Machine Learning

DISCHARGED:

EXAMINING HOSPITAL READMISSION

Group 45

December 2023

Catarina Reis, 20230981

Guilherme Curioso, 20230532

Eduardo Costa, r20201536

Mariana Cabral, 20230532

Tomás Castilho, 20230518

Prof. Roberto Rodrigues | Prof. Ricardo Santos | Prof. Rafael Pereira

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

ABSTRACT

This project, part of the Machine Learning course, addresses the pressing issue of hospital readmissions within the healthcare sector, recognized for its impact on care quality and costs. Our focus is on developing predictive models using patient and encounter information to answer key classification questions: identifying patients prone to readmission within 30 days post-hospital discharge and determining the precise readmission timeframe ('No,' '< 30 days,' or '>30 days').

Inspired by related research on disease-specific 30-day readmission rates in the US, our primary goal is to empower healthcare providers with personalized post-discharge care insights derived from machine learning models.

Throughout the project, we navigated challenges, notably computing limitations and the absence of medical expertise. Limited grid searches, necessitated by computational constraints, highlighted the trade-off between accuracy and runtime efficiency. The lack of medical expertise raised concerns about potential errors in feature selection, underscoring the importance of collaboration with domain experts in critical health prediction domains.

Despite challenges, the project proved highly educational, fostering deep discussions and investigations. In conclusion, our main findings underscore the need for collaborative efforts with medical experts to enhance model robustness. We suggest the implementation of stacking models, an ensemble learning technique, to overcome limitations, capitalize on individual model strengths, and mitigate weaknesses. Continuous refinement and optimization of individual models are crucial for advancing the accuracy and reliability of future predictive models in healthcare settings.

CONTENTS

1. INTRODUCTION

With the present study of the curricular unit of Machine Learning, we intend to analyse hospital readmissions, well-known for being a significant challenge in the healthcare sector, because they are not only an indicator of care quality but also, a driver of escalating costs.

The objective of this project is to develop predictive models using patient and encounter characteristics to address two critical classification inquiries: firstly, the identification of patients prone to readmission within 30 days post-hospital discharge; secondly, the determination of the precise timeframe of readmission, categorized as 'No,' '< 30 days,' or '>30 days'. Keeping that in mind, our main contribution involves using the models to extract valuable insights regarding patient risk levels, so that we could empower healthcare providers to implement a customization of post-discharge care.

By looking into analogous research, we found one[1] that uses machine learning models to predict disease-specific 30-day readmission rates in the US and also investigates the impact of factors such as demographics, disease types, hospital specialty, ownership, and locations on readmission rates.

In this context, the study revealed valuable insights. On the one hand, it identified correlations between readmission rates and factors such as demographics (e.g., gender, age), disease types, and specific hospital-related factors. For instance, readmission rates varied significantly across different diseases: ranging from 1.832% for Pneumonia to 8.761% for Diabetes. On the other hand, regarding machine learning and predictive modelling, the study highlighted the effectiveness of ensemble techniques, particularly Gradient Boosting, in predicting disease-specific hospital readmission outcomes.

2. EXPLORATION

For this research, we had available two datasets one for the training of our models and another for the model's assessment and testing. We directed our initial focus on the training dataset to gain a comprehensive understanding of the data.

Our training dataset has 71,236 records and 31 unique features. In the initial phase of our project, we conducted an exploration to identify potential issues and assess data coherence. For that purpose, we started by searching for undesirable characters and duplicate records, followed by an analysis of the data types, and a general overview of the missing values, concluding with some data statistics and visual analysis.

Firstly, we chose to use the *encounter_id* as the index instead of the *patient_id*, since the same patient could visit the hospital more than once and each record represents an encounter. Our next step was to look for duplicate records, which we didn´t find any.

Regarding the undesirable characters and some specific issues in the feature's data types our approach was to treat them as missing values. More specifically, "Unknown/Invalid" in *gender*, and "Not Available" in *admission_type* and *admission source*. In the case of *a1c_test_result* and *glucose_test_result*, the missing values were replaced with "None" to indicate that the respective tests were not conducted, being accurate with the feature's description and offering a more meaningful representation for machine learning analysis.

After analyzing the presence of missing values in our features, we found some critical cases such as *weight*, *medical_specialty,* and *payer_code,* with 96.8%, 49.0%, and 39.6%, respectively. The case of *payer_code* wasn´t considered as a problematic one, because we assumed that a missing value represented no use of insurance. Moreover, less significant missing values were found, all below 10%.

For further investigation, we decided to divide our variables into numerical and categorical [Table 1: Definition of numerical and categorical features] to structure our statistical and visual analysis.

Starting with the numerical features, we focused on differences in the values of some key metrics such as mean, median, max, and min to identify potential outliers. Some critical cases such as *outpatient_visits_in_previous_year, emergency_visits_in_previous_year, inpatient_visits_in_previous_year,* and *number_of_medications,* made us suspect a larger presence of outliers which was corroborated by looking the respective histograms, which had skewed data, and the box plots.

Alternatively, the analysis of categorical features revealed some more information, for instance, the fact that all records are from patients living in the USA, which means that this data was extracted from the same country as the one in the previous research that we found about this topic, and that makes us more confident about exploring their results. Thus, it tells us that *country* is a redundant feature, so that, we will exclude this feature after the feature selection for further analysis. Furthermore, some additional insights from the categorical data include:

- **Unbalanced data** on the **classification of the readmission** (predominance of negative cases, "No", with 88.83 %) and on **classes for timeline readmissions** (predominance of negative cases over the ones with "<30 days'" with 53.91%).
- **Patient's most common profile:** *gender* "Female", *race* "Caucasian", with an *age* within "[70-80)", that is diabetic (since the most frequent class for *prescribed_diabetes_meds* is "Yes").
- **Most common encounter's characteristics:** *admission_type* "Emergency", *medical_specialty,* and "Internal Medicine", the top medication administrated during the encounter was "['insulin']".

Before we get into the preprocessing, we decided to display some visualizations to better understand the data and identify patterns and relevant correlations.

First, we examined the distributions of numerical features, revealing the presence of outliers [Figure 1: Numeric Variables' Histograms Before Outlier Removal]. The skewed nature of the majority of graphs was evident, and boxplots further confirmed the existence of noise [Figure .2: Numeric Variables' Box Plots before Outlier Removal]. In terms of correlations between numerical features [Figure 3: Correlation Matrix], we didn´t find any meaningful value in the matrix that highlighted potential correlations between certain numerical features.

Moving on to categorical features, we decided to analyze the contribution of each value of a categorical feature to the classification of the status of the readmissions ("Yes" or "No"). For that purpose, we displayed some graphs that show for each categorical feature the number of observations for each value, and the status of the readmission [Figure 4: Influence of categorical features values in the status of the readmitted]. By looking at this, we extracted

some useful insights, such as the values for each feature that showed a relevant contribution for a certain patient to be readmitted: *race* "Caucasian", followed by "AfricanAmerican"; *age* within "[20-30)"; *a1c_test_result* "None" *glucose_test_result* ">300"; prescribed_diabetes_meds "Yes".

Moreover, we finalize our analyses by looking at the 15 most common combinations of medication administrated during the encounter [Figure 5: Top 15 Medication Combinations], being insulin and no medication the ones that stand out the most, which once again reinforces the presence of diabetic patients in our sample.

3. DATA PREPARATION AND PREPROCESSING

Regarding the data preprocessing phase, our focus was on meticulous data cleaning to ensure we had a well-prepared dataset, setting the stage for testing our models.

1. DATA COHERENCE

Initially, we analysed instances where medications for diabetes were prescribed, but the *medication* list was empty, and fortunately, no such discrepancies were found.

Additionally, we looked for inconsistencies in *gender*, *race* and *age* between records linked to the same patient. In the case of different genders, we decided to consider the first appearance. Since we had more than two races, we decided to select the most frequently occurring race. In the case of a tie, we opted for the first appearance, as well. Finally, we also found a potential incoherence with the *age*, but since we don't have a temporal relationship between encounters, we can't assume that one visit was before the other. Knowing that the *encounter_id* is random, a higher id does not indicate a later encounter. Thus, we will not consider it an incoherence.

2. FEATURE ENGINEERING

Our original dataset had a lot of features in formats that were difficult to work with, so in order to simplify our analysis and extract more valuable information from our data, we created the following features:

- ***insurance***: Based on the *payer_code* description, we assumed that a missing value indicated a lack of insurance. So, we introduced a new binary variable that takes the value 1 if a patient has insurance and 0 otherwise.
- ***T_num_visits***: Since we have multiple encounters for each patient, and the order of these visits is unknown, we've introduced a variable that represents the total number of visits. This variable is calculated as the sum of all encounters associated with the same patient.

Additionally, we made some changes in certain features to reduce their cardinality, so we could get a clearer interpretation:

- ***primary_diagnosis***, ***secondary_diagnosis, additional_diagnosis***: All of these features are related to medical diagnosis represented in a format of code given by

the ICD-9-CM[1]. Since there were a lot of different codes related to the same type of disease, we aggregated them according to the ICD-9-CM**.** Regarding the diseases that barely appeared (less than 5% in all the diagnosis), we combined them in the category "other diseases". Thus, we were left with only 9 unique values, in contrast to the more than 600 values that these three variables could take individually.

- ***medication****:* This feature was presented as a list of strings, displaying different combinations of medications. We thought it would be more useful to create variables with the most administrated medications (which appeared in our data set at a rate above 5%), using a similar approach to one-hot encoding. So, if during an encounter the patient was associated with certain medications, the value would be one to those particular medications, otherwise, it would be 0.

3. SPLITTING DATA

Before further transformations and to ensure that we provide a reliable model assessment, we needed to slip our data set into two sets: validation and training.

It´s important to mention that even though we had two variables we aimed to predict (*readmitted_multiclass* and *readmitted_binary*), it was irrelevant the target selected to make the split at this stage. It is in the Feature Selection that the approach will change according to each target variable.

4. OUTLIERS

Taking into consideration the boxplots and histograms analysed in Data Visualization, we already know that some cases stood out, having a concerning presence of outliers. To solve this issue and also search for more outliers that were not already detected, we started by applying some methods to deal with the outlier removal, including **IQR Method**, **Z-Score Method, Manual Filters,** remaining with approximately 62.20%, 92.68% and 98.77% of the data, respectively.

To enhance accuracy, we combined all methods, removing rows identified as outliers by all techniques. This yielded a dataset with 98.77% retention, ensuring a more robust outlier removal strategy. Then we checked the result in the boxplots after outlier removal [Figure 6: Numerical Variables' Boxplots After Outlier Removal].

5. SCALING

To make our data suitable for machine learning models, we needed to make sure that all our data were on the same scale. We chose **MinMaxScaler**, scaling values between 0 and 1. Other methods like Standard Scaler assumed a normal distribution, which didn't fit our data, and Robust Scaler wasn't needed as we had already addressed outliers separately.

[1] [1] *ICD-9-CM refers to the* **International Classification of Diseases, 9th Revision, Clinical Modification***, a coding system widely used for classifying medical conditions and procedures. It is commonly employed in medical records and clinical research.*

6. MISSING VALUES

In the previous steps, we conducted some data transformation so that, there were significant differences in the number of features with missing data and in the percentage of that data for each feature [Table 2: Missing Values Preprocessing]. Given that scenario, we started by removing the *weight* feature, which exhibited approximately 97% missing values.

Upon realizing that only categorical variables in our dataset had missing values, we proceeded with the following approach: for features that could be shared among patients with the same id (*gender* and race) we filled them with the values of the available value for the corresponding patient; then, for features with missing values below the 5% rate, we opted for a straightforward imputation strategy, filling the missing values with the **mode** of the respective feature. For the most critical cases (features with more than 5% rate of missing values – *medical_specialty*, *admission_type, admission_source*, and *race*) we used the **KNNClassifier**, meaning that we treated each of those critical features as the target variable and used numerical features as training data for a KNN model, filling the missing values with the predictions.

