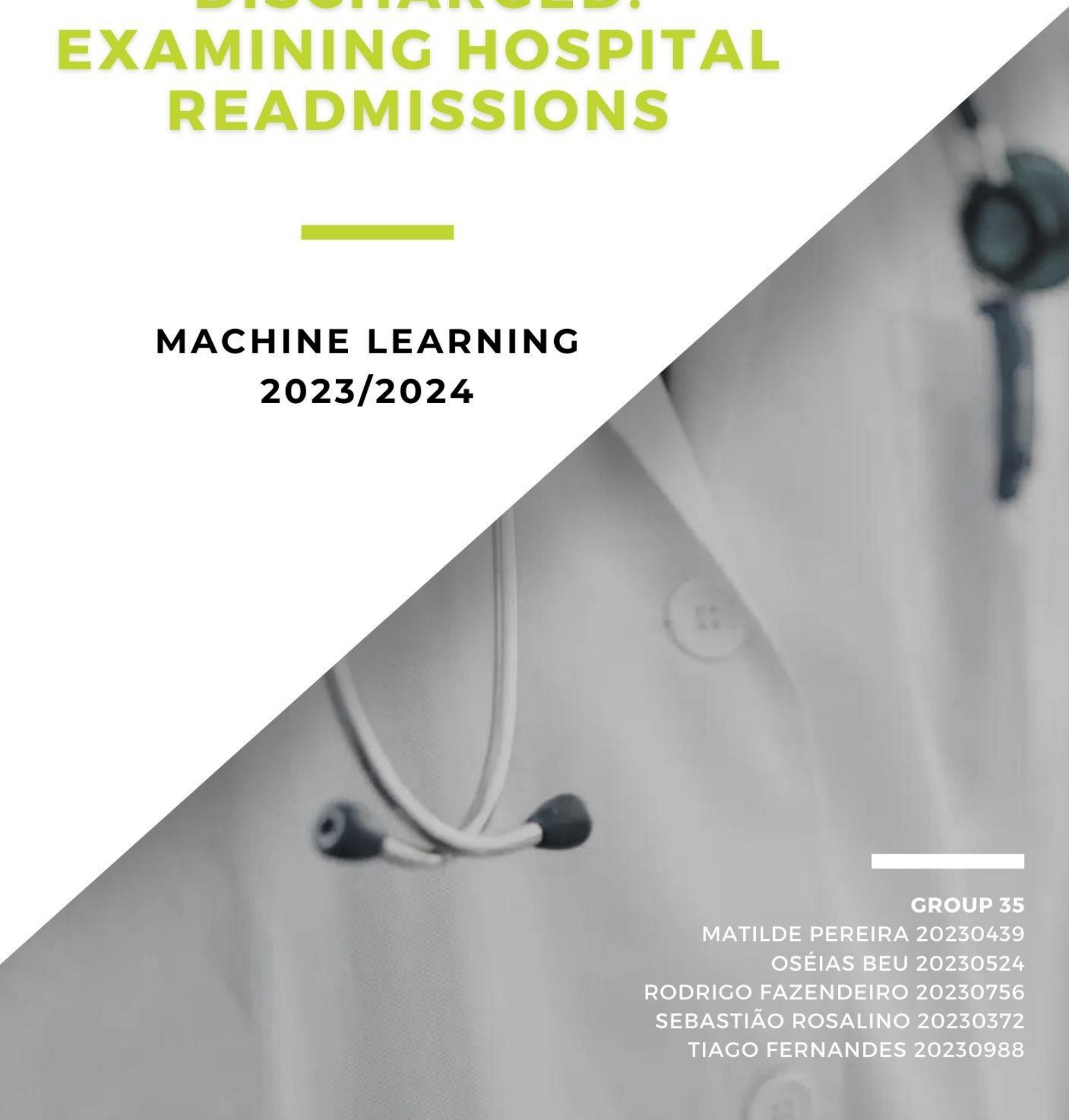
Master in Data Science and Advanced Analytics

# READY TO BE DISCHARGED: EXAMINING HOSPITAL READMISSIONS

## MACHINE LEARNING
## 2023/2024




**GROUP 35**
MATILDE PEREIRA 20230439
OSÉIAS BEU 20230524
RODRIGO FAZENDEIRO 20230756
SEBASTIÃO ROSALINO 20230372
TIAGO FERNANDES 20230988

# INDEX

## ABSTRACT

This project, conducted as part of a Machine Learning course, is centred around the development of predictive models for hospital readmission, particularly targeting diabetic patients due to their substantial impact on healthcare expenses. The primary hypothesis posits that machine learning algorithms can effectively predict readmissions, thus contributing to cost reduction and enhanced patient care.

The study entails two predictive tasks: binary classification to ascertain the likelihood of a patient's readmission within 30 days post-discharge, and multiclass classification to predict the readmission timeframe, categorized as 'No', '<30 days', or '>30 days'.

An extensive data exploration and preprocessing phase, encompassing data cleaning, outlier removal, and missing data imputation, was essential for optimizing prediction performance. The models incorporated diverse features, including patient demographics, medical history, and diagnostic data. Feature selection techniques were applied to distil the most impactful factors, thus bolstering the models' performance and interpretability. The outcome of this study is highlighted by the models' performance, measured in terms of the F1-Score. This metric was chosen due to its balanced consideration of precision and recall, with an emphasis on minimizing false negatives. In a healthcare context, this approach aligns with prioritizing the identification of patients at higher risk of readmission, as failing to predict a 'Yes' (readmission) is more critical than inaccurately predicting a 'Yes'.

The binary classifier showed modest effectiveness in identifying at-risk patients for 30-day readmissions, while the multiclass classifier provided valuable insights into varying readmission timeframes.

In summary, the findings affirm the potential of machine learning in predicting hospital readmissions, underlining its pivotal role in informing healthcare decision-making. The results emphasize the importance of data-driven methodologies in enhancing healthcare management and patient care outcomes.

# I. INTRODUCTION

The ability to accurately predict hospital readmissions, particularly for diabetic patients, is crucial in enhancing the quality of healthcare services. Effective predictions can prevent unnecessary hospitalizations, facilitate optimal resource allocation, and identify previously unknown readmission risk factors. For patients, this translates to accurate diagnoses, effective treatments, shorter hospital stays, and minimized psychological impacts. The ratio of readmissions to total hospital admissions serves as a significant metric, offering insights into service quality and patient care, while also enhancing discharge processes through better medication reconciliation and follow-up procedures. Traditionally, readmission predictions have relied on standard statistical models. However, recent research has pivoted towards the potential of machine learning. Machine learning's unique ability to train precise and general models from datasets makes it particularly suited for complex, non-linear data challenges in healthcare. An extensive overview study conducted by Huang, et al.,[1] concluded that tree-based models, neural networks, regularized logistic regressions, and Support Vector Machines are believed to be the most efficient Machine Learning algorithms for predicting hospital readmissions in the USA. The present project aims to harness this potential by developing a binary classification model to accurately predict readmissions within 30 days and a multiclass classifier to predict the timeframe of patient readmissions, offering a nuanced understanding of patient risk levels and care requirements. Initially, it was expected to obtain high performance results in both binary and multiclass classification, using one of the (or several) algorithms reviewed as the best for this task.

# II. DATA EXPLORATION AND PREPROCESSING

## i. Data Description

Provided by Professor Ricardo Santos, the study's dataset includes a training set with 71,236 records and 29 predictors, and a test set with 30,530 records, originating from the USA. Details on data collection and timeframe are unspecified. The test dataset, excluding the target variables *'readmitted_binary*' and *'readmitted_multiclass'*, mirrors the training set. Data analysis will categorize features into numerical, categorical ordinal, and categorical nominal groups:

- Analysis shows many zero values in *'outpatient'*, *'emergency'*, and *'inpatient'* visit features, indicating limited hospital interactions. The *'average_pulse_bpm'* feature shows high variability around a mean of 100, and *'length_of_stay_in_hospital'* averages at 4.4 days. On average, there are 43 lab tests and 1.34 non-lab procedures per encounter, with patients taking around 16 medications. Diagnoses range from 1 to 16. The consistency in numerical features' statistics across both datasets ensures data reliability and model generalizability.
- The categorical ordinal features, *'age'* and *'weight*', both have missing values, with *'weight'* being particularly notable as 96.8% of its data is missing.
- Categorical nominal features show limited variation in *'gender'* and *'race'*, while *'admission_type'*, *'discharge_disposition'*, and *'admission_source'* have diverse categories. The *'medical_specialty'* feature has many missing values (identified by '?'), and diagnosis features (*'primary'*, *'secondary'*, *'additional'*) include various unique codes. The *'glucose_test_result'* and *'a1c_test_result'* features each have three categories. Medication-related features are binary or a list of prescribed medications per encounter.

  In Annex 1, Table 1, it is presented the 29 features of both datasets (excluding the targets).

[1] Huang Y, Talwar A, Chatterjee S, Aparasu RR. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. BMC Med Res Methodol. 2021 May 6

## ii. Data Cleaning and Preparation

**Addressing Data Inconsistencies**

The training dataset revealed inconsistencies in two features: *'admission_type'* and *'admission_source'*. Records indicated patients labelled as "Newborn" under *'admission_type'*, suggesting infancy, yet their age was outside the 0-10 years interval. Similarly, two records under *'admission_source'* were classified as "Sick baby" and "Extramural birth," but the patients' ages were between 70 and 80 years. These discrepancies were deemed data entry errors, leading to the removal of a total of 7 records from the training dataset. A proof of this is provided in Annex 1, Figure 1.

**Data Leakage Issue**

The test dataset exhibited data leakage [2]due to patients having multiple encounters. The issue was that non-final encounters (for the same patient) suggested future readmissions, influencing the binary and multiclass targets. To address this, each encounter was analysed independently, disregarding *'patient_id'*, thus eliminating foresight into future admissions and focusing solely on available data per encounter. This strategy prevented data leakage and maintained the model's relevance and applicability to real-world, future-blind situations.

**Data Cleaning**

Each feature was meticulously cleaned according to its specific requirements, ensuring they were optimized for use in both models. This process entailed a range of techniques to refine the dataset:

- Dropping irrelevant features.
- Converting '?' characters to *NaNs* for appropriate missing data imputation.
- Changing '?' into 'Uncategorized' for class absence, without representing a missing value.
- Applying mappings to enhance the model's ability to generalize.
- Implementing binarization to further improve generalization.
- Executing other cleaning transformations for data consistency and integrity.

Each of these steps are elaborated upon in the Annex 1, Table 2, providing a comprehensive view of the data cleaning process feature by feature.

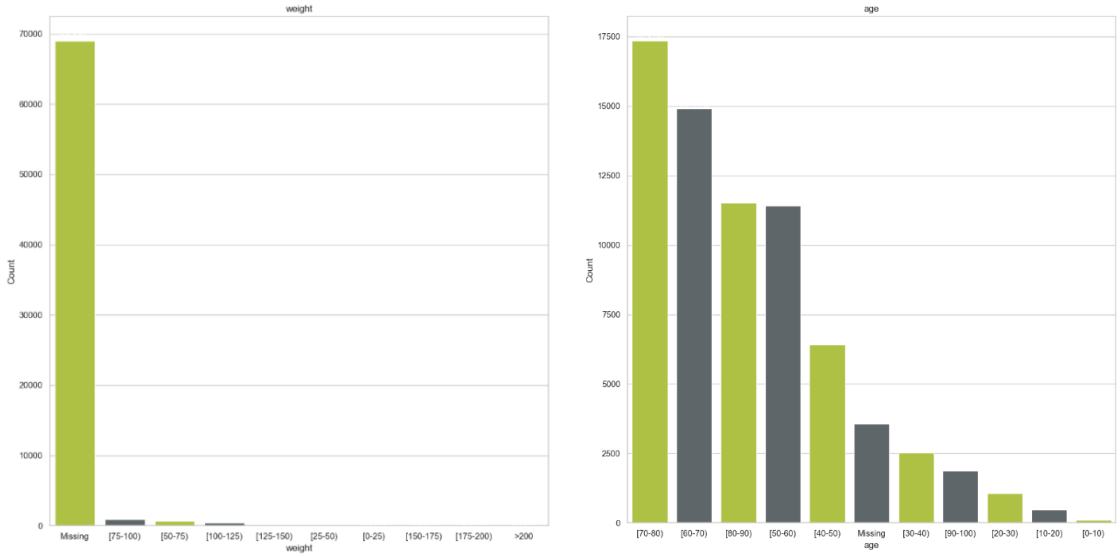
**Feature Engineering**

Two new features, *'number_of_prescribed_medications'* and *'average_visits_in_previous_year'*, were introduced for improved data understanding and model performance. The first, calculated from *'medication'*, indicates 1 to 6 medications prescribed per encounter, with most patients receiving just one. The second averages the previous year's *'inpatient', 'outpatient', and 'emergency'* visits, ranging from 0 to 27 in the training set and up to 23 in the test set, reflecting patient visit trends.

