 'admission_source_Transfer from a hospital'
- 'has_weight'

Annex 16 – Final selected features (multi class classification):

- 'gender'
- 'age'
- 'payer_code'
- 'inpatient_visits_in_previous_year'
- 'average_pulse_bpm'
- 'length_of_stay_in_hospital'
- 'number_lab_tests'
- 'non_lab_procedures'
- 'number_of_medications'
- 'number_diagnoses'
- 'glucose_test_result'
- 'a1c_test_result'
- 'change_in_meds_during_hospitalization'
- 'prescribed_diabetes_meds'
- 'medication'
- 'race_Asian'
- 'race_Caucasian'
- 'race_Hispanic'
- 'race_Other'
- 'race_Unknown'
- 'admission_type_Emergency'
- 'admission_type_Other'
- 'admission_type_Urgent'
- 'medical_specialty_Gastroenterology'
- 'medical_specialty_General Practice'
- 'medical_specialty_Nephrology'
- 'medical_specialty_Neurology'
- 'medical_specialty_Oncology/Hematology'
- 'medical_specialty_Orthopedics'
- 'medical_specialty_Other Specialties'
- 'medical_specialty_Pediatrics'
- 'medical_specialty_Surgery'
- 'discharge_disposition_Discharged to home'
- 'discharge_disposition_Discharged to home with health service'
- 'discharge_disposition_Expired'
- 'discharge_disposition_Left Against Medical Advice'
- 'discharge_disposition_Other'
- 'discharge_disposition_Transferred to Long Term Care'
- 'discharge_disposition_Transferred to another hospital/institution'
- 'admission_source_ Not Available'
- 'admission_source_ Not Mapped'
- 'admission_source_ Physician Referral'
- 'admission_source_ Transfer from a Skilled Nursing Facility (SNF)'
- 'admission_source_ Transfer from another health care facility'
- 'admission_source_Clinic Referral'
- 'admission_source_HMO Referral'
- 'admission_source_Other'
- 'admission_source_Transfer from a hospital'
- 'has_weight'

Annex 17 – Feature Importance for Multiclass Classification using Random Forest:



Annex 18 – Performance metrics for binary classification:

	Time	avg_f1_train	avg_f1_test	avg_acc_train	avg_acc_test	avg_rec_train	avg_rec_test
LogReg	8.543663	0.272762	0.27093	0.655819	0.655833	0.572005	0.572131
Perceptron	5.544535	0.19763	0.196878	0.594931	0.594931	0.520235	0.520235
SGD	11.55025	0.267436	0.266593	0.67248	0.690541	0.516852	0.548156
RF	614.141273	0.44486	0.239547	0.680102	0.680668	0.537045	0.53654
DT	2.88465	0.282681	0.25222	0.631309	0.631295	0.578459	0.578459
AdaBoost	37.122499	0.017592	0.016355	0.887592	0.887592	0.010602	0.010602

Annex 19 – Performance metrics for multi class classification:

	Time	avg_f1_train	avg_f1_test	avg_acc_train	avg_acc_test	avg_rec_train	avg_rec_test
LogReg	12.576073	0.415424	0.413883	0.485779	0.48585	0.458064	0.458053
Ridge	0.181495	0.367804	0.367448	0.578062	0.578062	0.396304	0.396304
Perceptron	9.017371	0.312642	0.311205	0.473817	0.473817	0.386937	0.386937
SGD	8.698813	0.39566	0.39417	0.560977	0.563721	0.426814	0.425561
RF	513.658434	0.582764	0.419789	0.525464	0.524969	0.456056	0.455938
DT	8.327899	0.487765	0.395987	0.452077	0.452105	0.443683	0.443675
AdaBoost	94.951553	0.3643	0.363211	0.579334	0.579334	0.406133	0.406133