7. ENCODING

In our data preprocessing, we employed specific encoding techniques according to the characteristics of the variables. **Ordinal encoding** was applied to *age* and test results (*ac1_test_result* and *glucose_test_result*) to represent their inherent order or scale. For the other categorical variables, all of which were nominal, we adopted **One-Hot Encoding**. This involved creating a new binary column for each unique value, effectively transforming categorical data into a binary format. To streamline the dataset and reduce complexity, categories occurring below a 1% threshold were grouped into an "Other" category.

8. FEATURE SELECTION

The more we progressed in our research, the more we realized that the main significant improvements in our model results would lie within this step. Hence, it became crucial to define a clear and efficient strategy. Our approach relied on the following steps: beforehand, we divided this section into two, each focused on a specific target. In each of these sections, we applied methods tailored to a specific type of variable—first numerical, then categorical—enabling us to remove certain features in advance. To conclude, on the remaining variables, we opted to apply methods capable of handling both types.

In more detail, for the numerical variables we start by applying the **univariate method** that verified the variance of numerical variables to ensure none exhibited zero variance (constant values); then the **Correlation of Spearmen**, ending with the **ANOVA Test**.

The results associated with the first two methods revealed that: none of the numerical variables was a constant, although the *emergency_visits_in_previous_year*, *outpatient_visits_in_previous_year* and *T_num_visits*, demonstrated relatively poor variance. In terms of correlation, no high correlation was verified, the only one that stood out, nevertheless not high enough, was between *length_of_stay_in_hospital* and *number_of_medications* and between *T_num_visits* and inpatient_*visits_in_previous_year,* 0.5 in both cases [Figure 7: Spearmen's Correlation Matrix]. Regarding the third method, as we employed ANOVA utilizing the F-value test scores, a higher F-score indicates a feature that

better predicts the target. For the binary classification, four features presented really low values, *number_lab_tests non_lab_procedures, average_pulse_bpm, outpatient_visits_in_previous_year,* thus they were eliminated [Figure 8: ANOVA Test Binary Target]. In the multiclass classification, just the *number_of_medications* was removed [Figure 9: ANOVA Test Multiclass Target].

Concerning categorical variables, we utilized the **Chi-square test** to assess the independence between these variables and understand if a statistically significant relationship exists between a feature and the target class. To assess this, we considered significance levels of 5% (for binary) [Table 3: Chi-Square Test Binary – Insignificant Variables] and 10% (for multiclass) [Table 4: Chi-square Test Multiclass – Insignificant Variables]. We concluded by removing all variables with p-values above the designated level of significance. In these cases, the failure to reject the null hypothesis indicates a lack of statistical evidence to support the significance of a particular feature in predicting the target.

Finally, in terms of feature selection in both types, we used several methods in parallel, such as **Mutual Information**, **Lasso**, and **Embedded Methods**, using **Logistic regression**, and **Random Forest.** After reviewing the combined results, we reached a consensus to begin by removing all the features that were rejected by every method and end the feature selection with 34 features [Table 5: Combined Result for Binary Target] for the binary classification and 16 [Table 6: Combined Result for Multiclass Target] for the multiclass classification. Other combinations of possible removals at this step were tried, however, we remained with this, because it allowed us to get better results on the model section.

4. MODELLING

Moving forward to the application of the models, we started with a simple approach: test every model learned during classes with their default parameters (except *class_weight* = "balanced" which we felt was necessary to account for class imbalance since it gives higher weights to the class with fewer observations) to see which ones were promising enough to pursue and perform a grid search. Although there was a problem, we were in the presence of an imbalanced class, which we approached using oversampling. We tested all our selected models using **regular sampling** and **oversampling (SMOTE)** and chose the best 3 models for each target variable.

SMOTE, Synthetic Minority Oversampling Technique, it's an oversampling technique that generates new artificial rows for the minority class with the objective of balancing that feature. It works by selecting two points of the minority class and generating a new point of that class on the gap between the two original points, there is a random factor associated with this technique that is regarding the place where the new point will be generated. This became extremely relevant for our dataset since we had a huge class imbalance in targets, especially in the binary target. Therefore, models such as the random forest which are ensemble algorithms using bagging techniques, keep retrieving with replacement small fractions of the dataset to create the various decision trees that will be used as estimators. However, because we had this huge imbalance in our target, most of the time the estimators wouldn't even have both binary classes in the sample retrieved, resulting in an estimator that would always give the same prediction. Therefore, although this technique allowed the creation of some balance

in our dataset in some of our models, in other cases, the creation of these artificial rows ended up hurting the performance of others, such as the Logistic Regression.

Our goal was to achieve the highest possible *F1-Score* for our binary target variable and the highest possible *Weighted F1-Score* for our multiclass target variable, so we kept that in mind in our grid searches. The formula for the Weighted *F1-Score* is $\frac{\sum_i [w_i \times F_1 Score_i]}{\sum_i w_i}$, where w_i is the weight of each class and $F_1 Score_i$ is the *F1-Score* of each class. The weights are proportional to the inverse of the class frequencies, meaning that less frequent classes contribute more to the overall score. This helps prevent the dominant class from overwhelming the evaluation metric and provides a more balanced measure of a model's performance across all classes. Of course, it wouldn't be finished if we didn't explore outside of the syllabus and find a different model that performed efficiently.

4.1. MODELLING – BINARY TARGET

Concerning the binary target and following the approach that we have mentioned before, we started by testing with and without oversampling: **Logistic Regression**, **Gaussian Naïve Bayes**, **KNN**, **Decision Trees**, **Support Vector Machine (SVM), Ensemble Models** (**AdaBoost** and **Random Forest**) and **Neural Networks**. Taking an overview in [Figure 10: F1 Scores for Models without SMOTE (Binary Target)] and [Figure 11: F1 Scores for Models with SMOTE (Binary Target)] we noticed that in general, SMOTE technique allows us to get better results.

More specifically, looking into the two different perspectives the **models that stood out** were: **Logistic Regression without SMOTE**, **SVM with SMOTE**, and **Neural Network with SMOTE**. Subsequently, we conducted an in-depth grid search in those models, assessing hyperparameter impacts on model performance and ensuring robust findings through systematic parameter variation and cross-validation.

For **Logistic Regression**, our hyperparameter tuning focuses on addressing *penalty* that influenced sparsity, *C* that determined fitting strength, *class_weight* to deal with imbalances, and also, the *solver* which impacted convergence, finalizing with a fixed *random_state* ensured reproducibility.

Moving to **SVM**, the grid search highlighted that *C* determined boundary smoothness, the choice of *kernel* introduced trade-offs, and *gamma* shaped the boundary. Including *class_weight* to balance class biases, and a fixed *random_state* to ensured consistency.

For **Neural Networks**, the grid search emphasized the importance of the *solver*="adam" setting, which optimized performance for large datasets. Parameters like *max_iter* influenced convergence, early stopping prevented overfitting, and activation and *hidden_layer_sizes* tailored the architecture. The fixed *random_state* ensured reproducibility.

Our systematic methodology, integrating meticulous **grid searches** and **cross-validation**, provided insights into hyperparameter impact, oversampling effects, and model performance on imbalanced data. This approach not only identified optimal models but also shed light on parameter interactions within our specific problem domain.

After conducting grid searches for our top three models, it is evident from the results, depicted in [Figure 12: F1 Scores for the three best Models for Binary Target], that **Logistic Regression** emerged as the **best-performing model** for our binary target. The meticulous

exploration of hyperparameters showcased its superiority in terms of predictive accuracy and overall effectiveness compared to the other models in consideration.

When analyzing the **confusion matrix for Logistic Regression**, it revealed the model's proficiency in predicting True Positives comparing to True Negatives, given a high percentage of False positives and low of False negatives. To enhance model performance, addressing the substantial miss in detecting actual negatives and positives and reducing false positives would be critical. [Figure 13: Confusion Matrix for Logistic Regression (Binary Target)].

4.2. MODELLING – MULTICLASS TARGET

The models subjected to testing, both with and without oversampling, were the same as the previous case, however an initial observation indicated **that oversampling negatively impacted scores.** Consequently, grid searches were exclusively performed on models without SMOTE, revealing that **KNN**, **Decision Trees**, and **Random Forest** outperformed others [Figure 14: F1 Scores for Models without SMOTE (Multiclass Target) and Figure 15: F1 Scores for Models with SMOTE (Multiclass Target)].

In our exploration of **Random Forest**, our grid search focused on crucial parameters such as *n_estimators* to determine the number of trees in the forest, *criterion* to define the measure for quality during tree splitting (specifically set to "entropy"), *max_depth* to influence the maximum depth of the trees, *min_samples_split* to set the minimum number of samples required to split an internal node, *class_weight* to address class imbalance by assigning weights to different classes, *min_samples_leaf* to specify the minimum number of samples required to form a leaf node, and *max_features* to determine the subset of features considered for each split, influencing tree diversity. This systematic exploration allows us to understand the impact of each parameter and identify the optimal configuration for maximizing Random Forest effectiveness in our specific problem domain.

Turning to **KNN**, the grid search meticulously tuned key parameters such as *n_neighbors* (5, 7, 9), *weights* ("uniform", "distance"), and *algorithm* ("auto", "ball_tree", "kd_tree"). This systematic exploration facilitated the identification of optimal combinations, optimizing KNN's performance for our specific problem.

In the case of **Decision Trees**, the grid search delved into critical hyperparameters, varying *min_samples_split* (2, 10, 50, 200), *min_samples_leaf* (1, 10, 50), *criterion* ("gini'", "entropy"), and *max_features* (None, 0.5). The insights garnered from the grid search provided clarity on optimal configurations for Decision Trees, considering factors like node splitting criteria and feature importance.

Following the grid searches for our top three models in multiclass target, the results in [Figure 16: F1 Scores for the three best Models for Multiclass Target] indicated that **Decision Tree** emerged as the **best-performing model**. However, it's important to note that KNN demonstrated competitive performance and was a close contender. The nuanced choice between Decision Tree and KNN highlights their effectiveness in addressing our specific multiclass problem.

Upon examining the **confusion matrix** for the **Decision Trees** [Figure 17: Confusion Matrix for Decision Tree (Multiclass Target)], we observed that its ability to predict the first class is very strong but it is very weak to predict the second class. Regarding the third class, the model

wrongly predicted that many observations were from the first class probably due to the high number of observations from that class.

4.3. HISTOGRAM-BASED GRADIENT BOOSTING

Given the optimistic results of the research we found in terms of utilization of Bosting algorithms, for instance, Gradient Bosting, in these situations, we decided to search about more about it and found **Histogram-based Gradient Boosting**, which is an optimization of the traditional Gradient Boosting algorithm, where histograms are used to efficiently represent and manipulate the data during the training process. This is particularly useful when dealing with large datasets which is part of the reason why we chose to test this model. Another reason would be the ability of this model to treat missing values: "This estimator has native support for missing values (*NaNs*). During training, the tree grower learns at each split point whether samples with missing values should go to the left or right child, based on the potential gain. When predicting, samples with missing values are assigned to the left or right child consequently. If no missing values were encountered for a given feature during training, then samples with missing values are mapped to whichever child has the most samples."[2]. Considering this, we thought it would be interesting to apply this model on the train dataset in which missing values were not treated. We applied the same pre-processing steps as we did originally, with the exception of the missing values imputation, and we trained both versions to see which one was better. The outcomes were rather similar. The scores of the binary target were almost the same, as well as the multiclass, therefore as it would be time efficient to not input the missing values, we decided to choose the model without the imputation of missing values in both cases.

In the comparison of confusion matrices, for the binary target, HGB [Figure18: Confusion Matrix for HGB - Binary Target] the model seems to have the same limitation as observed in Logistic Regression although we can see a slightly improving, having lower rates of misclassifications. In the multiclass target, HGB [Figure 19: Confusion Matrix for HGB - Multiclass Target] showcased a balanced distribution of correct predictions across all classes, contrasting with the Decision Tree's matrix, where misclassifications were more concentrated.

Lastly, we further extended our evaluations by comparing HGB directly to our best model for the binary target and our best model for the multiclass target, as you can see in [Figure 20: F1 Scores of Logistic Regression and HGB (Binary Target)] and [Figure 21: F1 Scores of Decision Tree and HGB (Multiclass Target)] respectively. These comprehensive comparisons provided a nuanced understanding of HGB's performance relative to our top models in each target category.