[2] Danb, August 2017, "Data Leakage – Credit Card Data from book "Econometric Analysis"

**Outlier Removal**

Outliers were identified in numerical features to prevent skewed results, using boxplots and histograms (Annex 1, Figure 2 & 3) to visualize data distribution, and set specific outlier thresholds (Annex 1, Table 6). This identified 0.16% of the data as outliers but removing them decreased model performance and exacerbated class imbalance, especially in binary classification where the minority class was about 11%. Consequently, outliers were retained to represent the minority class accurately. The distribution of categorical ordinal features (*'age'* and *'weight'*) was analysed using bar charts:




Graphics 1 (right) and 2 (left) displaying the distributions of different values from categorical ordinal features 'weight' and 'age', respectively. Graphic 1 scale is set to 70,000 and has 10 categories. Graphic 2 scale is set to 17,500 and has 11 categories.

Despite low representation in age intervals [0-10] and [10-20], these were kept maintaining valuable insights and ensure pattern recognition within these groups. Conversely, the *'weight'* feature, with 96.85% missing values, was removed for lacking analytic value. No values were omitted from categorical nominal features to preserve their unique information, ensuring a robust dataset for enhanced model performance.

**Missing Data Imputation**

- Numerical Features: No missing values were found in either the train or test datasets.
- Categorical Ordinal Features: In the training and testing datasets, *'age'* had 3557 and 1531 missing values, respectively. These were imputed using an ordinal encoder for age intervals from [0-10] to [90-100]. The imputation was based on standardized numerical features and employed a KNN Imputer with 15 neighbours, selected to align the method closely with the mode of age intervals while capturing possible data patterns. Post-imputation, *'age'* was rounded to the nearest integer and reclassified as a numerical feature, ensuring compatibility with future Feature Selection methods. This process, applied to both training and test sets, aimed to prevent data leakage.
- Categorical Nominal Features: The *'race'* feature, with 5069 and 2191 missing values in the training and test datasets, was imputed with its most frequent value, 'Caucasian'. This was due to the difficulties in predicting *'race'* from medical data and its weak correlation with other features. For *'discharge_disposition'*, a label encoder and KNN Classifier (11 neighbours) were used to impute values, aligning closely with this feature's mode, and capturing possible data patterns. The same model trained on the training set was used for the test set to maintain consistency and avoid data leakage.

# III. BINARY CLASSIFICATION

## i. Feature Selection

This section will display the results of feature selection and the conclusions taken. Feature Selection methods applied to the binary classification task were exclusively based on the training dataset. All details about the results and methodologies can be found in Annex 2.

**Numerical and Categorical Ordinal Features**

- Pearson's correlation matrix detected low correlations between features, with the highest being a moderate to weak association of 0.46 between '*number_of_medications*' and '*length_of_stay_in_hospital*'.
- The Kruskal-Wallis test [3]rejected the null hypothesis of similar median values between the features and the target feature for all features except '*average_pulse_bpm*' at a significance level of 0.05. The p-value for '*average_pulse_bpm*' was 0.9694.
- For Kendall's Tau [4]test all features except '*average_pulse_bpm*' rejected the null hypothesis that the correlation coefficient Tau equals 0. The Tau value and p-value of '*average_pulse_bpm' of* 0.000125 and 0.967612, respectively.
- LASSO Regression with Cross-Validation proposed the removal of *'average_pulse_bpm'* and *'number_lab_tests'* features.
- RFE|LR, RFE|SVC and RFE|RF gave very poor F1-Score results and were therefore deemed as untrustworthy.

Departing from the results obtained, it was observed that Pearson's correlation matrix did not reveal significant linear associations. Kruskal-Wallis test and Kendall's Tau test supported the non-significant association of '*average_pulse_bpm*' with the target feature. The elimination of *'average_pulse_bpm'* and *'number_lab_tests'* was suggested by LassoCV. While RFE results were not deemed trustworthy due to low F1-Scores, potential factors influencing the results included feature relevance, data quality issues (such as outliers), and the sensitivity of the F1-Score to class imbalance. The removal of *'average_pulse_bpm'* was supported by the consistency of results from different methodologies, which increased the confidence in the decision.

**Categorical Nominal Features**

- For the Chi-Squared test, features *'race'* (p-value = 0.0634) and *'gender'* (p-value = 0.1214) had p-values above the significance level established of 0.05.
- Mutual Information (MI) and Random Forest (RF) analysis yielded poor feature importance scores across all features.

The Chi-square test indicated that *'race'* and *'gender'* had p-values above the 0.05 significance level, suggesting no strong association with the binary target. MI and RF also showed minimal contribution of these features to target prediction. The low importance scores in RF and MI might stem from testing these categorical nominal features in isolation. Their predictive power could potentially be more meaningful if assessed alongside numerical features. Based on these insights, *'race'* and *'gender'* were removed following the Chi-Squared test results, underscoring the necessity of a comprehensive methodological approach in feature selection.

[3] Laerd Statistics, Kruskal-Wallis H Test using SPSS Statistics

[4] Ishan, October 3, 2023, Virginia Tech & IIT Delhi, Understanding Kendall's Tau Rank Correlation

## ii. Additional Preprocessing Steps

**Feature Encoding**

Insights from *'medication'* were extracted by turning unique medications into dummy variables and substituting *NaNs* with 'no_medication'. After removing the original *'medication'* feature, both training and testing datasets were updated. One-hot encoding created dummy variables for each category in the categorical-nominal features, with dataset-specific medications excluded for consistency (Annex 2, Table 2). The final training and test datasets comprised 71,229 and 30,530 instances, each with 133 predictors.

**Tuning the Model based on Patient Outcomes**

In the pursuit of refining the predictive model for hospital readmissions, particularly in the binary context, an insightful observation was made regarding patient encounters labelled with 'Expired' in their *'discharge_disposition'* feature. This observation led to a strategic adjustment in the model's approach to handling these specific cases.

Rationale Behind the Adjustment

It was observed that cases marked as 'Expired' in *'discharge_disposition'* always had *'readmitted_binary'* set to 0 ('No'). This reflects the reality that deceased patients cannot be readmitted. However, to highlight the severity of these cases, *'readmitted_binary'* was changed from 0 to 1 ('Yes') in the training dataset. This change is not to suggest deceased patients would be readmitted, but to help the model recognize critical health patterns that led to deceased outcomes and associate them with similar clinical profiles that might lead to readmission within 30 days, for the predictions in the test dataset.

Application to the Test Dataset

In the test dataset, a strategy was implemented to improve the binary F1-Score. A mask identified cases with 'Expired' in *'discharge_disposition'*, and predictions for these were directly set to 'No' for readmission. This method acknowledges the reality of such cases while concentrating the model's prediction on feasible readmission instances.

Overview

This methodological adjustment has dual benefits. On one hand, it equips the model to identify and learn patterns associated with patients experiencing critical health conditions. By compelling the model to focus on scenarios involving patient deaths, it becomes adept at recognizing similar high-risk profiles that could lead to potential readmissions in unseen encounters. On the other hand, it allows for immediate and logical prediction of deceased patients as non-readmitted, streamlining the model's decision-making process. This balanced approach enhances the model's predictive binary F1-Score and relevance to real-world healthcare situations.

## iii. Results and Discussion of Main Findings

In the context of constructing binary classification models, a decision was made to employ 5-Fold Cross-Validation Grid Search for the purpose of optimizing hyperparameters efficiently. In the subsequent stage of model evaluation on the validation dataset, the utilization of the holdout method was deemed necessary due to computational limitations. This choice was driven by the estimation that implementing 5-Fold Cross-Validation would entail an additional processing time of approximately 160 minutes.

**K-Nearest Neighbours**

The KNN model, for classification, predicts based on the nearest neighbours' majority vote. Hyperparameters fine-tuning was carried out through a 5-Fold Cross-Validation Grid Search with the binary F1-Score as the metric:

- neighbours (3): Balances bias and variance by considering three nearest neighbours.
- metric ('manhattan'): Calculates the sum of absolute differences in coordinates.
- weights ('distance'): Weighs predictions based on the inverse of distance, enhancing the influence of closer neighbours.

The model showed a weak performance on the validation set, reflected by an F1-Score of 0.12.

**Multi-Layer Perceptron**

The Multi-Layer Perceptron (MLP), an artificial neural network for regression and classification, features interconnected neurons across an input layer, a single or multiple hidden layers, and an output layer. The MLP refines connection weights through backpropagation during training to enhance prediction performance, capable of learning complex patterns due to its non-linear design. Hyperparameters fine-tuning was carried out through a 5-Fold Cross-Validation Grid Search with the binary F1-Score as the metric:

- activation ('tanh'): Uses the hyperbolic tangent function, normalizing neuron outputs between -1 and 1.
- alpha (0.05): Regularizes to prevent overfitting by penalizing larger weights.
- hidden_layer_sizes ((100,)): A single hidden layer with 100 neurons, optimized for complex pattern detection.
- learning_rate_init (0.001): Sets a learning rate that balances weight adjustment speed with convergence performance.
- solver ('adam'): Employs the 'adam' optimizer, suitable for large datasets and complex spaces.

The MLP model improved validation performance, achieving an F1-Score of 0.51.

**Stacking Ensemble**

The Stacking Ensemble model, aiming to bolster prediction accuracy, combines two base models with a final estimator for integrated predictions. The composition is as follows:

<u>Base Models</u>

- Gaussian Naive Bayes[5]: A probabilistic classifier based on applying Bayes' theorem, particularly effective for large feature spaces.
- Histogram Gradient Boosting Classifier[6]: An efficient implementation of the gradient boosting framework, with max_depth=4 and learning_rate=0.05, offering robust performance on complex datasets.

[5] Martins, Carla, 2023, Gaussian Naive Bayes Explained With Scikit-Learn

[6] Brownlee, Jason, 2021 April

Final Estimator

- Logistic Regression: With C=0.05 and class_weight='balanced', this model serves as the final decision layer, effectively integrating the predictions from the base models.

Sampling Strategy

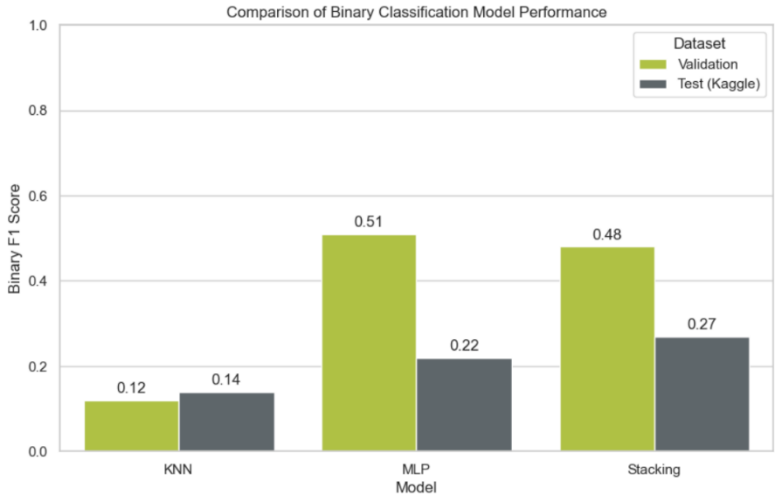
- Random Under Sampler: Applied to balance the dataset, with a sampling strategy of 0.8, ensuring a more representative training process. This means that the minority class/majority class ratio is 0.8.

Cross-Validation

- Utilizing Stratified K-Fold with 5 splits, the model's performance was evaluated to ensure its robustness and generalizability.

The Stacking Ensemble achieved a Binary F1-Score of 0.48 on validation via Stratified K-Fold, reflecting a more controlled overfitting and potential for stronger generalization than the MLP.

**Model Performance Comparison in Test Dataset**




Graphic 3 – Comparison of model performance between K-Nearest Neighbours, Multi-Layer Performance and Stacking Ensemble for binary classification.

The bar chart compares binary F1 Scores of three binary classifiers. KNN scored the lowest with 0.12 on validation and slightly better on the test set. MLP outperformed on validation with 0.51 but fell to 0.22 on the test set, hinting at overfitting. Stacking, with a 0.48 validation score and a consistent 0.27 test score, suggests better generalization than MLP, balancing training fit and new data prediction. Stacking likely surpassed MLP and KNN due to its integration of diverse models, each correcting the others' errors and capturing unique data patterns. This approach reduces overfitting, seen in complex models like MLP, and balances bias and variance better than KNN. Stacking's robustness to noisy data and its ability to combine strengths while mitigating weaknesses of individual models contribute to its enhanced performance on unseen data. Therefore, Stacking was chosen as the final predictive model.

# IV. Multiclass Classification

## i. Feature Selection

This section summarizes the feature selection results for the multiclass scenario, based exclusively on the training dataset. Detailed information on the outcomes and methods is in Annex 2.

**Numerical and Categorical Ordinal Features Results**

- Pearson's correlation matrix detected low correlations between features, with the highest being a moderate association of 0.46 between '*number_of_medications*' and '*length_of_stay_in_hospital*'.
- The Kruskal-Wallis test rejected the null hypothesis of similar median values between the features and the multiclass target for all features except '*average_pulse_bpm*' at a significance level of 0.05. The p-value for '*average_pulse_bpm*' was 0.2404.
- For Kendall's Tau test, all features except *'average_pulse_bpm'* and *'number_of_prescribed_medications'* rejected the null hypothesis that the correlation coefficient Tau equals 0. The Tau value and p-value of *'average_pulse_bpm'* were -0.00493 and 0.096255, respectively. Also, the Tau value and p-value of *'number_of_prescribed_medications'* were 0.006041 and 0.083812, respectively.
- LASSO Regression with Cross-Validation proposed the maintenance of all numerical features.
- RFE|LR and RFE|SVC suggested retaining only the '*average_visits_in_previous_year*' feature, while RFE|RF recommended the inclusion of every numerical feature.

In feature selection for the multiclass target, various methods yielded mixed results. Pearson's correlation showed limited linear relationships. Both Kruskal-Wallis and Kendall's Tau tests suggested removing *'average_pulse_bpm'* due to its weak significance. Conversely, LASSO Regression recommended keeping all numerical features. RFE results were mixed, with RFE|LR and RFE|SVC highlighting *'average_visits_in_previous_year'* as the only important feature, while RFE|RF agreed with LASSO for retaining all features. Ultimately, the recommendation to drop *'average_pulse_bpm'* across most of the models due to its non-predictive nature drove the final decision. This highlights the value of integrating diverse analytical perspectives.

**Categorical Nominal Features Results**

- For the Chi-Squared test, all features had p-values below the significance level of 0.05.
- Mutual Information (MI) and Random Forest (RF) analysis yielded poor feature importance scores across all features.

Chi-Squared, MI, and RF results were cautiously interpreted, leading to retaining all features. The absence of removal recommendations from Chi-Squared and unreliable low scores from MI and RF, possibly due to isolated testing, emphasized assessing these features alongside numerical ones for potential predictive power. Consequently, no categorical-nominal features were removed.

## ii. Additional Preprocessing Steps

**Feature Encoding**

Unique medications in *'medication'* were converted into dummy variables, with *NaNs* replaced by 'no_medication'. After removing the original feature, the datasets were updated using one-hot encoding for categorical-nominal features, excluding dataset-specific medications for consistency (Annex 2, Table 2). The final training and test datasets now have 71,229 and 30,530 instances, respectively, each with 140 predictors.

**Tuning the Model based on Patient Outcomes**

The multiclass model for predicting hospital readmissions was refined by adjusting for 'Expired' labels in *'discharge_disposition'*, a critical observation that guided strategic changes.

Rationale Behind the Adjustment

Every instance of 'Expired' in *'discharge_disposition'* matched with a *'readmitted_multiclass'* value of 0 ('No'), reflecting that deceased patients cannot be readmitted. Recognizing the importance of these cases for understanding clinical condition severity, the *'readmitted_multiclass'* label was changed from 0 ('No') to 1 ('<30') in the training dataset. This was to signal critical cases to the model, not to suggest deceased patients would be readmitted, but to force the model to identify patterns associated with critical health conditions, helping it to recognize potential readmissions, within 30 days, for similar clinical profiles in unseen data.

Application to the Test Dataset

In the test dataset, to improve the Weighted F1-Score, a mask identified 'Expired' cases in *'discharge_disposition'*, directly setting their readmission predictions to 'No'. This strategy acknowledges the reality of these cases, focusing the model on predicting readmission where it is plausible.

Overview

This methodological adjustment serves two purposes: it trains the model to identify patterns in critical health conditions, including high-risk profiles that might lead to readmissions in less than 30 days, and enables immediate, logical predictions for deceased patients as non-readmitted. This balanced approach not only improves the model's Weighted F1-Score but also its applicability in real-world healthcare scenarios.

## iii. Results and Discussion of Main Findings

For optimizing hyperparameters in multiclass classification models, 5-Fold Cross-Validation Grid Search was used, while the holdout method was chosen for validation due to computational constraints, saving about 144 minutes. The training dataset showed that 53.9% of patients were not readmitted, 34.9% readmitted within 30 days, and 11.2% after 30 days. Lacking labels in the test set, the training set was used for model fitting. To mitigate overfitting, the training dataset was split into training (50%), validation (30%), and testing (20%), balancing pattern learning, hyperparameter tuning, and real-world applicability. Four models were developed with this approach.

**K-Nearest Neighbours**

The KNN model, for classification, predicts based on the nearest neighbours' majority vote. Hyperparameters fine-tuning was carried out through a 5-Fold Cross-Validation Grid Search with the Weighted F1-Score as the metric:

- neighbours (7): Balances bias and variance by considering seven nearest neighbours.
- p (2): A value of 2 implies the use of the Euclidean distance, which calculates the straight-line distance between two points in space.
- weights ('distance'): Weighs predictions based on the inverse of distance, enhancing the influence of closer neighbours.

The model showed moderate validation performance with a Weighted F1-Score of 0.51 closely reflected in the test set with a score of 0.51.

**Gaussian Naive Bayes**

The Gaussian Naive Bayes algorithm, a probabilistic classifier utilizing Bayes' theorem under the feature independence assumption, was chosen for its ability to efficiently process many features. Its hyperparameters were optimized through 5-Fold Cross-Validation focusing on optimizing the Weighted F1-Score (explanation of the Gaussian Naive Bayes model in Annex 2):

- var_smoothing (1e-07): Selected to enhance the model's probability estimates, making them more resilient to outliers by smoothing the data.

On the validation set, the Gaussian Naive Bayes model achieved a low Weighted F1-Score of 0.12, indicating difficulties in balancing precision and recall. A similar performance was observed in the test set, where it recorded a Weighted F1-Score of 0.12, underscoring its limitations in predicting unseen data.

**Histogram-Based Gradient Boosting Classification Tree**

Histogram-Based Gradient Boosting, an ensemble method within the boosting algorithm family, improves performance by sequentially combining multiple decision trees, each addressing the errors of its predecessor. This approach is efficient for large datasets and effectively handles missing data and categorical features. Its hyperparameters were fine-tuned using 5-Fold Cross-Validation with the Weighted F1-Score as the metric (explanation of the Histogram-Based Gradient Boosting model in Annex 2):

- learning_rate (0.2): Determines the learning speed, with a higher rate ensuring quicker convergence but risking surpassing the optimal solution.
- max_depth (5): Controls each tree's depth to balance model complexity and the risk of overfitting, allowing the model to learn detailed patterns.
- max_iter (50): Sets the number of boosting stages, where more iterations enhance model refinement but increase computation time and overfitting risk.

The Histogram-Based Gradient Boosting model demonstrated a robust performance on both validation and test sets, with Weighted F1-Scores of 0.55 and 0.55, respectively. These scores reflect a balanced compromise between precision and recall, showcasing the model's effectiveness in classifying the instances for this specific dataset. The model's strength lies in its ability to capture complex nonlinear relationships within the data, making it a suitable choice for diverse and challenging classification tasks.

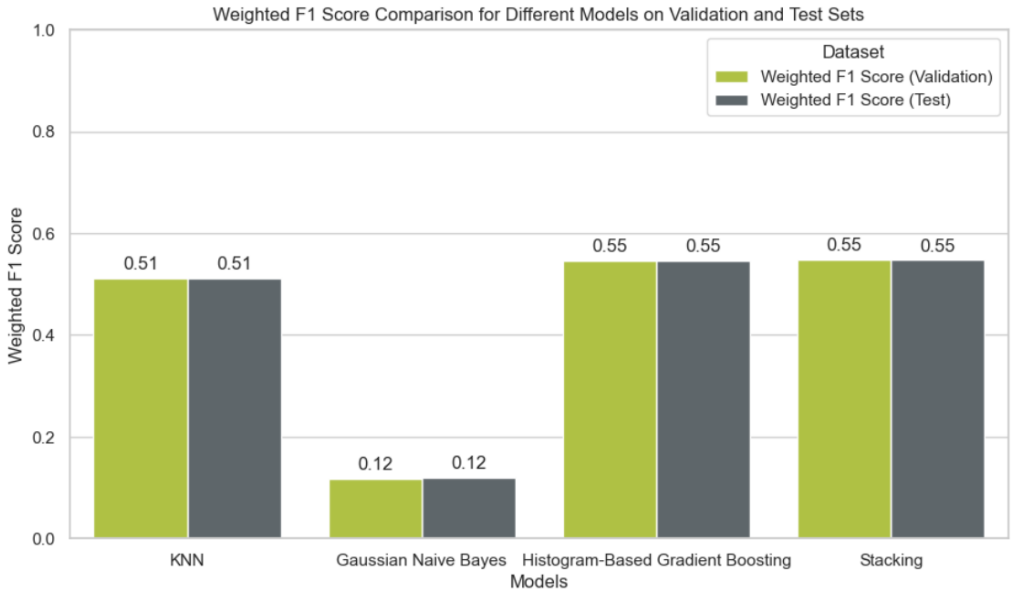
**Stacking Ensemble**

The Stacking Classifier, effectively combining the already-tuned base models (K-Nearest Neighbours, Gaussian Naive Bayes, Histogram-Based Gradient Boosting), leverages their collective strengths to improve performance. A meta-learner, specifically Linear Regression, is then used to make the final decision based on these models' outputs. This strategy ensures a balance between precision and recall, vital for nuanced medical predictions like hospital readmissions.

Achieving the Weighted F1-Scores of 0.55 on validation and 0.55 on the test set, the Stacking model demonstrates its robustness in both prediction quality and case detection. Its effectiveness in integrating insights from multiple models results in more reliable predictions, underscoring the power of ensemble methods in complex classification scenarios.

**Overall Performance and Predictions**

In selecting the final model, the Histogram-Based Gradient Boosting (HGB) and Stacking Classifier showed similar Weighted F1 Scores on validation and test datasets. The HGB model was chosen for predictions on the unseen test dataset due to its lower complexity and higher interpretability compared to the Stacking Classifier. HGB's single algorithm approach offers computational efficiency and quicker training times. Its interpretability is vital for healthcare applications where clear justification of predictions is essential. Despite the Stacking Classifier's robustness, HGB's simplicity and transparency make it more suitable for practical, real-world implementation.




Graphic 4 – Comparison of model performance between K-Nearest Neighbours, Gaussian Naïve Bayes, Histogram-base Gradient Boosting, and Stacking Ensemble for multiclass classification.

## V. CONCLUSION

In this project focused on predicting hospital readmissions, particularly for diabetic patients, the findings present a nuanced view of the challenges inherent in such a predictive task. The primary metric used to evaluate model performance were the binary F1-Score and the Weighted F1-Score, chosen for its balance between precision and recall, which is crucial in a healthcare context where false negatives (failing to predict a readmission) are more critical than false positives (incorrectly predicting a readmission).

The results, however, were modest, with the binary classification achieving an F1-Score of 0.27 and the multi-classification an F1-Score of 0.55. These scores, while not negligible, indicate the inherent difficulty of the prediction task. A significant contributing factor to this challenge was the noticeable class imbalance in the dataset, which tends to hinder the model's ability to effectively discern patterns among the less represented, but critical, positive cases.

The outcomes, particularly for the binary classification model, fell short of initial expectations. It was anticipated that the models would demonstrate a higher degree of predictive power. This expectation assumed that the extensive range of patient data and medical indicators available in the dataset would provide a robust foundation for the machine learning models to identify clear predictive patterns.

However, a primary limitation encountered during the project was related to a proven data leakage issue in the test dataset. It was observed that the records might not be chronologically ordered, which can significantly impact the performance and generalizability of the predictive models. Addressing data leakage is crucial in ensuring the validity of a model's predictive power on unseen data.

It's crucial to acknowledge the unconvincing results in feature selection, particularly given the exclusive testing of numeric and nominal categorical features. This approach may have introduced bias, potentially impacting the perceived significance of these features. A promising solution lies in assessing the features collectively, thereby offering a more comprehensive understanding of their importance. Moreover, it is essential to recognize the substantial influence of the preprocessing methodology on the lackluster performance observed in feature selection. A thoughtful reconsideration and refinement of the preprocessing steps may prove instrumental in enhancing the overall effectiveness of feature selection techniques.

Reflecting on what could have been done differently, a more rigorous approach in the collection and preparation of the dataset would be beneficial. Ensuring chronological ordering and scrutinizing for any potential data leakage scenarios should be integral steps in the data preparation phase. Additionally, more advanced feature engineering techniques could have been employed to derive more nuanced attributes from the existing data, potentially providing the models with richer insights for prediction.

For future work, exploring more complex and sophisticated models holds promise. The constraints of using only vanilla scikit-learn implementations in this project meant that more advanced techniques, such as deep learning models, or state-of-the-art algorithms, could not be explored.

Future studies could benefit from employing these advanced methodologies, which might be better equipped to handle the complexities and imbalances inherent in medical datasets. Moreover, integrating domain-specific knowledge during the feature engineering and model evaluation stages could further enhance the predictive capabilities of the models.

Lastly, a more comprehensive approach to handling class imbalance, such as advanced oversampling techniques or cost-sensitive learning, could be crucial in improving model performance, particularly in predicting the less frequent, but more critical, positive cases of readmission.

## BIBLIOGRAPHY

[1] Huang Y, Talwar A, Chatterjee S, Aparasu RR. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. BMC Med Res Methodol. 2021 May 6;21(1):96. doi: 10.1186/s12874-021-01284-z. PMID: 33952192; PMCID: PMC8101040.