4.4. DEPLOYMENT

In order to deploy our final model and make the predictions on the test dataset, we first had to deal with similar issues to those of the train dataset. We must also take into consideration that in the test we could take advantage of the fact that it possesses a temporal relationship regarding our train, meaning that we know for sure that the values in test were retrieved chronologically after the one from train.

[2] *https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html*

With this in mind, the first thing that we did was to verify if we had any patients that appeared in both datasets so that we could input values such as *race* and *age* from, due to the fact that we would know that the *age* had to be bigger than the *age* of the last visit of that patient and also that the *race* had to be the same on all encounters of that patient. However, we didn't find any case where this occurred, meaning that we would have to solve this incoherences by using the same ideology used in the train dataset, the only difference was that we trained the models for the imputation of the categorical values using the data from train, since it had more information.

Finally, we simply followed the extra same steps used in train, meaning that we had to apply the same encoders, create the new features and remove the ones that weren't selected. With all of this done, we simply made a prediction on our final dataset using the trained models from above.

5. CONCLUSION

Throughout the course of this project, we successfully integrated the knowledge acquired during the Machine Learning course and applied additional insights gained through self-directed study. Despite encountering challenges, the experience proved to be highly educational, fostering in-depth discussions and investigations into various topics.

One significant challenge we faced was the limitation in computing power. This became evident when we had to resort to limited grid searches instead of testing many different hyperparameters in our SVM, Neural Networks and Ensemble methods. The limitation was needed by the excessive time required for Grid Search, highlighting the trade-off between accuracy and runtime efficiency.

Another notable limitation in our project was our lack of expertise as medical professionals. This aspect became particularly evident during the feature selection process, where we acknowledge the possibility of making errors. There is a concern that we might have inadvertently excluded some crucial variables that are essential for accurate predictions, highlighting the importance of collaboration with domain experts to enhance the robustness of the model.

To enhance the effectiveness of future projects and models, a promising approach would be to implement stacking models. By employing ensemble learning techniques that combine multiple models, such as stacking, we can capitalize on the strengths of individual models and mitigate their weaknesses. Stacking involves training a meta-model that learns to weigh the predictions of various base models, optimizing for improved overall performance. This strategy not only fosters diversity in model architectures but also provides a means to capture complex relationships within the data. Moreover, doing a different encoding besides on-hot encoding to deal with categorical nominal variables that had a high cardinality, for instance target encoding will have a huge impact on feature selection and model assessment. Finaly, a focus on elevating the predictive capabilities of individual models should be prioritized. This can be achieved through continuous refinement and optimization, exploring techniques to enhance the precision of singular model predictions. Embracing stacking models, and different encoding techniques along with striving for heightened individual model predictions presents a comprehensive and promising avenue for advancing the accuracy and reliability of future predictive models.

6. REFERENCES

[1] Wang, S., & Zhu, X. (2022). Nationwide hospital admission data statistics and disease-specific 30-day readmission prediction. Health Informatics Science and Systems, 10(1), 25.

https://sharanramjee.github.io/files/projects/cs224w.pdf page 13

[2] Seo, S. (2006). A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets (Master's thesis). University of Pittsburgh, Pittsburgh, PA.

https://d-scholarship.pitt.edu/7948/1/Seo.pdf

[3] Peng, H.C.; Long, F. & Ding, C. (2005). "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." Journal of Machine Learning Research, 5, 1227-1251

7. ANNEXES

7.1 APPENDICES

[a] Z-score

Another method that can be used to detect outliers is the Z-Score, using the mean and standard deviation:

$\boldsymbol{Z_i = \frac{x_i - \bar{x}}{sd}}$, where $\boldsymbol{X_i \sim N(\mu, \sigma^2)}$, and sd is the standard deviation of data.

The basic idea of this method is that if X follows a normal distribution, $\boldsymbol{N(\mu, \sigma^2)}$, then Z follows a standard normal distribution, $\boldsymbol{N(0, 1)}$, and Z-scores that exceed 3 in absolute value are generally considered outliers. It presents a reasonable criterion for identification of the outlier when data follows the normal distribution.

[b] Mutual Information

Mutual information is a statistical measure that quantifies the dependence between two random variables. It's a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

The fundamental idea behind mutual information is to evaluate how much information about one variable (e.g., a feature) is gained by observing another variable (e.g., the target). It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

[c] Histogram-based Gradient Boosting

Histogram-based Gradient Boosting, a scikit-learn classification model, uniquely accommodates missing values without requiring imputation. Unlike traditional scikit-learn models, it seamlessly works with NaNs during training.

Its key deviation from standard Gradient Boosting lies in the handling of data. Instead of exhaustively searching for optimal split points, this model uses a histogram technique. It discretizes continuous features into bins, efficiently identifying the best splits.

The advantages are substantial: reduced memory usage and faster training, particularly valuable for extensive datasets. This contrasts with Gradient Boosting, which lacks the efficiency gains achieved through histogram-based techniques.