[2] Danb, August 2017, "Data Leakage – Credit Card Data from book "Econometric Analysis", https://www.kaggle.com/datasets/dansbecker/aer-credit-card-data

[3] Laerd Statistics, Kruskal-Wallis H Test using SPSS Statistics, https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

[4] Ishan, October 3, 2023, Virginia Tech & IIT Delhi, Understanding Kendall's Tau Rank Correlation, https://ishanjainoffical.medium.com/understanding-kendalls-tau-rank-correlation-c959a7daea56

[5] Martins, Carla, 2023, Gaussian Naive Bayes Explained With Scikit-Learn, https://builtin.com/artificial-intelligence/gaussian-naive-bayes

[6] Brownlee, Jason, 2021 April, https://machinelearningmastery.com/histogram-based-gradient-boosting-ensembles/

[7] Anne Monteiro Mendes de Senna, July 2022, Predictive Modelling Of Hospital Readmissions In Diabetic Patients Clusters, (https://run.unl.pt/bitstream/10362/145706/1/TGI1636.pdf)

# ANNEXES

## Annex 1

**Table 1 - Data description**

| No | Features | Description | Characterization |
|---|---|---|---|
| | | **Panel 1: Patient Information** | |
| **1** | Patient ID | Identifier of the patient | Numerical |
| **2** | Country | Country | Categorical Nominal |
| **3** | Encounter ID | Unique identifier of the encounter | Numerical |
| **4** | Race | Patient's race | Categorical Nominal |
| **5** | Gender | Patient's gender | Categorical Nominal |
| **6** | Age | Patient´s age interval | Categorical Ordinal |
| **7** | Weight | Patient´s weight | Categorical Ordinal |
| | | **Panel 2: Healthcare Utilization and Insurance** | |
| **8** | Payer Code | Code of the health insurance provider | Categorical Nominal |
| **9** | Outpatient Visits in previous year | Number of visits the patient made with the intention of leaving on the same day to the hospital in the year preceding the encounter | Numerical |
| **10** | Emergency Visits in Previous Year | Number of emergency visits the patient made to the hospital in the year preceding the encounter | Numerical |
| **11** | Inpatient Visits in Previous Year | Number of visits with the intention to stay overnight the patient made to the hospital in the year preceding the encounter | Numerical |

| Panel 3: Admission and Stay Details | | | |
|---|---|---|---|
| **12** | Admission Type | Type of admission of the patient (e.g. Emergency, Urgent, etc) | Categorical Nominal |
| **13** | Medical Specialty | Medical specialty on which the patient was admitted (e.g. Cardiology, etc) | Categorical Nominal |
| **14** | Average Pulse bpm | Average pulse of the patient during their stay in the hospital in beats per minute | Numerical |
| **15** | Discharge Disposition | Destination given to the patient after being discharged | Categorical Nominal |
| **16** | Admission Source | Source of the patient before being admitted in the current encounter (e.g. Transfer from a hospital, etc) | Categorical Nominal |
| **17** | Length of Stay in Hospital | Number of days between admission and discharge | Numerical |

| Panel 4: Medical Procedures and Diagnoses | | | |
|---|---|---|---|
| **18** | Number Lab Tests | Number of lab tests performed during the encounter | Numerical |
| **19** | Non Lab Procedures | Number of non-lab procedures performed during the encounter | Numerical |
| **20** | Number Of Medications | Number of distinct types of medication administered during the encounter | Numerical |
| **21** | Medication | List containing all generic names for the medications prescribed to the patient during the encounter | Raw list of medications |
| **22** | Primary Diagnosis | Primary diagnosis | Categorical Nominal |
| **23** | Secondary Diagnosis | Secondary diagnosis | Categorical Nominal |
| **24** | Additional Diagnosis | Additional secondary diagnosis | Categorical Nominal |
| **25** | Number Diagnoses | Number of diagnoses entered to the system | Numerical |

| Panel 5: Diabetes-related Information | | | |
|---|---|---|---|
| **26** | Glucose Test Result | Range of the glucose test results or if the test was not taken | Categorical Nominal |
| **27** | A1c Test Result | Range of the A1C test results or if the test was not taken | Categorical Nominal |
| **28** | Change in Meds During Hospitalization | Indicates if there was a change in diabetic medications (dosage or generic name) | Binary |
| **29** | Prescribed Diabetes Meds | If the patient has prescribed medications for diabetes | Binary |

**Table 1** - Displays the 29 attributes of both datasets (excluding the target variables), delineating what each signifies and identifying whether they are numerical, ordinal categorical, or nominal categorical.

## Figure 1 – Inconsistencies in the '*admission_type*' Feature

| | admission_type | age |
|---|---|---|
| 248 | Newborn | [80-90) |
| 43925 | Newborn | [70-80) |
| 46759 | Newborn | [40-50) |
| 49974 | Newborn | [70-80) |
| 64818 | Newborn | [50-60) |

**Figure 1** - Records indicated patients labeled as "Newborn" under the *'admission_type'* feature, suggesting infancy, yet their age was outside the 0-10 years interval. These discrepancies were deemed data entry errors.

## Table 2 – Data cleaning

| Panel 1: Patient Information | | |
|---|---|---|
| **Feature** | **Approach** | **Reason** |
| **encounter_id** | Drop | Does not provide any valuable information for predictive modeling tasks |
| **country** | Drop | Univariate feature |
| **patient_id** | No transformation | Important for exporting the final predictions file |
| **race** | Replace '?' with NaNs | Assuming '?' means no recording |
| **gender** | No transformation | Future one-hot encoding |
| **age** | No transformation | Bins are suitable for modeling |
| **weight** | Replace '?' with NaNs | Assuming '?' means no recording |

| Panel 2: Healthcare Utilization and Insurance | | |
|---|---|---|
| **Feature** | **Approach** | **Reason** |
| **payer_code** | Binarization (has payer code or not) | It was assumed that having or not having health insurance is more important than knowing which health insurance provider a patient has |
| **outpatient_visits_in_previous_year** | No transformation | Numerical values are suitable for modeling |
| **emergency_visits_in_previous_year** | No transformation | Numerical values are suitable for modeling |
| **inpatient_visits_in_previous_year** | No transformation | Numerical values are suitable for modeling |
| **Panel 3: Admission and Stay Details** | | |
| **Feature** | **Approach** | **Reason** |
| **admission_type** | Replace missing, 'Not available' and 'Not Mapped' with 'Uncategorized' | Assuming that 'Not Mapped', 'Not available' and missings mean the absence of any admission type |
| **medical_specialty** | Map of categories by resemblance | Reducing the number of categories to improve model generalization |
| **average_pulse_bpm** | No transformation | Numerical values are suitable for modeling |
| **discharge_disposition** | Map of categories by resemblance | Reducing the number of categories to improve model generalization |
| **admission_source** | Map of categories by resemblance | Reducing the number of categories to improve model generalization |
| **length_of_stay_in_hospital** | No transformation | Numerical values are suitable for modeling |
| **Panel 4: Medical Procedures and Diagnoses** | | |
| **Feature** | **Approach** | **Reason** |
| **number_lab_tests** | No transformation | Numerical values are suitable for modeling |
| **non_lab_procedures** | No transformation | Numerical values are suitable for modeling |
| **number_of_medications** | No transformation | Numerical values are suitable for modeling |
| **primary_diagnosis** | Map based on the IC9 codes | Increase interpretability to the feature |
| **secondary_diagnosis** | Map based on the IC9 codes | Increase interpretability to the feature |
| **additional_diagnosis** | Map based on the IC9 codes | Increase interpretability to the feature |
| **number_diagnoses** | No transformation | Numerical values are suitable for modeling |
| **Panel 5: Diabetes-related Information** | | |
| **Feature** | **Approach** | **Reason** |
| **glucose_test_result** | Replace missings with 'Not measured' | Assuming a missing represents not taking the test |
| **a1c_test_result** | Replace missings with 'Not measured' | Assuming a missing represents not taking the test |
| **change_in_meds_during_hospitalization** | No transformation | Binary values suitable for modeling |
| **prescribed_diabetes_meds** | No transformation | Binary values suitable for modeling |
| **medication** | Separate different medications by a comma | Future one-hot encoding |
| **readmitted_binary** | Binarization<br>(0 - No, 1 - Yes) | Binary Classification Purposes |
| **readmitted_multiclass** | Conversion into numerical classes | Multiclass Classification Purposes |

**Table 3 - Medical Specialty Mapping**

**Anne Monteiro Mendes de Senna, July 2022, Predictive Modelling Of Hospital Readmissions In Diabetic Patients Clusters, (https://run.unl.pt/bitstream/10362/145706/1/TGI1636.pdf)**

| Significance | Categories |
|---|---|
| **Allergy_and_Immunology** | AllergyandImmunology |
| **Pathology** | Pathology |
| **Anesthesiology** | Anesthesiology<br>Anesthesiology-Pediatric |
| **Dermatology** | Dermatology |
| **Diagnostic_Radiology** | Radiologist<br>Radiology |
| **Emergency_Medicine** | Emergency<br>Trauma |
| **Family_Practice** | Family<br>GeneralPractice |
| **General_Surgery** | Surgeon<br>Surgery-Cardiovascular<br>Surgery-Cardiovascular/Thoracic<br>SurgeryColon&Rectal<br>Surgery-General<br>Surgery-Maxillofacial<br>Surgery-Neuro<br>SurgeryPediatric<br>Surgery-Plastic<br>Surgery-Thoracic<br>Surgery-Vascular<br>SurgicalSpecialty |
| **Internal_Medicine** | Cardiology<br>DCPTEAM<br>Endocrinology<br>Endocrinology-Metabolism<br>Gastroenterology<br>Hematology<br>Hematology/Oncology<br>Hospitalist<br>InfectiousDiseases<br>InternalMedicine<br>Nephrology<br>Neurophysiology<br>Oncology<br>Proctology<br>Pulmonology<br>Rheumatology<br>SportsMedicine<br>Urology |
| **Neurology** | Neurology |
| **Obstetrics_and_Gynecology** | Gynecology<br>Obsterics&Gynecology-GynecologicOnco<br>Obstetrics<br>ObstetricsandGynecology |
| **Ophthalmology** | Orthopedics<br>Orthopedics-Reconstructive |
| **Orthopaedic_Surgery** | Orthopedics |

| | |
|---|---|
| | Orthopedics-Reconstructive |
| **Osteopathics** | Osteopath |
| **Other** | Other<br>OutreachServices |
| **OtherHealthcarePractitioners** | Dentistry<br>Podiatry<br>Psychology<br>Resident<br>Speech |
| **Otolaryngology** | Otolaryngology |
| **Pediatrics** | Cardiology-Pediatric<br>Pediatrics<br>Pediatrics-AllergyandImmunology<br>PediatricsCriticalCare<br>Pediatrics-EmergencyMedicine<br>Pediatrics-Endocrinology<br>PediatricsHematology-Oncology<br>Pediatrics-InfectiousDiseases<br>Pediatrics-Neurology<br>PediatricsPulmonology<br>Perinatology<br>Psychiatry-Child/Adolescent |
| **PhysicalMedicine_and_Rehabilitation** | PhysicalMedicineandRehabilitation<br>PhysicianNotFound |
| **Psychiatry** | Psychiatry<br>Psychiatry-Addictive |

**Table 2** – Medical Specialty Mapping, in order to reduce dimensionality and improve model generalization.

**Table 3 - Discharge Disposition Mapping**

| Significance | Categories |
|---|---|
| **home_discharge** | Discharged to home<br>Discharged/transferred to home with home health service<br>Discharged/transferred to home under care of Home IV provider |
| **facility_transfer** | Discharged/transferred to SNF<br>Discharged/transferred to another short term hospital<br>Discharged/transferred to another rehab fac including rehab units of a hospital<br>Discharged/transferred to another type of inpatient care institution<br>Discharged/transferred to ICF<br>Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital<br>Discharged/transferred to a long term care hospital.<br>Discharged/transferred within this institution to Medicare approved swing bed<br>Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.<br>Discharged/transferred/referred to this institution for outpatient services |

| | |
|---|---|
| | Discharged/transferred/referred another institution for outpatient services<br>Discharged/transferred to a federal health care facility.<br>Neonate discharged to another hospital for neonatal aftercare |
| **hospice** | Hospice / medical facility<br>Hospice / home |
| **expired** | Expired<br>Expired in a medical facility. Medicaid only, hospice.<br>Expired at home. Medicaid only, hospice. |
| **other** | Not Mapped<br>Left AMA<br>Admitted as an inpatient to this hospital<br>Still patient or expected to return for outpatient services |

**Table 3** – Discharge Disposition Mapping, in order to reduce dimensionality and improve model generalization.

**Table 4 - Mapping Admission Source**

| CRITERIA | CATEGORY |
|---|---|
| **NAN** | Uncategorized |
| **EMERGENCY ROOM** | Emergency Room |
| **CONTAINS 'REFERRAL'** | Referrals |
| **CONTAINS 'TRANSFER'** | Transfers |
| **REMAINING CATEGORIES** | Other |

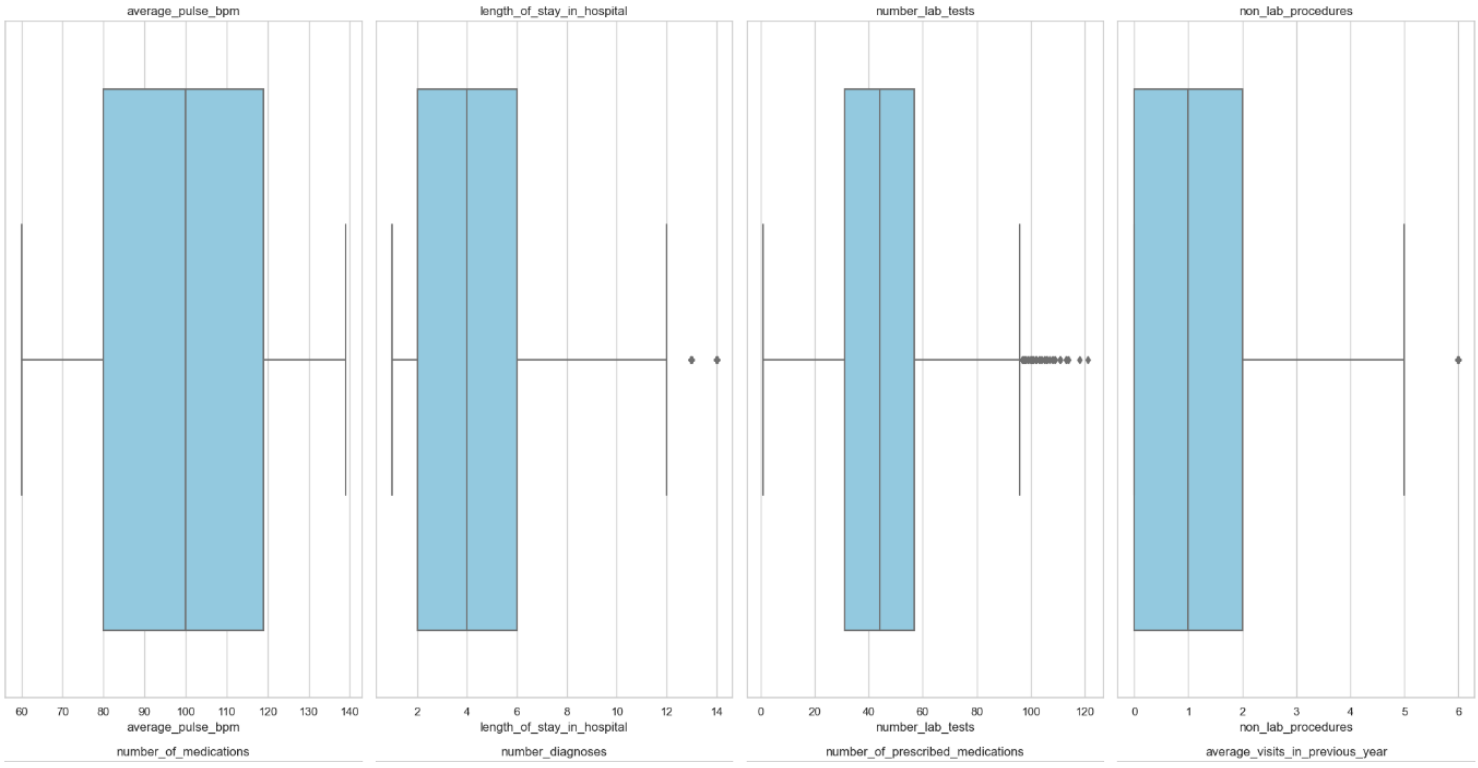
**Table 4** – Mapping Admission Source, to reduce dimensionality and improve model generalization.

**Table 5 - ICD 9 Code (https://en.wikipedia.org/wiki/List_of_ICD-9_codes)**

| Range of Intervals | ICD9 Code |
|---|---|
| **[1 - 139[** | infectious and parasitic diseases |
| **[140 - 240[** | neoplasms |
| **[240 - 280[** | endocrine, nutritional and metabolic diseases, and immunity disorders |
| **[280 - 290[** | diseases of the blood and blood-forming organs |
| **[290 - 320[** | mental disorders |
| **[320 - 390[** | diseases of the nervous system and sense organs |
| **[390 - 460[** | diseases of the circulatory system |
| **[460 - 520[** | diseases of the respiratory system |
| **[520 - 580[** | diseases of the digestive system |
| **[580 - 630[** | diseases of the genitourinary system |
| **[630 - 680[** | complications of pregnancy, childbirth, and the puerperium |
| **[680 - 710[** | diseases of the skin and subcutaneous tissue |
| **[710 - 740[** | diseases of the musculoskeletal system and connective tissue |
| **[740 - 760[** | congenital anomalies |
| **[760 - 780[** | certain conditions originating in the perinatal period |
| **[780 - 800[** | symptoms, signs, and ill-defined conditions |
| **[800 - 1000[** | injury and poisoning |
| **Starting with 'E' or 'V'** | external causes of injury and supplemental classification |
| **'?'** | No ICD-9 |

**Table 5** – IDC 9 CODE, to reduce dimensionality and improve model generalization.

**Figure 2 – Boxplots for outlier removal**



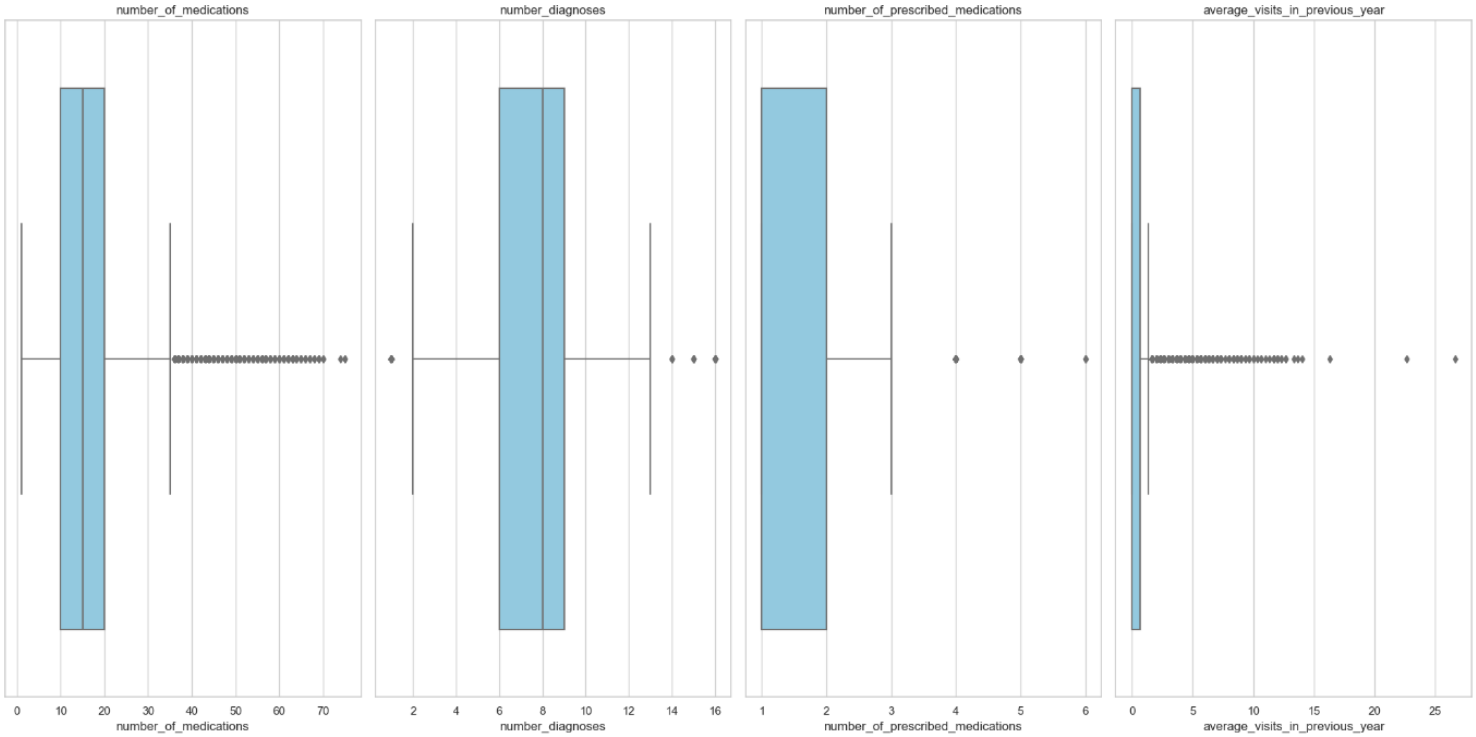
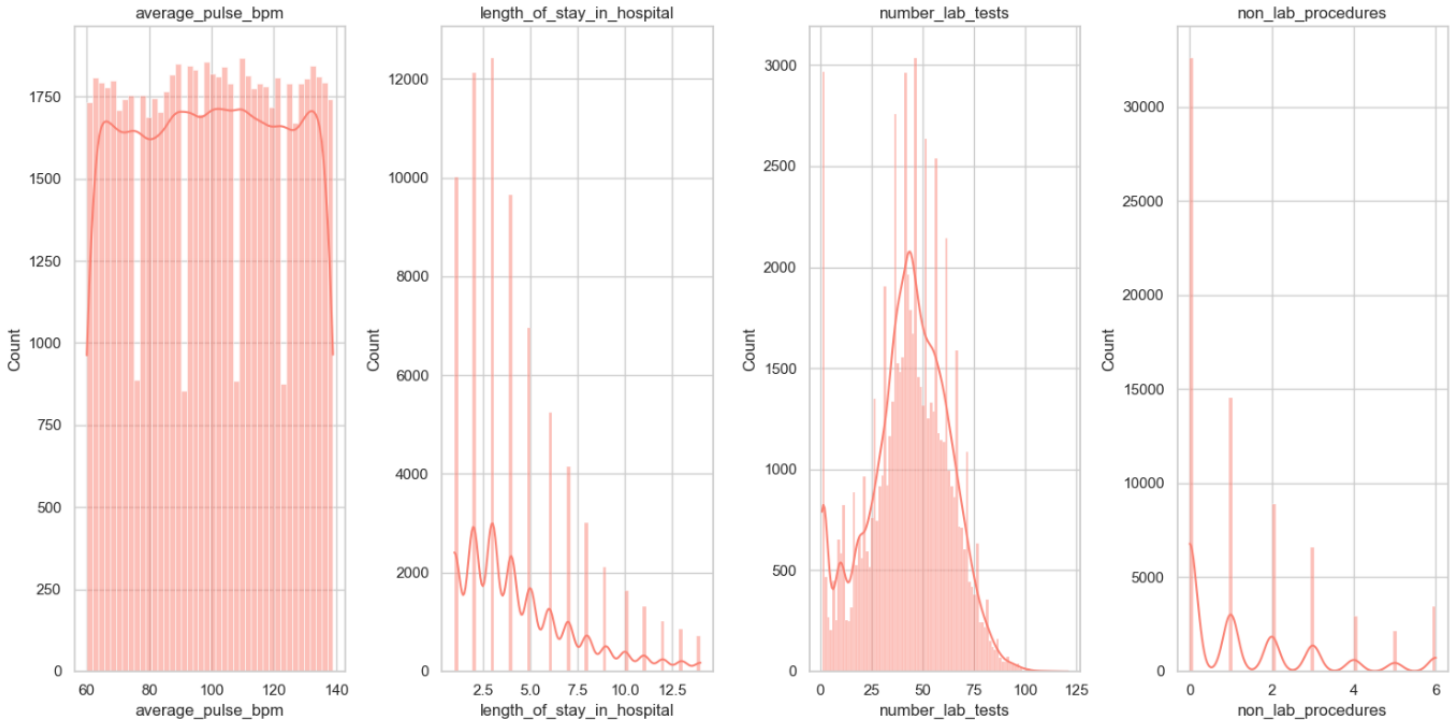



**Figure 2** - Boxplots of *'average_pulse_bpm', 'length_of_stay_in_hospital', 'number_lab_tests', 'non_lab_procedures', 'number_of_medications', 'number_diagnoses', 'number_of_prescribed_medications', 'average_visits_in_previous_year'*. The features' scales vary according to their distribution, and they are different from the Histogram's.

**Figure 3 – Histograms for outlier removal**



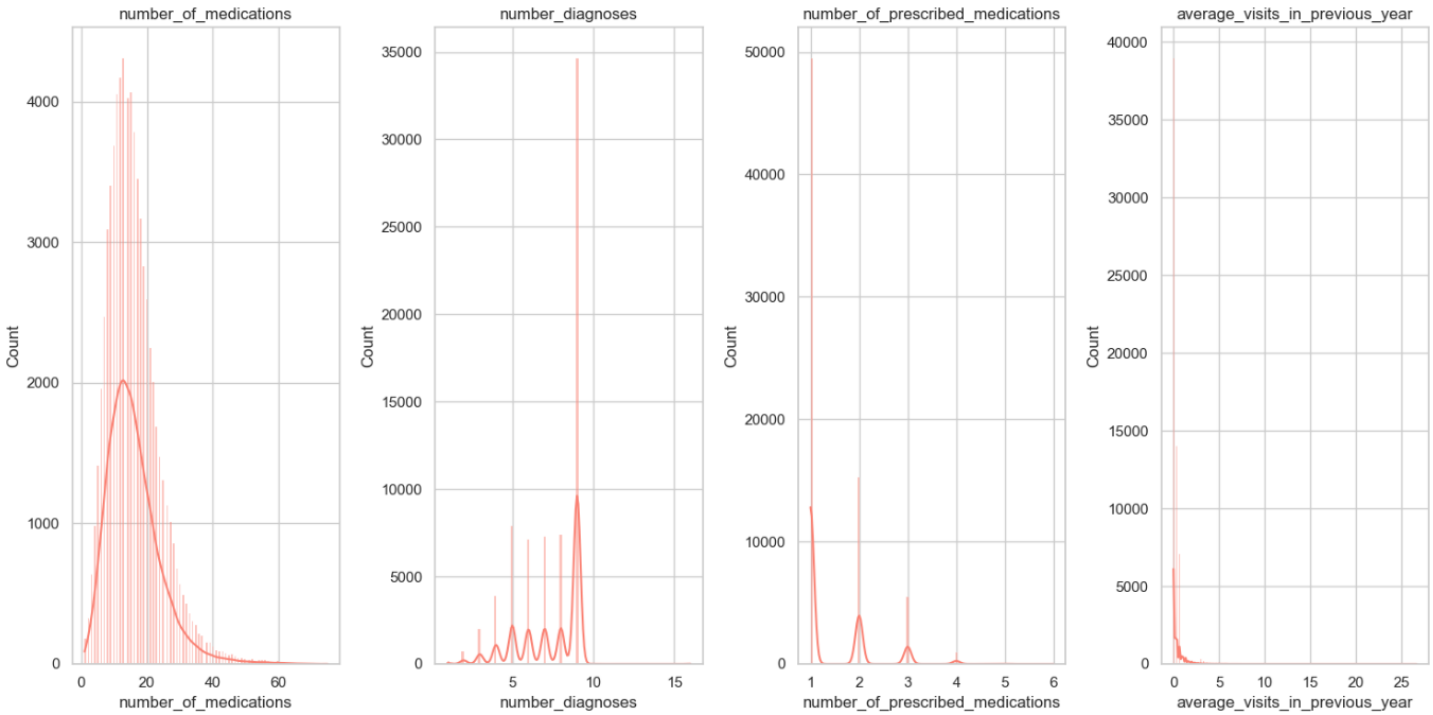



**Figure 3** - Histograms of '*average_pulse_bpm', 'length_of_stay_in_hospital', 'number_lab_tests', 'non_lab_procedures', 'number_of_medications', 'number_diagnoses', 'number_of_prescribed_medications', 'average_visits_in_previous_year'*. The features' scales vary according to their distribution, and they are different from the Boxplots'.