7.2 FIGURES

Figure 1: ***Numeric Variables' Histograms before Outlier Removal***

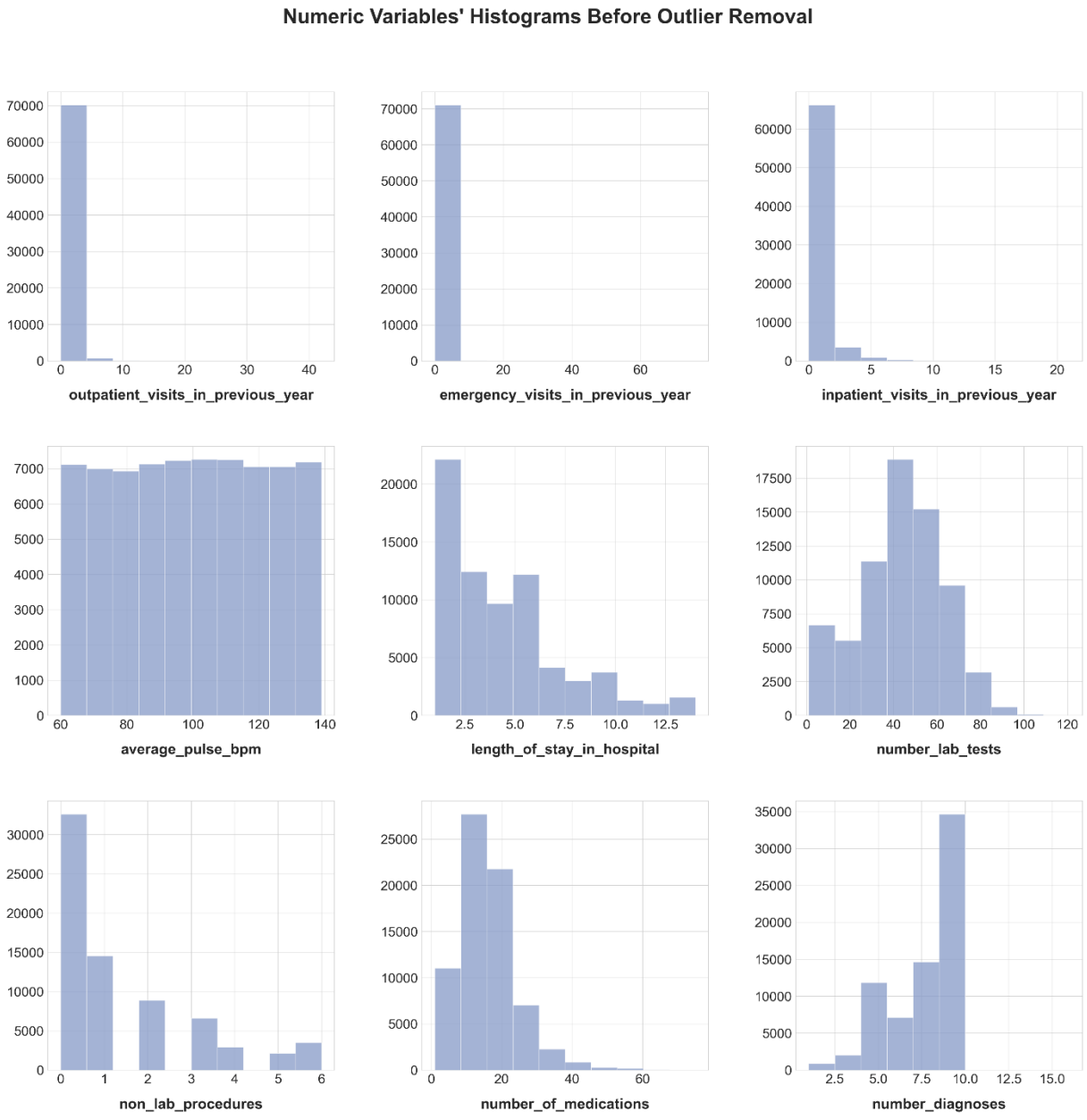


Figure 2: ***Numeric Variables' Box Plots Before Outlier Removal***

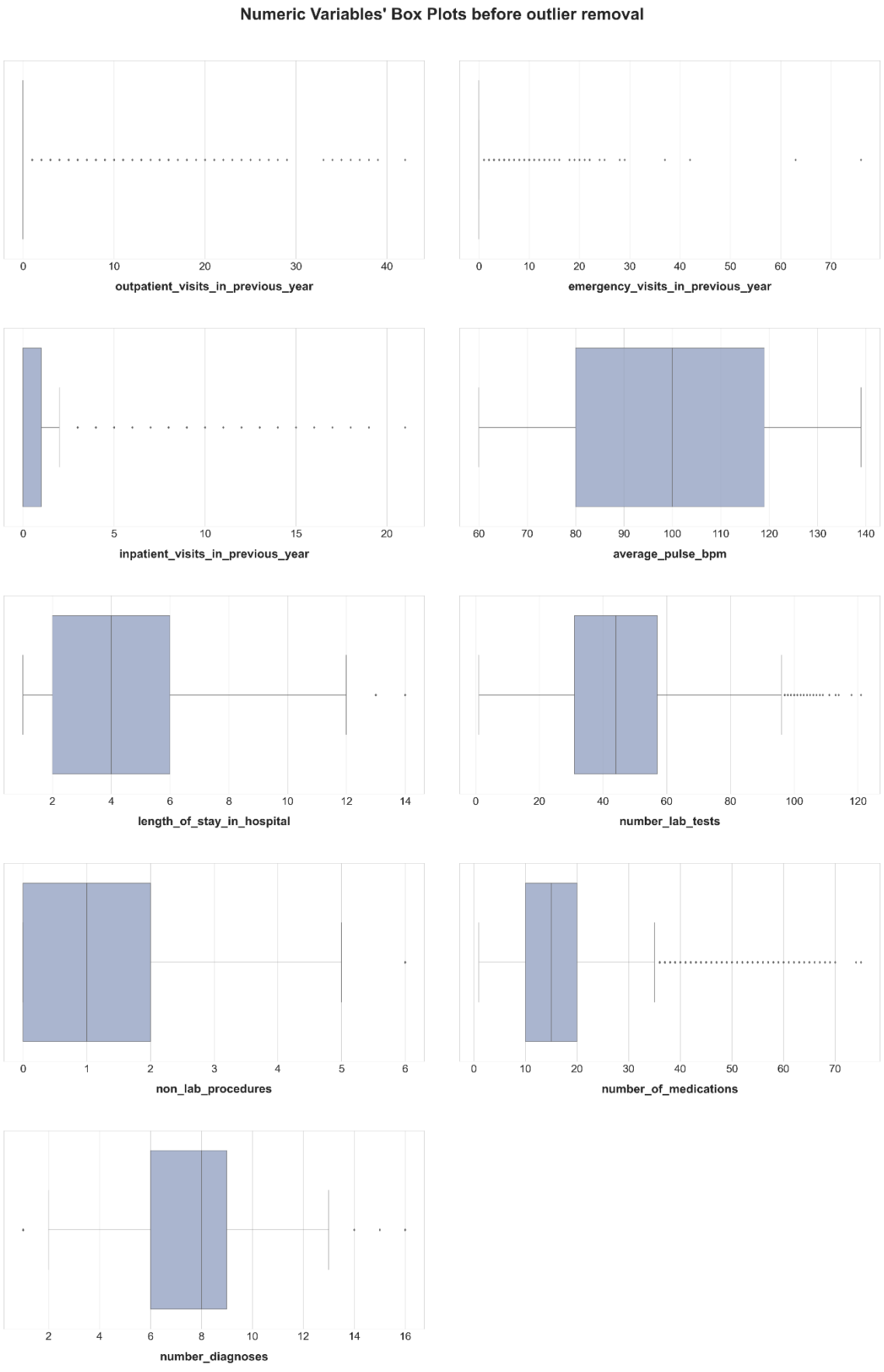


Figure 3: ***Correlation Matrix***

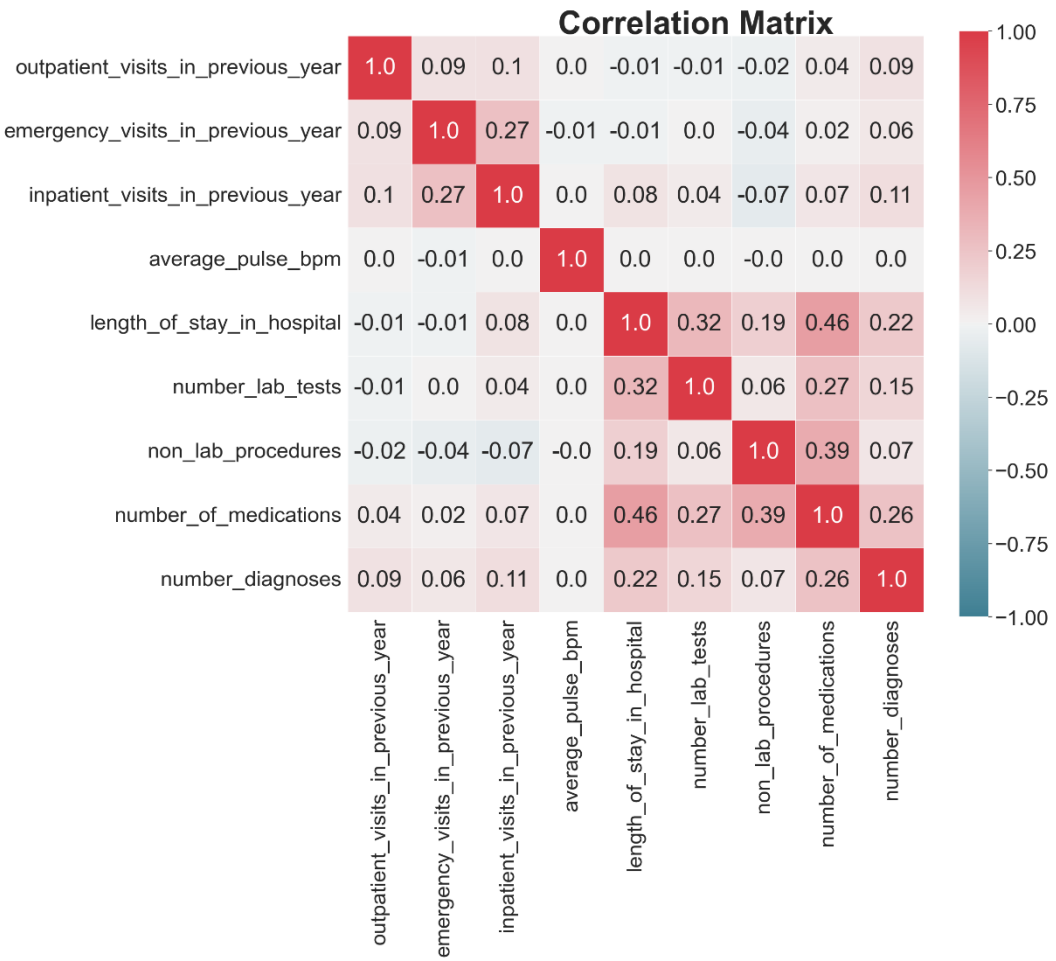
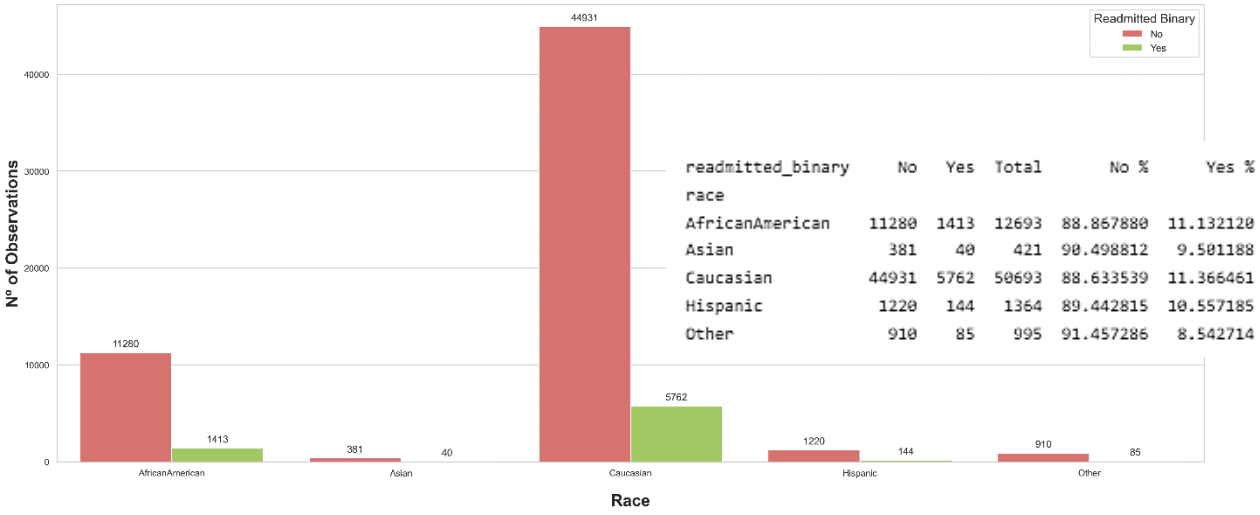
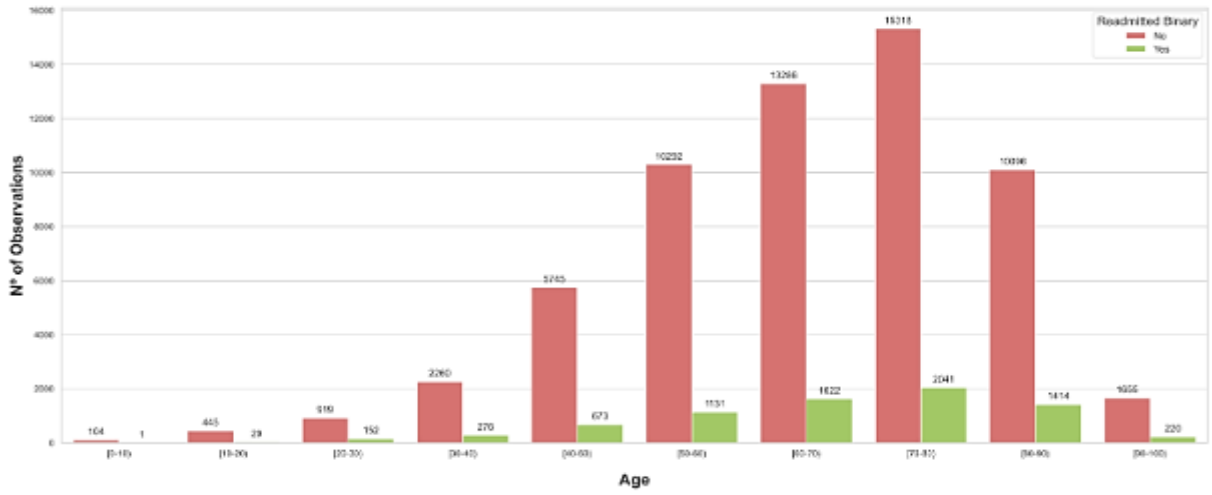


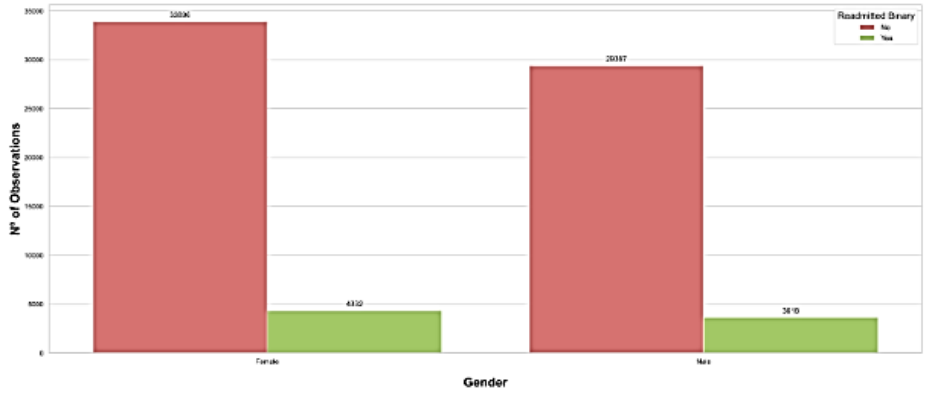
Figure 4: ***Influence of categorical features values in the status of the readmitted***



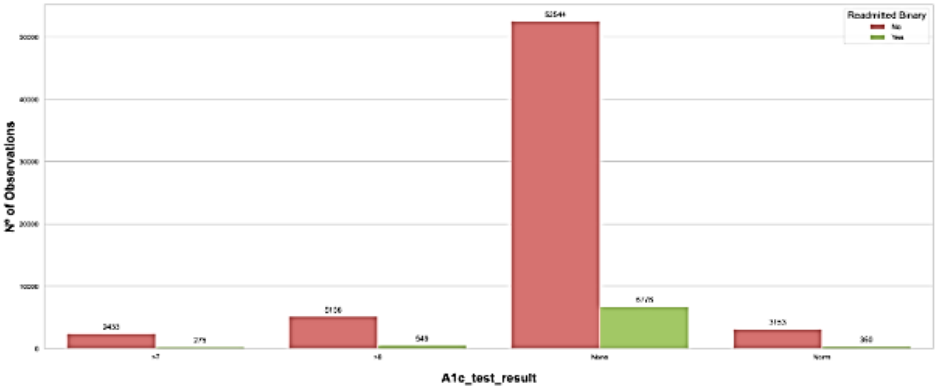
readmitted_binary	No	Yes	Total	No %	Yes %
race					
AfricanAmerican	11280	1413	12693	88.867880	11.132120
Asian	381	40	421	90.498812	9.501188
Caucasian	44931	5762	50693	88.633539	11.366461
Hispanic	1220	144	1364	89.442815	10.557185
Other	910	85	995	91.457286	8.542714



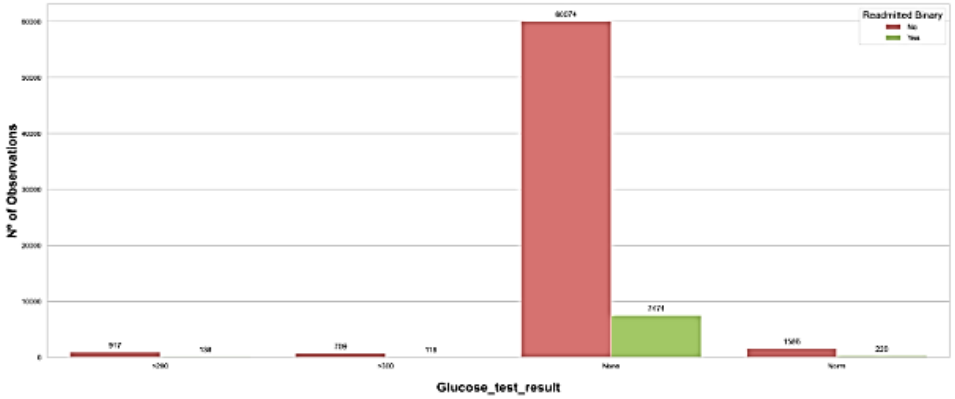
readmitted_binary	No	Yes	Total	No %	Yes %
age					
[0-10)	104	1	105	99.047619	0.952381
[10-20)	445	29	474	93.881857	6.118143
[20-30)	919	152	1071	85.807656	14.192344
[30-40)	2260	276	2536	89.116719	10.883281
[40-50)	5745	673	6418	89.513867	10.486133
[50-60)	10292	1131	11423	90.098923	9.901077
[60-70)	13286	1622	14908	89.119936	10.880064
[70-80)	15318	2041	17359	88.242410	11.757590
[80-90)	10096	1414	11510	87.715030	12.284970
[90-100)	1655	220	1875	88.266667	11.733333