**Table 6 – Outlier thresholds**

| Feature Name | Threshold | Number of deleted records |
|---|---|---|
| *average_pulse_bpm* | No threshold | 0 |
| *length_of_stay_in_hospital* | No threshold | 0 |
| *number_lab_tests* | >110 | 8 |
| *non_lab_procedures* | No threshold | 0 |
| *number_of_medications* | >65 | 22 |
| *number_diagnoses* | >14 | 39 |
| *number_of_prescribed_medications* | >4 | 44 |
| *average_visits_in_previous_year* | >15 | 3 |

**Table 6** - This table lists the feature names that are linked with set thresholds, along with the count of records removed for each feature.

## Annex 2

**Feature Selection Methodology for Binary and Multiclass Classification**

**Filter Methods**

- **Pearson's correlation coefficient**

Pearson's correlation coefficient is a filter method for measuring the linear correlation between two continuous variables. It is computed as the ratio of the covariance of the two variables to the product of their standard deviations. It ranges between -1 (inverse variation) to 1 (positive variation), with 0 indicating no correlation. It is beneficial for feature selection since features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. By selecting these features one can reduce the dimensionality of the dataset. Limitations are present in the fact Pearson's correlation coefficient only captures linear relationships and fails to detect complex, non-linear relationships.

- **Kruskal-Wallis Test**

The Kruskal-Wallis test serves as a non-parametric and robust alternative to the conventional one-way ANOVA test, particularly valuable in situations where the assumption of normality is compromised. This statistical method employs the H-statistic to assess disparities among group medians, diverging from ANOVA, which examines means and can be sensitive to outliers. The process of the Kruskal-Wallis test commences by consolidating data from all groups and ranking it collectively, ranging from the smallest to the largest values. Each data point is then substituted with its corresponding rank. Subsequently, the ranks are averaged within each group. If these average ranks demonstrate significant differences across groups, the test infers a statistically significant distinction among them. Key attributes of the Kruskal-Wallis test include its suitability for ordinal or continuous response variables, along with its robustness in scenarios involving violations of normality assumptions. The test assumes independence among observations and expects similar distributions across the compared groups. In the realm of feature selection for numeric features, the Kruskal-Wallis test proves advantageous by providing a reliable means of identifying statistically significant differences in medians, facilitating the discernment of impactful features within a dataset.

- **Chi-Squared Test**

The Chi-Squared test evaluates the significance of associations between categorical non-ordinal features and a binary target variable. It measures the difference between the observed frequencies of the categories and the frequencies that would be expected if there was no association between the variables. Features with the highest Chi-squared statistics are selected, as these are the features that are most likely to be dependent on the target and therefore contain information that is useful for prediction. This selection process diverges from the null hypothesis of independence. The Chi-Squared test is particularly suitable for binary classification problems. If the target variable is continuous, it should be binned prior to the test. The test assumes that the observations are independent. If this assumption is violated, the test results may not be valid. Additionally, the Chi-Squared test requires a sufficient sample size to ensure the reliability of the results.

- **Mutual Information (MI) Test**

Mutual Information (MI) is a statistical metric that quantifies the degree of dependence between two variables. It measures the extent to which information, in the form of entropy, about one variable provides information about another variable. MI is particularly suitable for use in

classification problems. The values of MI are zero if and only if two random variables are independent, with higher values indicating a greater degree of dependency. It's important to note that MI operates under the assumption that the observations are independent. If this assumption is violated, the validity of the results may be compromised.

- **Kendall's Tau Test**

Kendall's Tau test is a rank correlation coefficient, which measures the degree of correspondence between the rankings of two variables. It is used to determine the strength of the relationship between two variables of ranked data. Kendall's Tau starts by ranking all the data from all groups together, from the smallest to the largest. For each feature, Kendall's Tau statistic is calculated between the feature and the target variable. This statistic measures the difference between the number of concordant pairs (C) and discordant pairs (D), normalized by the total number of pairs. The formula to calculate Kendall's Tau is follows: $\tau$ = (C-D)/(C+D) . It operates under the null hypothesis of the correlation coefficient Tau = 0, meaning there is no correlation between variables.

However, Kendall's Tau test has some limitations. It assumes that the observations are independent. If this assumption is violated, the test results may not be valid. Kendall's Tau is sensitive to outliers, where they can significantly affect the ranking of the data, which in turn can affect the calculations of Kendall's Tau.

**Wrapper Methods**

- **Recursive Feature Elimination (RFE)**

Recursive Feature Elimination (RFE) is a widely employed machine learning technique for selecting pertinent features in predictive modeling. The process involves ranking features based on their importance scores, eliminating the least important ones, and iteratively refining the model until reaching the desired feature count or stability. RFE is compatible with various supervised learning methods, often paired with models like Support Vector Machines (SVM) or Random Forests. The ranking metric is contingent on the chosen algorithm. However, RFE has limitations, it assumes independent observations. If this assumption is violated, the test results may be compromised.

**Embedded Methods**

- **Least Absolute Shrinkage and Selection Operator (LASSO) regression**

Lasso Regression, is a form of linear regression that incorporates shrinkage, pulling data values toward a central point, typically the mean. Its notable feature is effective feature selection, wherein coefficients can shrink to zero as the regularization parameter increases, removing features and enhancing model interpretability, mitigating overfitting risk. Leveraging cross-validation, a resampling technique, further refines Lasso Regression. In k-fold cross-validation, the training set is partitioned into K subsets, with the model trained on K-1 partitions and tested on the remaining subset. This process is repeated, and test errors are averaged. In the context of Lasso Regression, cross-validation aids in tuning the regularization parameter (lambda), eliminating the need for manual optimization, enhancing model performance, and ensuring generalization to unseen data.

- **Random Forest Feature Importance**

Random Forest is an ensemble learning technique that combines the results of multiple decision trees to produce a single outcome. It is particularly effective in handling categorical data. Feature importance in Random Forest is determined using a metric known as Gini importance. This metric measures the reduction in Gini impurity that results from splitting the data based on a particular feature. A higher Gini importance signifies a greater importance of the feature for the

model. However, it's crucial to note that Random Forests may overstate the importance of highly correlated features. Moreover, the optimization of the Random Forest Classifier Algorithm can be done by the use of Grid Search Cross-Validation. This technique systematically examines a range of hyperparameter values to find the subset that delivers the best model performance. In the context of Random Forest, cross-validation plays a key role in fine-tuning the regularization parameter, ensuring the selection of values that boost the overall performance of the model. It works by training the model with different parameter values on the K-1 partitions and evaluating on the Kth partition, one it finds the parameter value that gives the best model performance.

**Figure 1 – Peason's Correlation Matrix for Binary Classification**

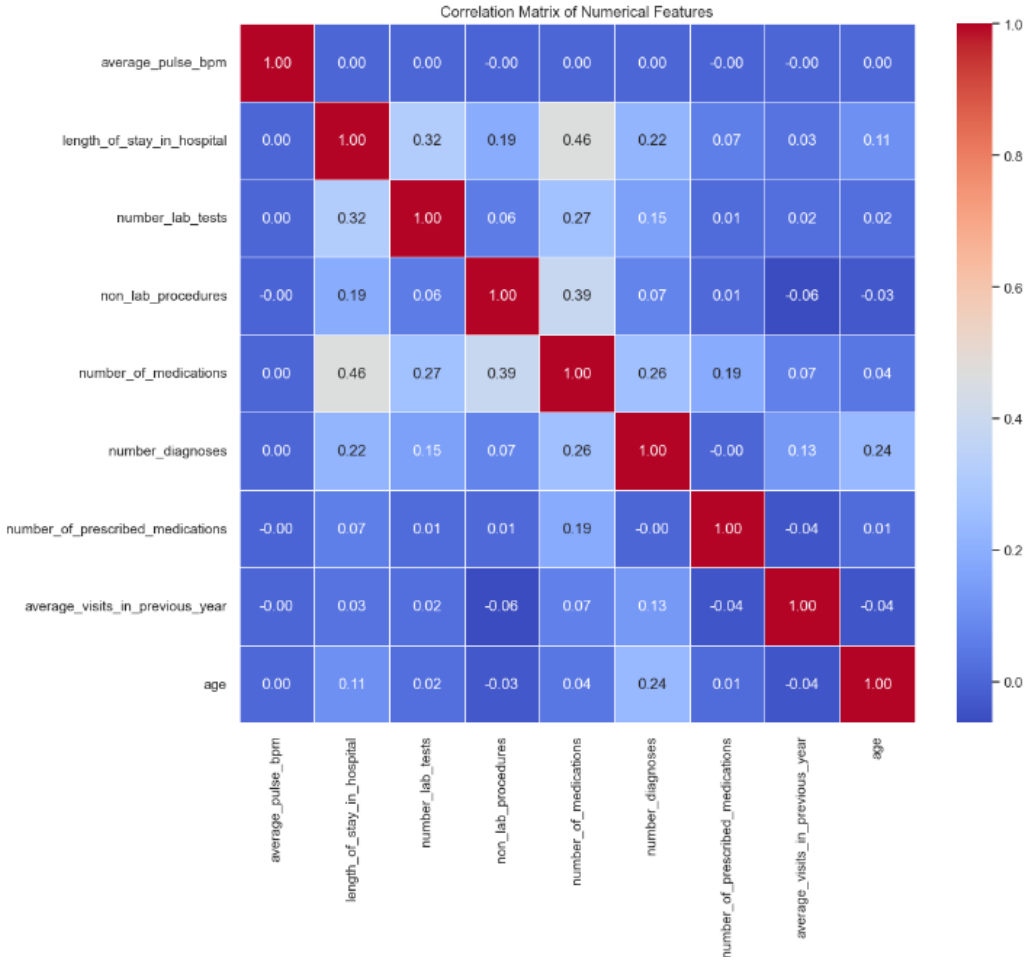



**Figure 1** – Peason's Correlation Matrix with the numeric features and categorical ordinal feature *('age')* for Binary Classification. Strong red represents strong positive correlations, strong blue represents strong negative correlations.

## Figure 2 – Results of the Kruskal-Wallis test for Binary Classification

| | Input_Features | Score | P_Value |
|---|---|---|---|
| 7 | average_visits_in_previous_year | 1146.887962 | 0.0000 |
| 5 | number_diagnoses | 191.235807 | 0.0000 |
| 1 | length_of_stay_in_hospital | 120.302321 | 0.0000 |
| 4 | number_of_medications | 99.932035 | 0.0000 |
| 8 | age | 28.731750 | 0.0000 |
| 2 | number_lab_tests | 25.169623 | 0.0000 |
| 3 | non_lab_procedures | 15.701909 | 0.0001 |
| 6 | number_of_prescribed_medications | 13.496956 | 0.0002 |
| 0 | average_pulse_bpm | 0.001468 | 0.9694 |

**Figure 2** – Kruskal-Wallis results of the H-test score and the correspondent p-values for each numeric feature plus the '*age*' feature.

## Figure 3 – Results of the Kendall's Tau test for Binary Classification

| | tau | p-value |
|---|---|---|
| average_pulse_bpm | 0.000125 | 0.967612 |
| length_of_stay_in_hospital | 0.041271 | 0.0 |
| number_lab_tests | 0.015521 | 0.0 |
| non_lab_procedures | -0.007319 | 0.031721 |
| number_of_medications | 0.038972 | 0.0 |
| number_diagnoses | 0.044692 | 0.0 |
| number_of_prescribed_medications | -0.012623 | 0.000504 |
| average_visits_in_previous_year | 0.116524 | 0.0 |
| age | 0.019544 | 0.0 |

**Figure 3** – Kendall's Tau results of the Tau score and the correspondent p-values for each numeric feature plus the '*age*' feature.

## Figure 4 – Plots showing the F1-score on Training and Validation Set for the Optimal Number of Features for Recursive Feature Selection Using the 3 different algorithms for Binary Classification

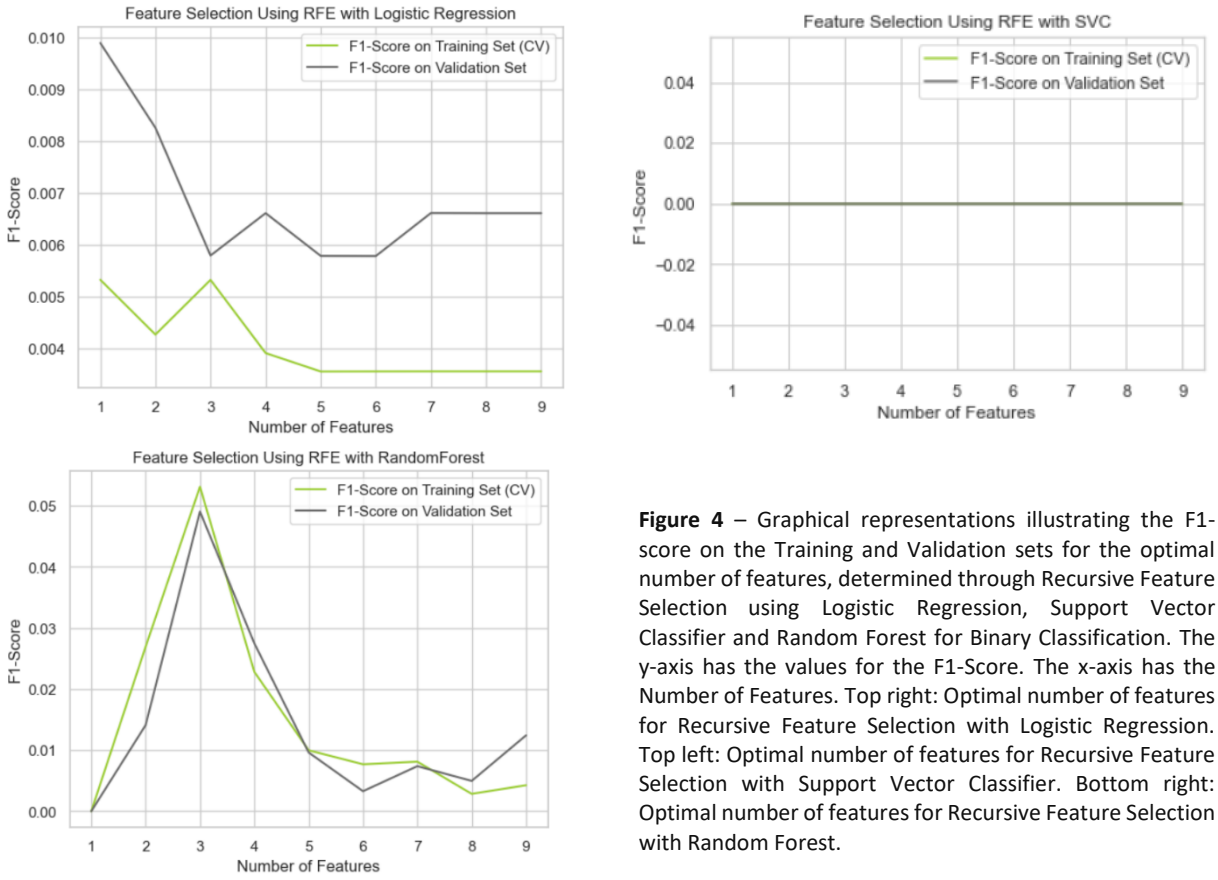



**Figure 4** – Graphical representations illustrating the F1-score on the Training and Validation sets for the optimal number of features, determined through Recursive Feature Selection using Logistic Regression, Support Vector Classifier and Random Forest for Binary Classification. The y-axis has the values for the F1-Score. The x-axis has the Number of Features. Top right: Optimal number of features for Recursive Feature Selection with Logistic Regression. Top left: Optimal number of features for Recursive Feature Selection with Support Vector Classifier. Bottom right: Optimal number of features for Recursive Feature Selection with Random Forest.

## Figure 5 – Bar plot of Feature Importance using Lasso Regression for Binary Classification

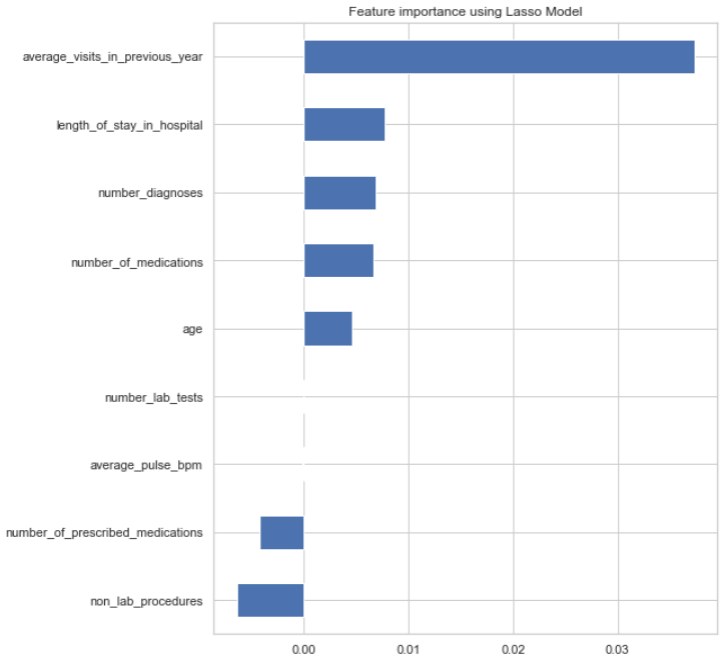



**Figure 5** – Graphical representations illustrating Lasso Regression Feature importance for Binary Classification. The y-axis has the name of the numerical features plus the 'age' feature. The x-axis has the values of the L1 regularization

## Table 1 & 2 – Feature Selection for Binary Classification Final Results

| PREDICTOR | KRUSKAL-WALLIS | RFE LOGISTIC REGRESSION | RFE SUPPORT VECTOR MACHINE | RFE RANDOM FOREST | LASSO | KENDALL'S TAU |
|---|---|---|---|---|---|---|
| *AVERAGE_PULSE_BPM* | Discard | No trust | No trust | No trust | Discard | Discard |
| *LENGTH_OF_STAY_IN_HOSPITAL* | Keep | No trust | No trust | No trust | Keep | Keep |
| *NUMBER_LAB_TESTS* | Keep | No trust | No trust | No trust | Discard | Keep |
| *NON_LAB_PROCEDURES* | Keep | No trust | No trust | No trust | Keep | Keep |
| *NUMBER_OF_MEDICATIONS* | Keep | No trust | No trust | No trust | Keep | Keep |
| *NUMBER_DIAGNOSES* | Keep | No trust | No trust | No trust | Keep | Keep |
| *NUMBER_OF_PRESCRIBED_MEDICATIONS* | Keep | No trust | No trust | No trust | Keep | Keep |
| *AVERAGE_VISITS_IN_PREVIOUS_YEAR* | Keep | No trust | No trust | No trust | Keep | Keep |
| *AGE* | Keep | No trust | No trust | No trust | Keep | Keep |

| PREDICTOR | CHI-SQUARED | RANDOM FOREST | MUTUAL INFORMATION |
|---|---|---|---|
| *RACE* | Discarded | No trust | No trust |
| *GENDER* | Discarded | No trust | No trust |
| *PAYER_CODE* | Keep | No trust | No trust |
| *MEDICAL_SPECIALTY* | Keep | No trust | No trust |
| *DISCHARGE_DISPOSITION* | Keep | No trust | No trust |
| *ADMISSION_SOURCE* | Keep | No trust | No trust |
| *PRIMARY_DIAGNOSIS* | Keep | No trust | No trust |
| *SECONDARY_DIAGNOSIS* | Keep | No trust | No trust |
| *ADDITIONAL_DIAGNOSIS* | Keep | No trust | No trust |
| *CHANGE_IN_MEDS_DURING_HOSPITAL* | Keep | No trust | No trust |
| *PRESCRIBED_DIABETES_MEDS* | Keep | No trust | No trust |
| *GLUCOSE_TEST_RESULT* | Keep | No trust | No trust |
| *A1C_TEST_RESULT* | Keep | No trust | No trust |

**Table 1 & 2** – Summary of the retained, discarded, and unreliable features for selection in the final results of Binary Classification for each model.

## Feature Encoding – Binary and Multiclass Classification

### Table 2 – Names of the missing features in the training and testing datasets

| *Features present in the train dataset that are absent in the test dataset* | *Features present in the test dataset that are absent in the train dataset* |
|---|---|
| – 'glimepiride-pioglitazone'<br>– 'metformin-pioglitazone'<br>– 'acetohexamide'<br>– 'readmitted binary'<br>– 'readmitted multiclass' | – *'medical specialty dermatology'* |

**Table 2** – Summary of the names absent between training and testing datasets after feature encoding of the *'medication'* feature.

### Figure 6 – Peason's Correlation Matrix for Multiclass Classification

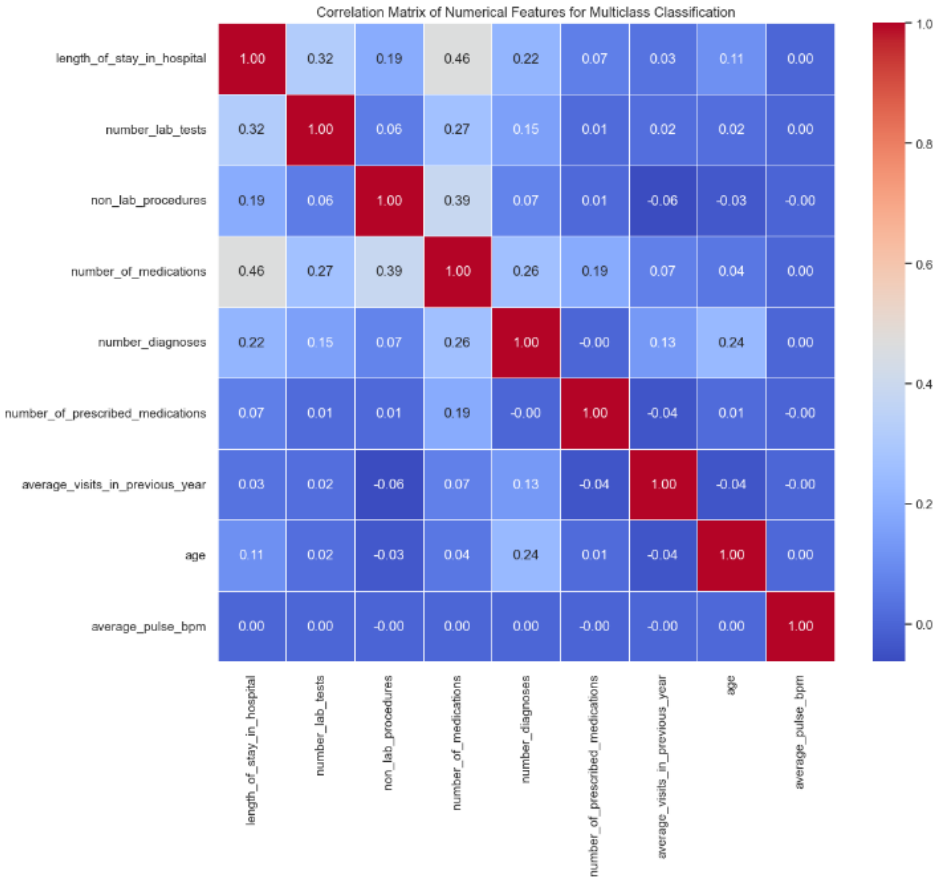



**Figure 6** – Peason's Correlation Matrix with the numeric features and categorical ordinal feature *('age')* for Multiclass Classification. Strong red represents strong positive correlations, strong blue represents strong negative correlations.

## Figure 7 – Results of the Kruskal-Wallis test for Multiclass Classification

| | Input_Features | Score | P_Value |
|---|---|---|---|
| 6 | average_visits_in_previous_year | 1707.600492 | 0.0000 |
| 4 | number_diagnoses | 494.883577 | 0.0000 |
| 0 | length_of_stay_in_hospital | 109.421455 | 0.0000 |
| 3 | number_of_medications | 91.347464 | 0.0000 |
| 2 | non_lab_procedures | 75.310011 | 0.0000 |
| 1 | number_lab_tests | 54.850049 | 0.0000 |
| 7 | age | 41.660057 | 0.0000 |
| 5 | number_of_prescribed_medications | 8.112430 | 0.0003 |
| 8 | average_pulse_bpm | 1.425456 | 0.2404 |

**Figure 7** – Kruskal-Wallis results of the H-test score and the correspondent p-values for each numeric feature plus the '*age*' feature.