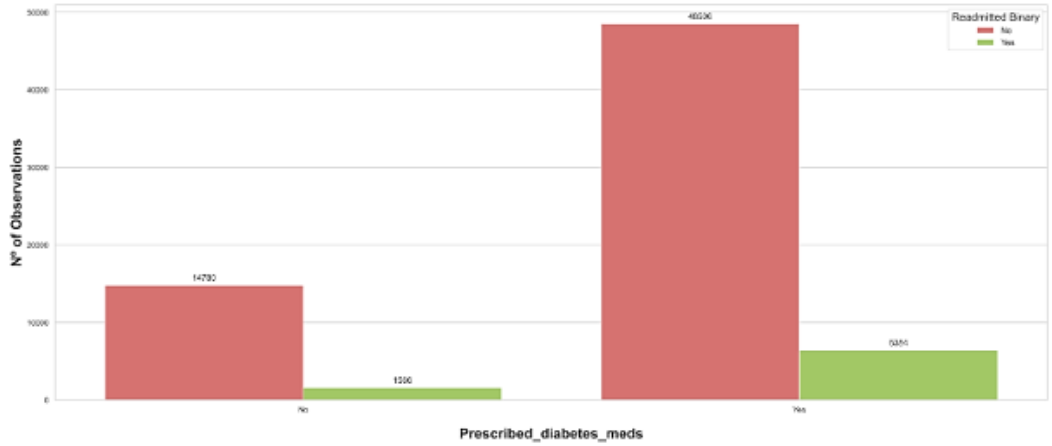
readmitted_binary	No	Yes	Total	No %	Yes %
gender					
Female	33896	4332	38228	88.667992	11.332008
Male	29387	3618	33005	89.038025	10.961975



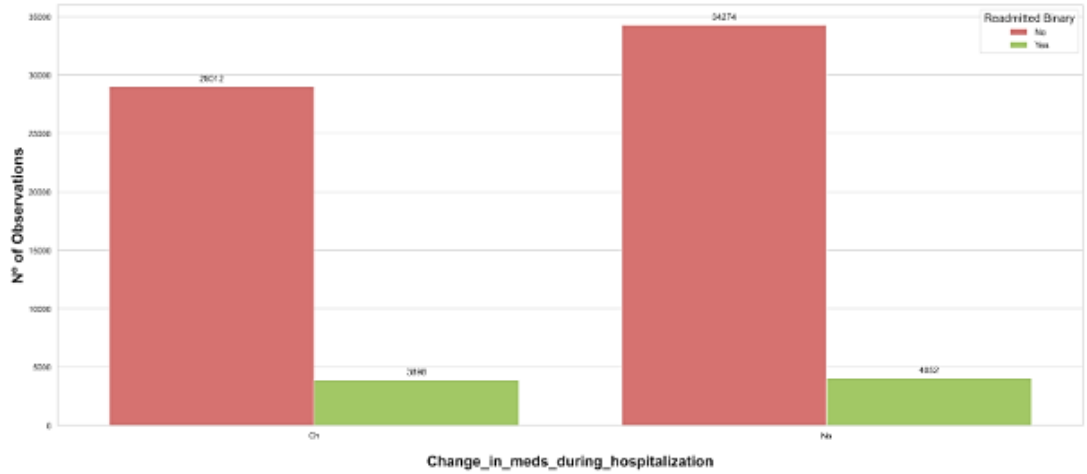
readmitted_binary	No	Yes	Total	No %	Yes %
a1c_test_result					
>7	2433	275	2708	89.844904	10.155096
>8	5156	549	5705	90.376862	9.623138
None	52544	6776	59320	88.577208	11.422792
Norm	3153	350	3503	90.008564	9.991436



readmitted_binary	No	Yes	Total	No %	Yes %
glucose_test_result					
>200	917	138	1055	86.919431	13.080569
>300	709	118	827	85.731560	14.268440
None	60074	7474	67548	88.935276	11.064724
Norm	1586	220	1806	87.818383	12.181617



readmitted_binary	No	Yes	Total	No %	Yes %
prescribed_diabetes_meds					
No	14780	1566	16346	90.419675	9.580325
Yes	48506	6384	54890	88.369466	11.630534



readmitted_binary	No	Yes	Total	No %	\
change_in_meds_during_hospitalization					
Ch	29012	3898	32910	88.155576	
No	34274	4052	38326	89.427543	

readmitted_binary	Yes %
change_in_meds_during_hospitalization	
Ch	11.844424
No	10.572457

Figure 5: ***Top 15 Medication Combinations***

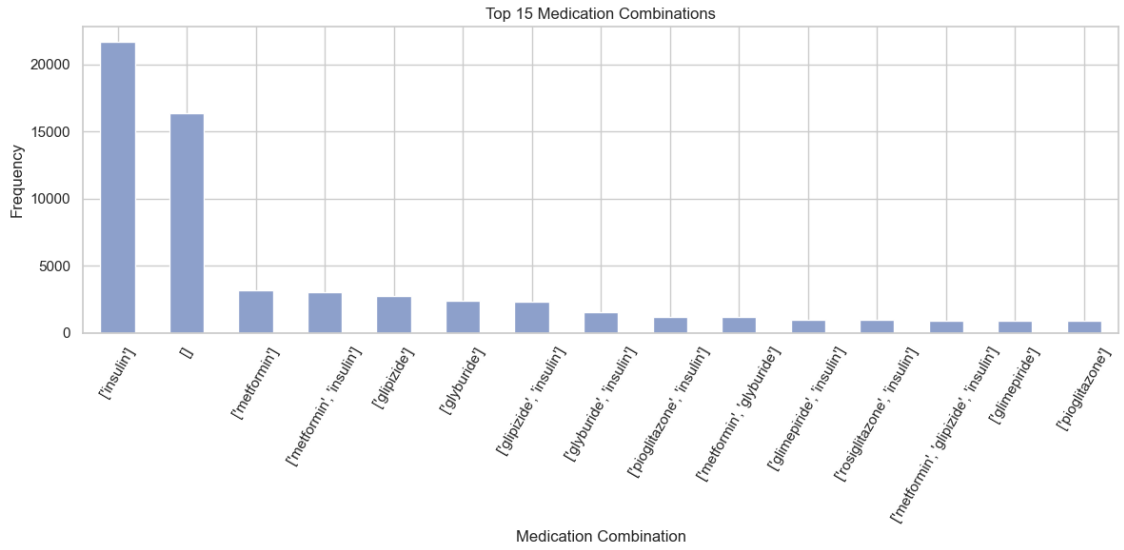


Figure 6: ***Numeric Variables' Box Plots After Outlier Removal***

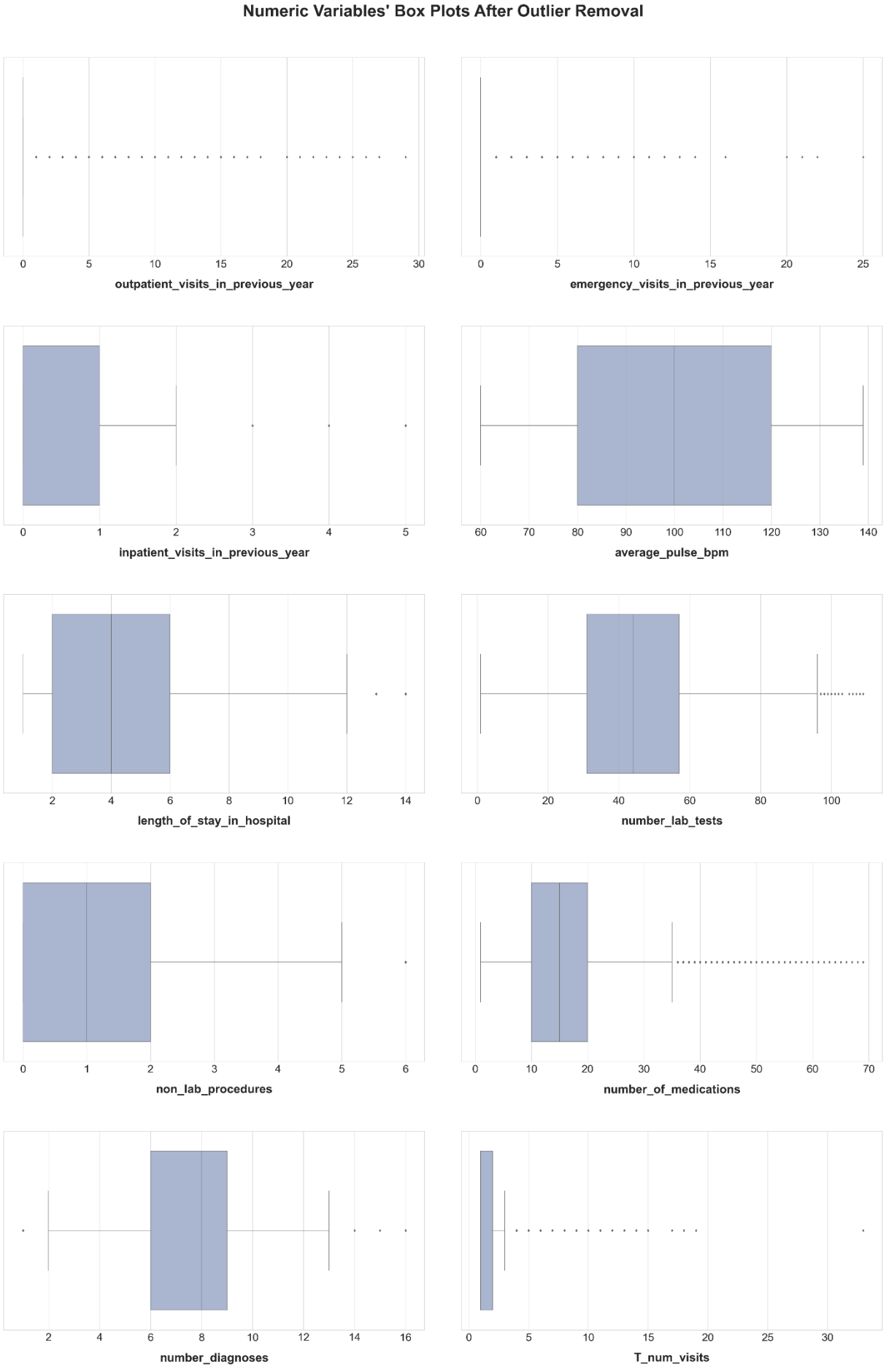


Figure 7: ***Spearmen's Correlation Matrix***

Figure 8: ***ANOVA Test Binary Target***

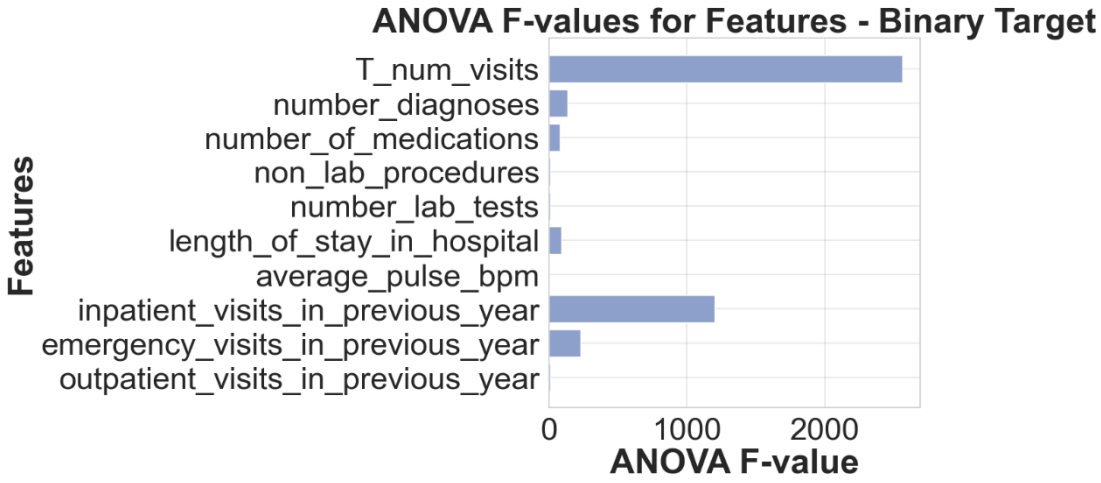


Figure 9: ***ANOVA Test Multiclass Target***

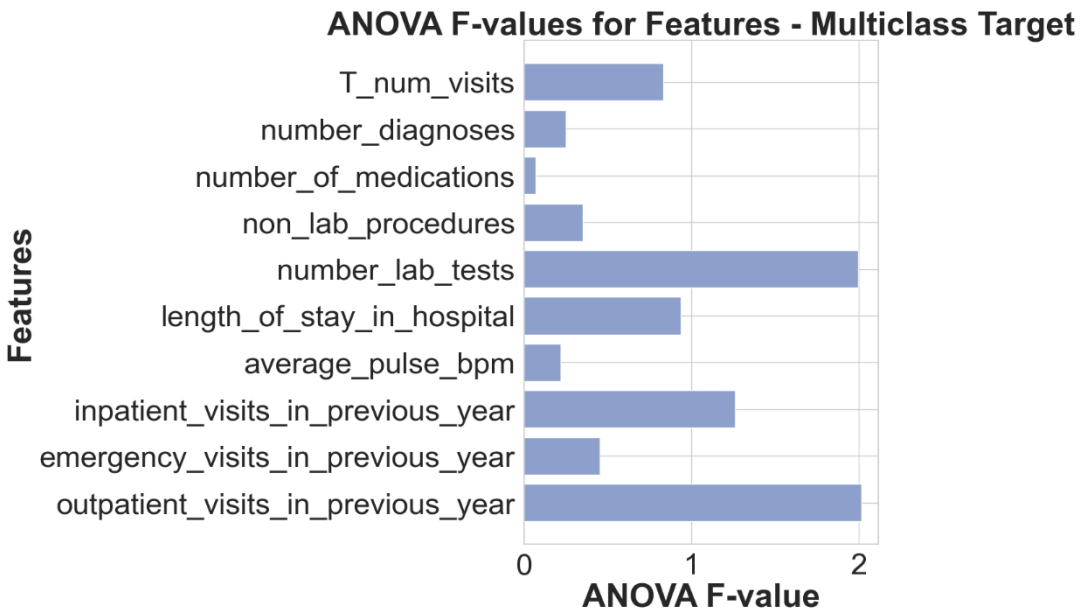


Figure 10: ***F1 Scores for Models without SMOTE (Binary Target)***

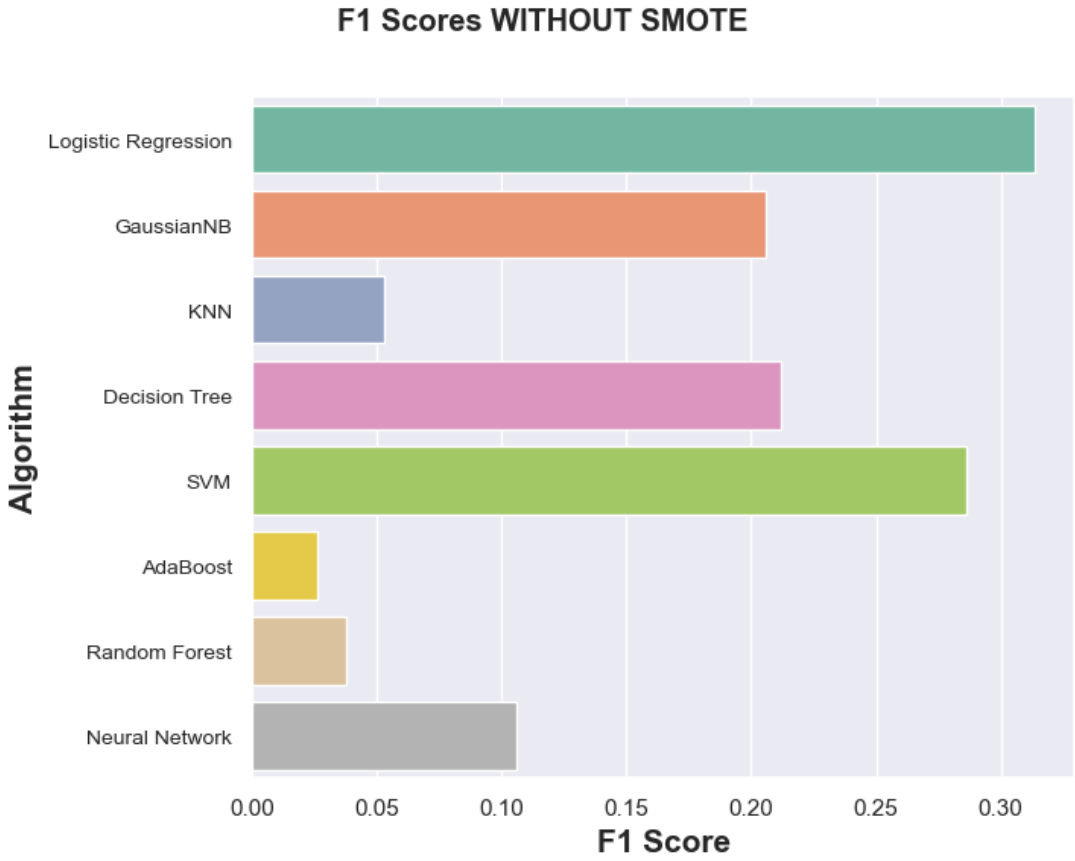


Figure 11: ***F1 Scores for Models with SMOTE (Binary Target)***

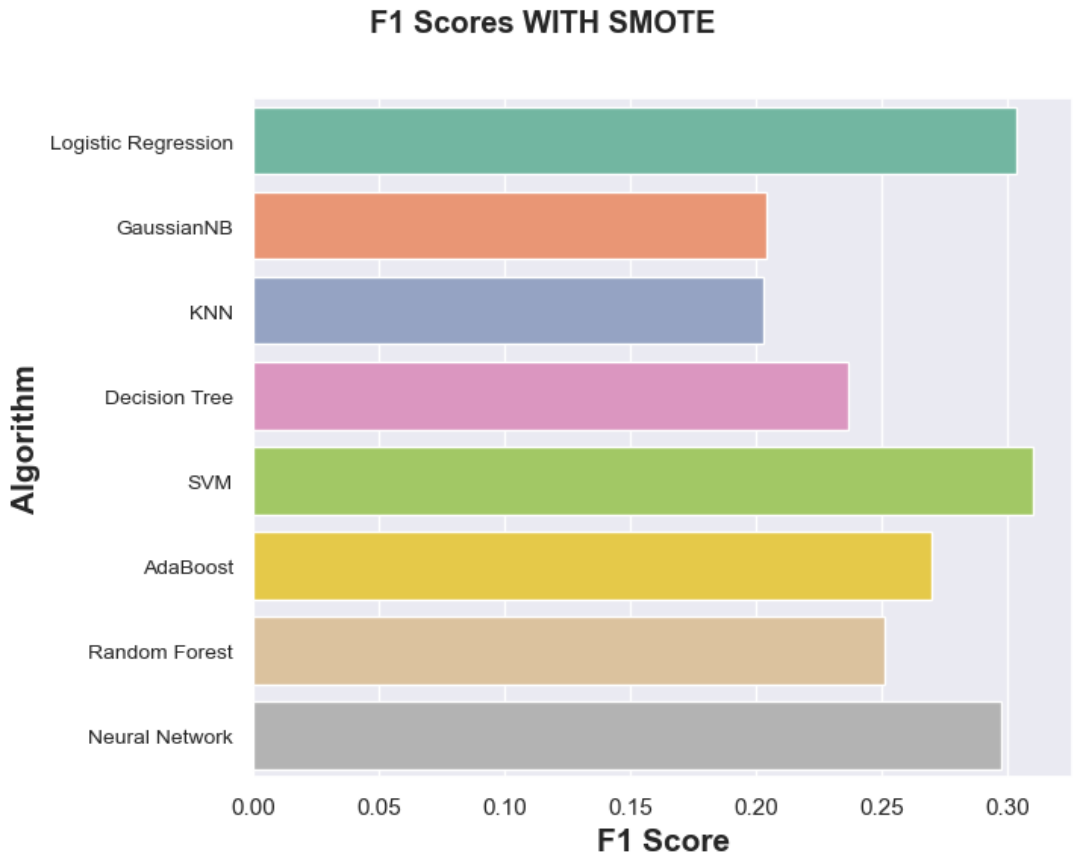


Figure 12: ***F1 Scores for the three best Models for Binary Target***

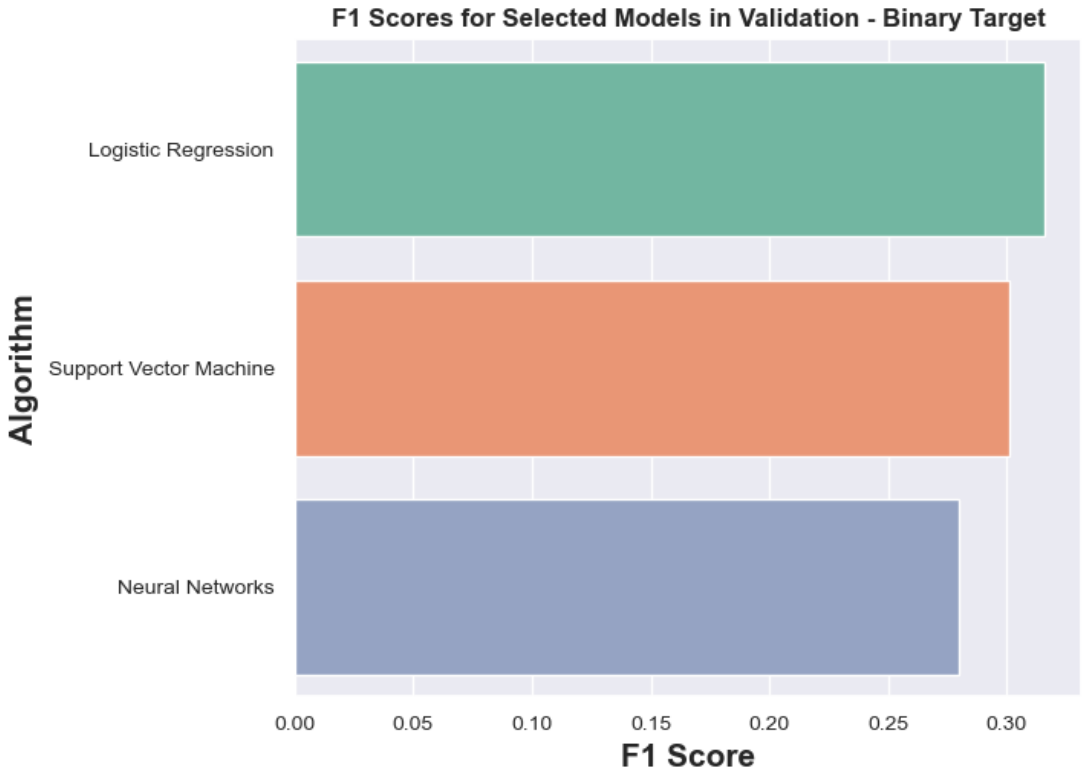


Figure 13: ***Confusion Matrix for Logistic Regression (Binary Target)***