## Figure 8 – Results of the Kendall's Tau test for Multiclass Classification

| | tau | p-value |
|---|---|---|
| length_of_stay_in_hospital | 0.042592 | 0.0 |
| number_lab_tests | 0.030538 | 0.0 |
| non_lab_procedures | -0.041432 | 0.0 |
| number_of_medications | 0.050249 | 0.0 |
| number_diagnoses | 0.093246 | 0.0 |
| number_of_prescribed_medications | 0.006041 | 0.083812 |
| average_visits_in_previous_year | 0.200476 | 0.0 |
| age | 0.025187 | 0.0 |
| average_pulse_bpm | -0.00493 | 0.096255 |

**Figure 8** – Kendall's Tau results of the Tau score and the correspondent p-values for each numeric feature plus the '*age*' feature.

## Figure 9 – Plots showing the F1-score on Training and Validation Set for the Optimal Number of Features for Recursive Feature Selection Using the 3 different algorithms for Multiclass Classification

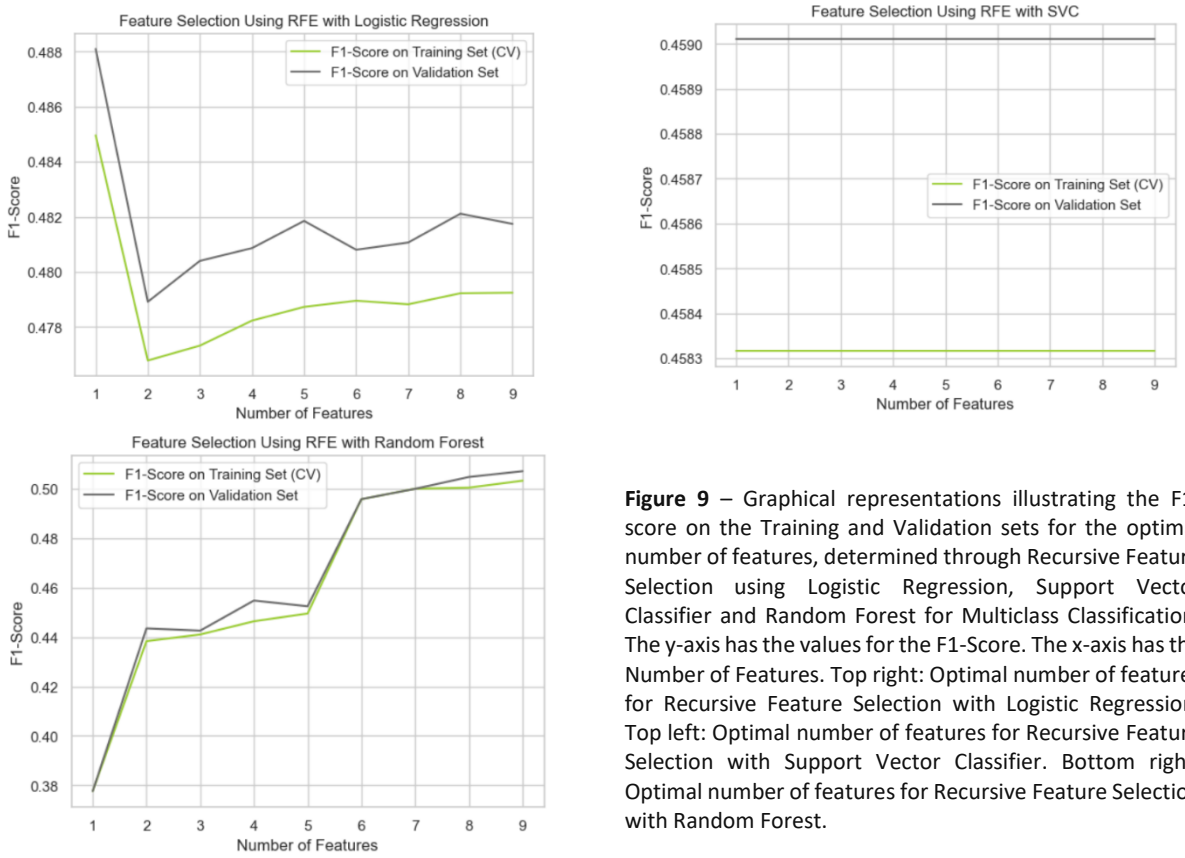



**Figure 9** – Graphical representations illustrating the F1-score on the Training and Validation sets for the optimal number of features, determined through Recursive Feature Selection using Logistic Regression, Support Vector Classifier and Random Forest for Multiclass Classification. The y-axis has the values for the F1-Score. The x-axis has the Number of Features. Top right: Optimal number of features for Recursive Feature Selection with Logistic Regression. Top left: Optimal number of features for Recursive Feature Selection with Support Vector Classifier. Bottom right: Optimal number of features for Recursive Feature Selection with Random Forest.

## Figure 10 – Bar plot of Feature Importance using Lasso Regression for Binary Classification

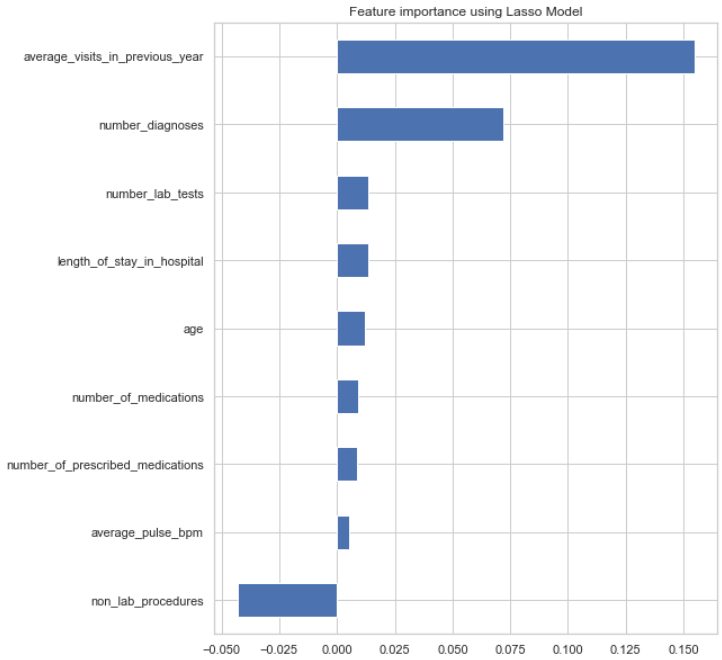



**Figure 10** – Graphical representations illustrating Lasso Regression Feature importance for Multiclass Classification. The y-axis has the name of the numerical features plus the 'age' feature. The x-axis has the values of the L1 regularization

## Table 3 & 4 – Feature Selection for Multiclass Classification Final Results

| PREDICTOR | KRUSKAL-WALLIS | RFE LOGISTIC REGRESSION | RFE SUPPORT VECTOR MACHINE | RFE RANDOM FOREST | LASSO | KENDALL'S TAU |
|---|---|---|---|---|---|---|
| *AVERAGE_PULSE_BPM* | Discard | Discard | Discard | Keep | Keep | Discard |
| *LENGTH_OF_STAY_IN_HOSPITAL* | Keep | Discard | Discard | Keep | Keep | Keep |
| *NUMBER_LAB_TESTS* | Keep | Discard | Discard | Keep | Keep | Keep |
| *NON_LAB_PROCEDURES* | Keep | Discard | Discard | Keep | Keep | Keep |
| *NUMBER_OF_MEDICATIONS* | Keep | Discard | Discard | Keep | Keep | Keep |
| *NUMBER_DIAGNOSES* | Keep | Discard | Discard | Keep | Keep | Keep |
| *NUMBER_OF_PRESCRIBED_MEDICATIONS* | Keep | Discard | Discard | Keep | Keep | Discard |
| *AVERAGE_VISITS_IN_PREVIOUS_YEAR* | Keep | Keep | Keep | Keep | Keep | Keep |
| *AGE* | Keep | Discard | Discard | Keep | Keep | Keep |

| PREDICTOR | CHI-SQUARED | RANDOM FOREST | MUTUAL INFORMATION |
|---|---|---|---|
| *RACE* | Keep | No trust | No trust |
| *GENDER* | Keep | No trust | No trust |
| *PAYER_CODE* | Keep | No trust | No trust |
| *MEDICAL_SPECIALTY* | Keep | No trust | No trust |
| *DISCHARGE_DISPOSITION* | Keep | No trust | No trust |
| *ADMISSION_SOURCE* | Keep | No trust | No trust |
| *PRIMARY_DIAGNOSIS* | Keep | No trust | No trust |
| *SECONDARY_DIAGNOSIS* | Keep | No trust | No trust |
| *ADDITIONAL_DIAGNOSIS* | Keep | No trust | No trust |
| *CHANGE_IN_MEDS_DURING_HOSPITAL* | Keep | No trust | No trust |
| *PRESCRIBED_DIABETES_MEDS* | Keep | No trust | No trust |
| *GLUCOSE_TEST_RESULT* | Keep | No trust | No trust |
| *A1C_TEST_RESULT* | Keep | No trust | No trust |

**Table 3 & 4** – Summary of the retained, discarded, and unreliable features for selection in the final results of Binary Classification for each model.

## Models for Binary and Multiclass Classification

### Gaussian Naive Bayes

The Gaussian Naive Bayes (GNB) algorithm is a popular machine learning classifier that operates on the principles of Bayes' theorem, specifically designed for handling large feature sets efficiently. It belongs to the family of Naive Bayes classifiers, which assume that the features used to describe instances are conditionally independent given the class label. The 'naive' assumption simplifies the computation and enables the algorithm to scale effectively to datasets with a substantial number of features. In the context of GNB, it is particularly well-suited for continuous data where features follow a Gaussian (normal) distribution. This makes it applicable to a wide range of real-world problems, especially those involving numerical and continuous variables. The algorithm calculates the probability of a given instance belonging to a particular class based on the conditional probabilities of each feature given the class. One of the strengths of GNB lies in its simplicity and computational efficiency. The independence assumption allows the algorithm to estimate parameters for each feature independently, significantly reducing the computational burden compared to more complex models. This makes GNB particularly useful in scenarios where computational resources are limited or when dealing with large-scale datasets. However, the assumption of independence may not always hold in real-world data, and this is a limitation of the model. Despite this simplification, GNB often performs remarkably well in practice, especially when the independence assumption is not severely violated. To enhance the performance of the GNB model, hyperparameter tuning was employed to identify an optimal configuration. Hyperparameter tuning involves systematically searching through a predefined set of hyperparameter values to find the combination that yields the best model performance. In the case of GNB, hyperparameters such as the smoothing parameter (if applicable) and other configuration options were fine-tuned to achieve optimal classification accuracy. The application of hyperparameter tuning ensures that the GNB model is finely adjusted to the characteristics of the specific dataset, maximizing its predictive capabilities. This iterative process involves training and evaluating the model with different hyperparameter configurations to find the combination that strikes the right balance between bias and variance, resulting in a well-generalized and robust model. In conclusion, the Gaussian Naive Bayes algorithm, with its inherent simplicity and efficiency, coupled with careful hyperparameter tuning, presents a powerful tool for classification tasks involving a large number of features. Its ability to handle continuous data and scalability make it a versatile choice in various domains, from text classification to medical diagnosis, where interpretability and computational efficiency are paramount.

**Histogram-based Gradient Boosting Classification Tree**

Histogram-Based Gradient Boosting (HGB) is an advanced ensemble learning technique that falls within the broader category of boosting algorithms. Developed as an evolution of traditional gradient boosting methods, HGB introduces innovative strategies to enhance predictive accuracy, especially in the context of large and complex datasets. At its core, HGB leverages the principles of decision trees, creating an ensemble of these trees to collectively improve the model's overall performance. The boosting algorithm is characterized by its sequential training process, where each tree is built to correct the errors made by its predecessors. This iterative refinement of the model enables the algorithm to learn complex relationships within the data and adapt to non-linear patterns effectively. The sequential nature of boosting allows the model to focus on instances that were misclassified in previous iterations, leading to a progressive improvement in predictive accuracy with each added tree. One of the key advantages of HGB is its efficiency, particularly when dealing with large datasets. Traditional gradient boosting algorithms may become computationally expensive as the dataset size increases, but HGB addresses this challenge by utilizing histogram-based techniques to efficiently represent and manipulate the data. This leads to faster training times and makes HGB well-suited for applications where both computational efficiency and model accuracy are crucial considerations. Another notable feature of HGB is its intrinsic ability to handle missing data and categorical features. The histogram-based approach allows the algorithm to efficiently process and split the data based on feature histograms, accommodating missing values without the need for imputation. Additionally, categorical features can be incorporated seamlessly into the decision tree structure, eliminating the need for one-hot encoding, and simplifying the preprocessing steps. The combination of boosting principles, histogram-based optimizations, and robust handling of missing data and categorical features makes HGB a powerful tool in various machine learning applications. It excels in scenarios where the data is diverse, non-linear relationships are prevalent, and computational efficiency is a priority. To further enhance the performance of the HGB model, hyperparameter tuning can be employed. Adjusting parameters such as the learning rate, tree depth, and regularization settings allows practitioners to fine-tune the model's behavior to match the characteristics of the specific dataset. This iterative tuning process ensures that the HGB algorithm is optimized for the task at hand, striking an appropriate balance between model complexity and generalization. In summary, Histogram-Based Gradient Boosting stands out as an advanced ensemble learning method that effectively addresses the challenges posed by large datasets, missing data, and categorical features. Its sequential learning process and innovative histogram-based techniques contribute to its ability to capture intricate patterns within the data, making it a valuable tool in the machine learning practitioner's toolkit.