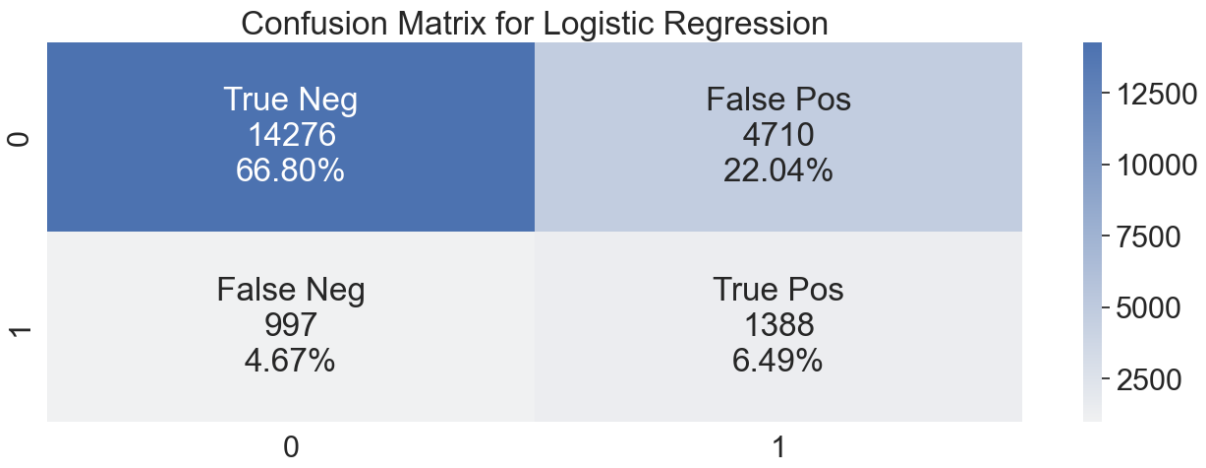


Figure 14: ***F1 Scores for Models without SMOTE (Multiclass Target)***

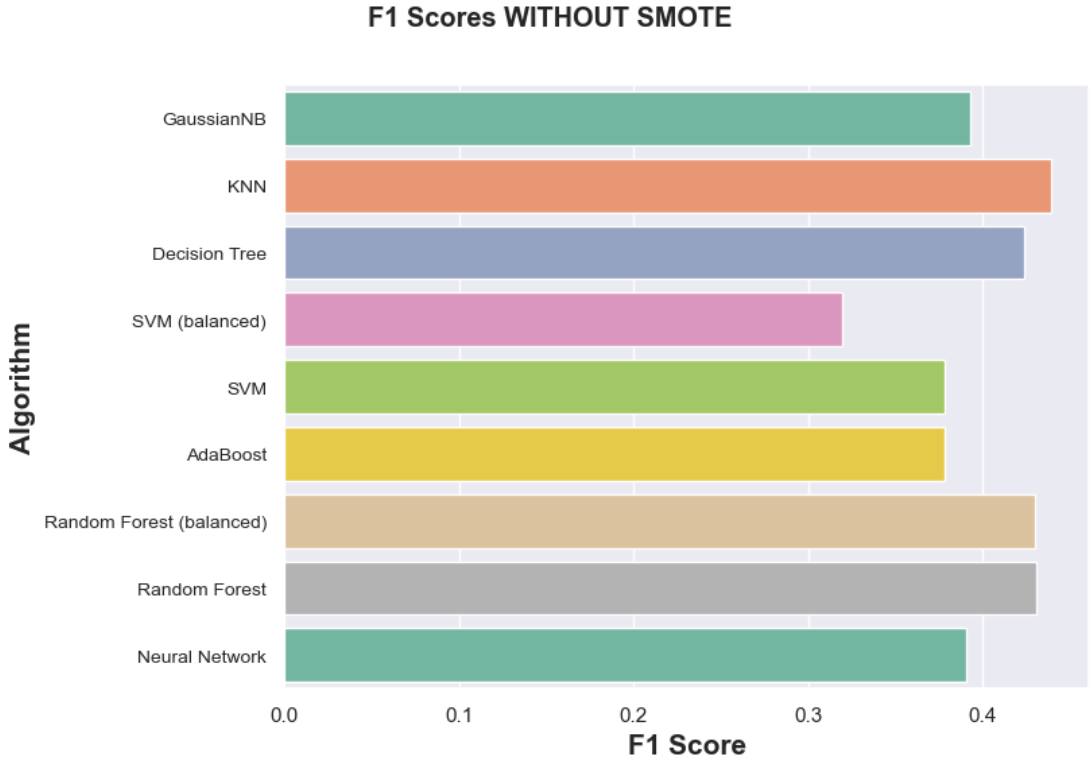


Figure 15: ***F1 Scores for Models with SMOTE (Multiclass Target)***

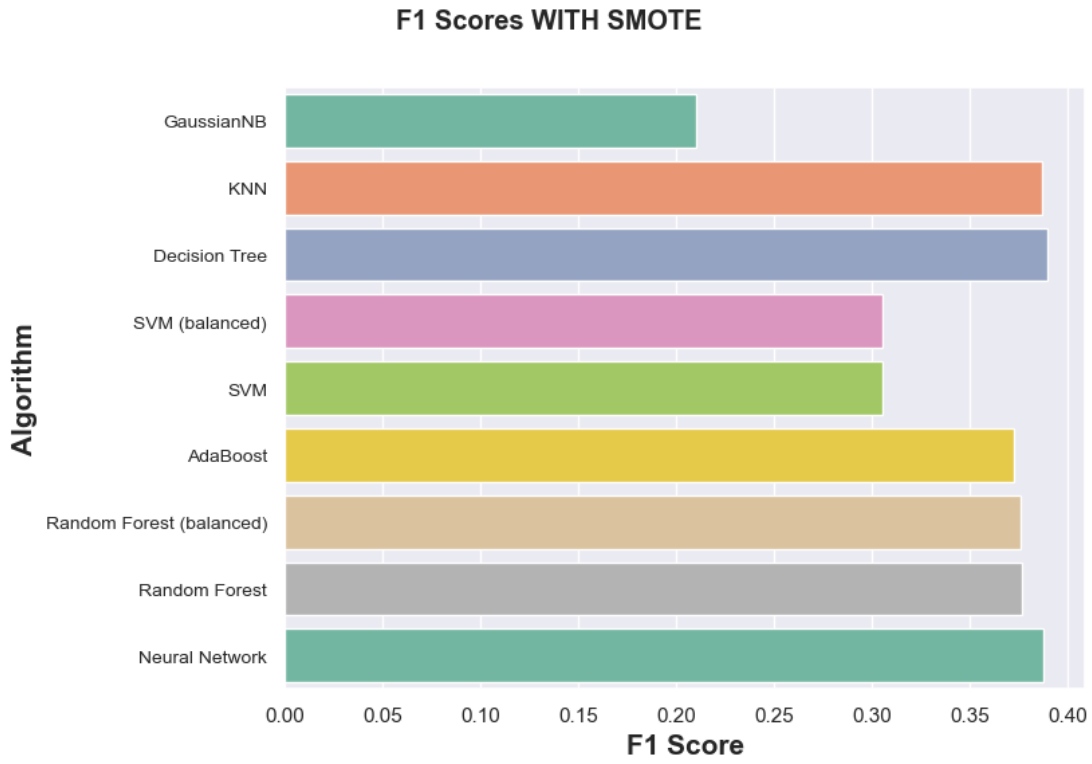


Figure 16: ***F1 Scores for the three best Models for Multiclass Target***

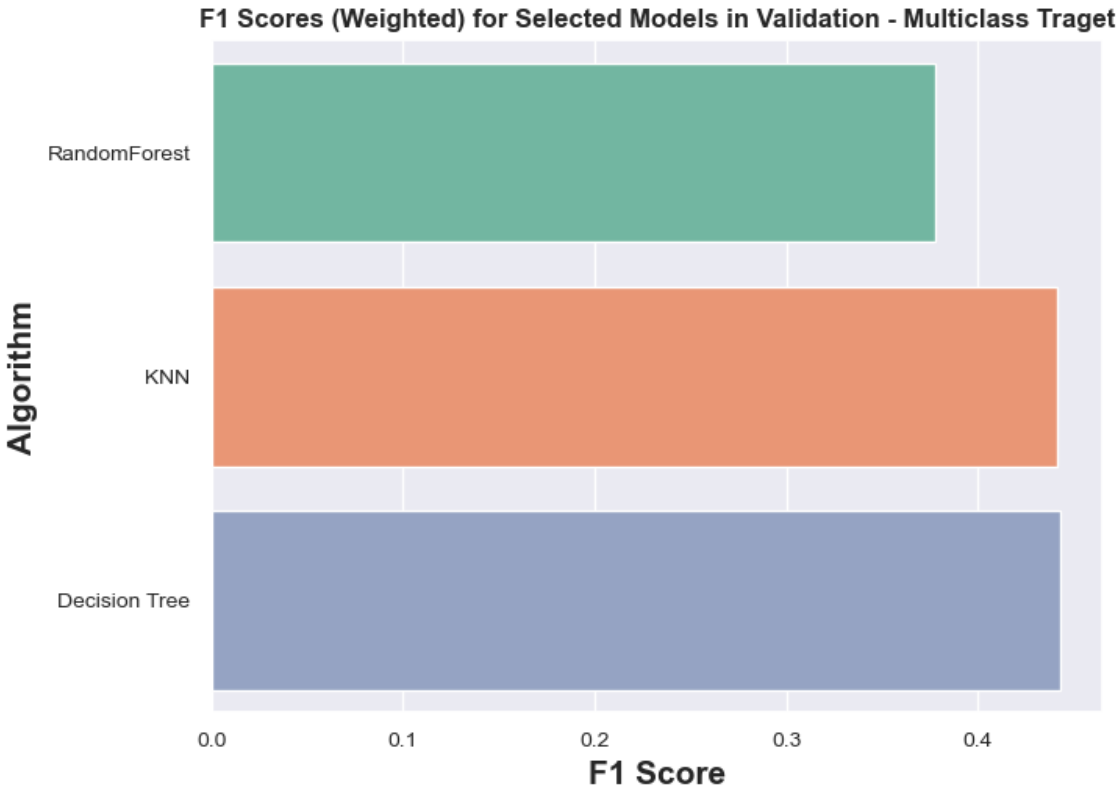


Figure 17: ***Confusion Matrix for Decision Tree (Multiclass Target)***

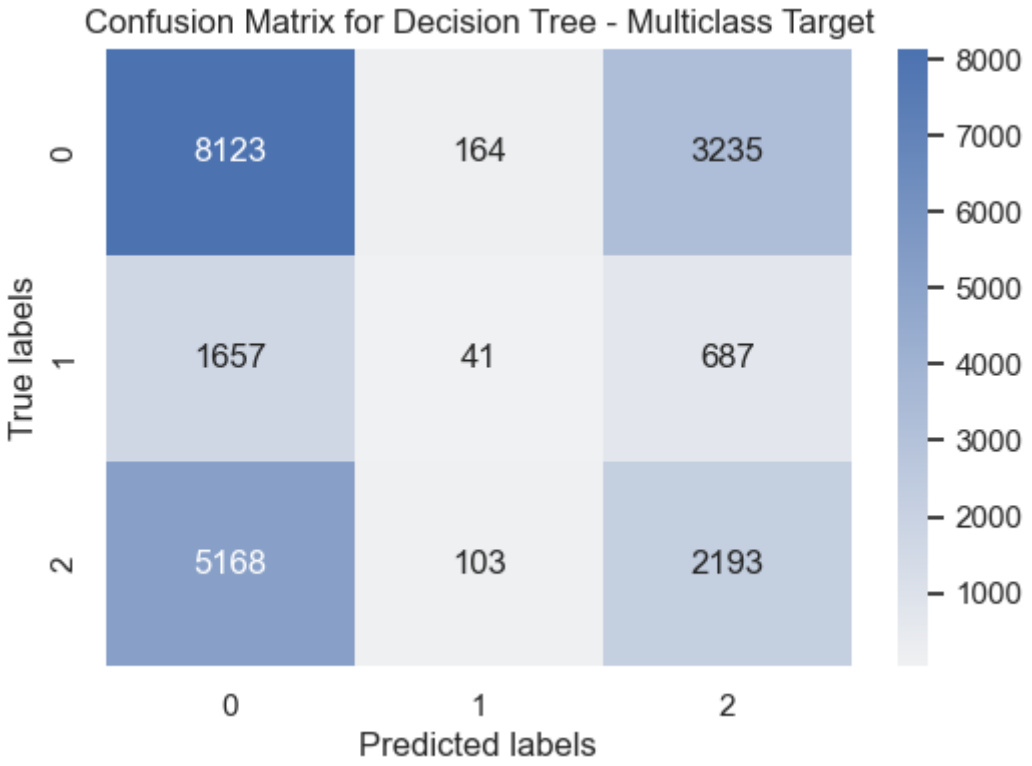


Figure 18: ***Confusion Matrix for HGB - Binary Target***

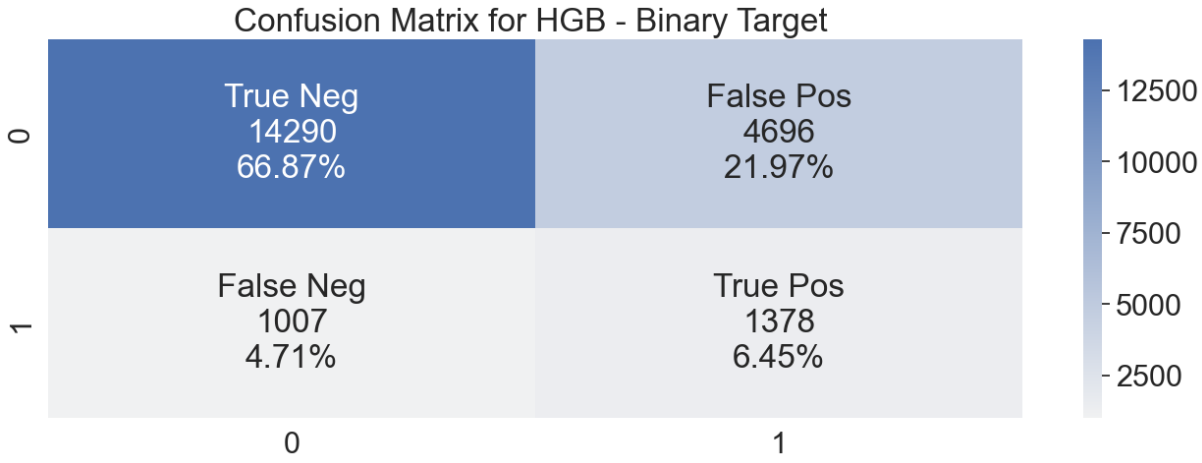


Figure 19: ***Confusion Matrix for HGB - Multiclass Target***

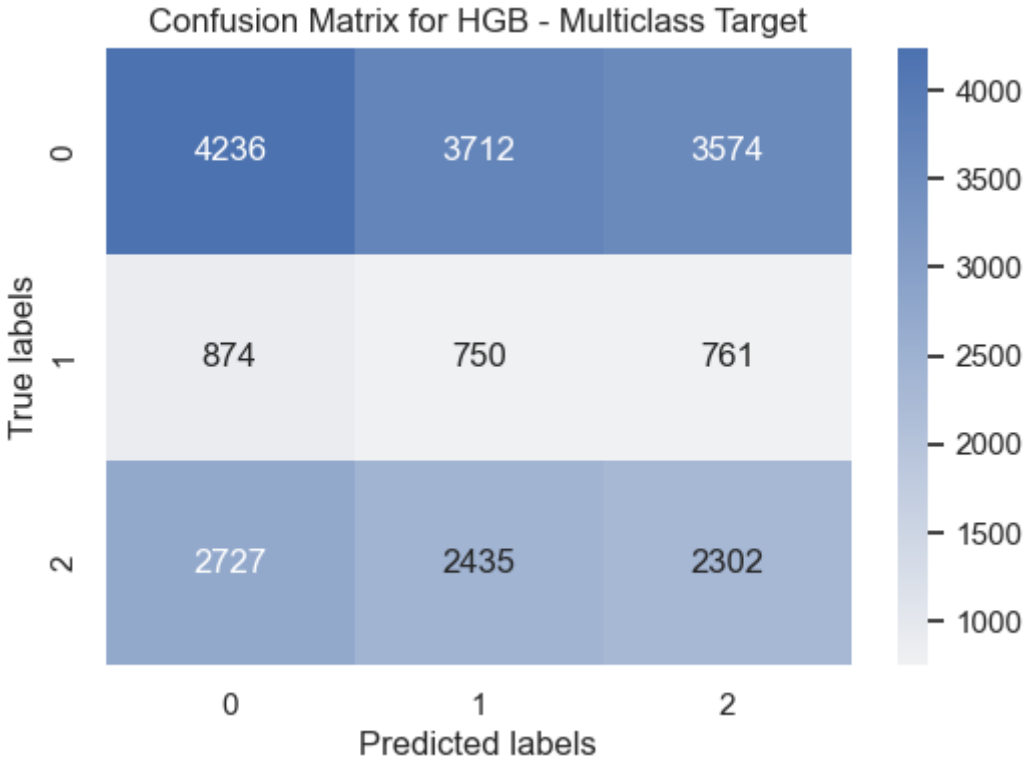


Figure 20: ***F1 Scores of Logistic Regression and HGB (Binary Target)***

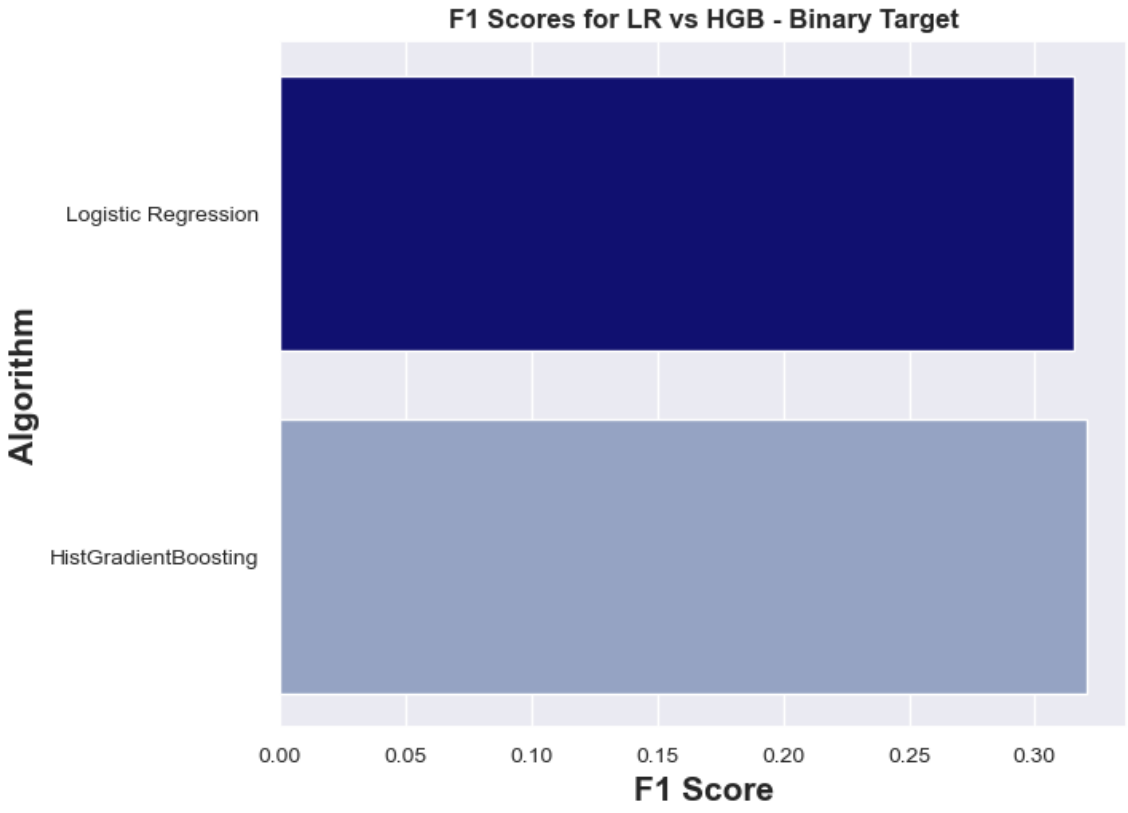
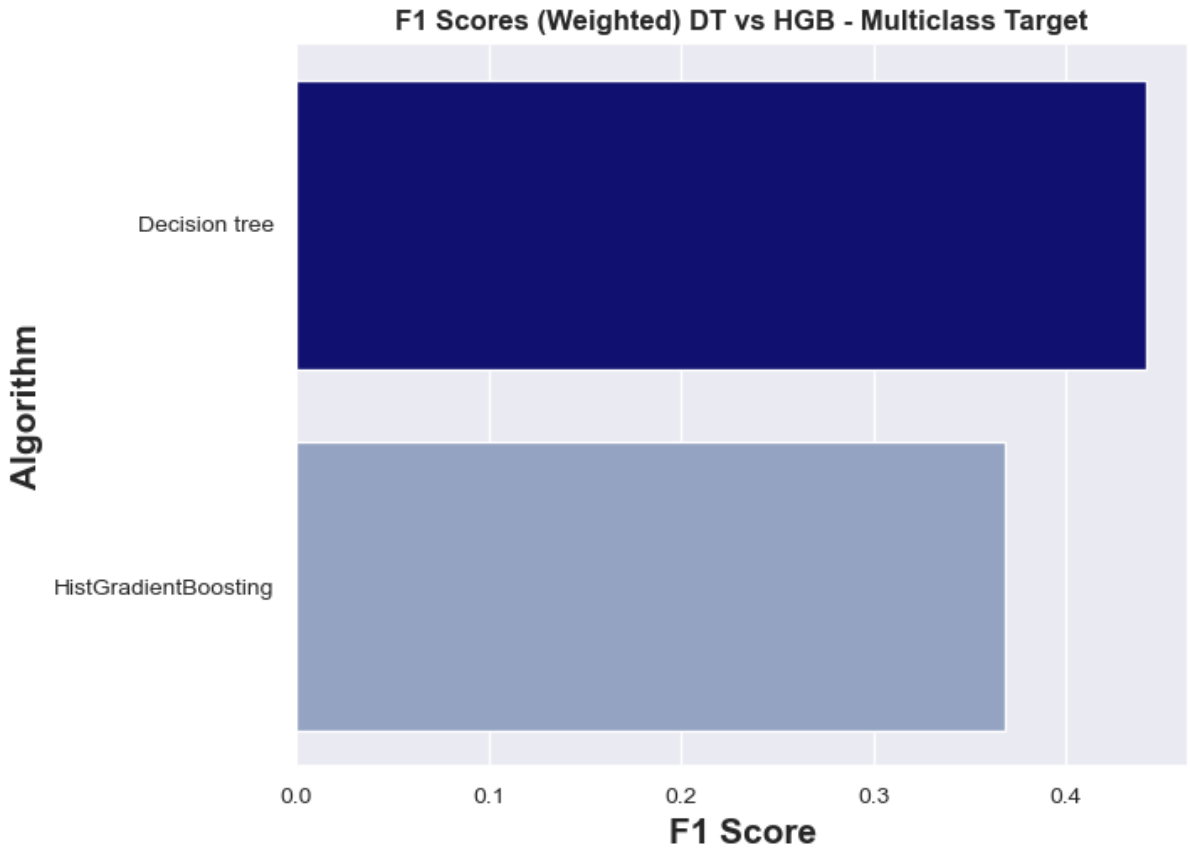


Figure 21: ***F1 Scores of Decision Tree and HGB (Multiclass Target)***



7.3 TABLES

Table 1: ***Definition of numerical and categorical features***

Metric (Continuous) Features	Non-Metric (Categorical or Discrete) Features
outpatient_visits_in_previous_year	*readmitted_multiclass*
emergency_visits_in_previous_year	*readmitted_binary*
inpatient_visits_in_previous_year	*weight*
average_pulse_bpm	*country*
length_of_stay_in_hospita	*age*
number_lab_tests	*medication*
non_lab_procedures	*prescribed_diabetes_meds*
number_of_medications	*change_in_meds_during_hospitalization*
number_diagnoses	*a1c_test_result*
	glucose_test_result
	additional_diagnosis
	secondary_diagnosis
	primary_diagnosis
	admission_source
	discharge_disposition
	medical_specialty
	admission_type
	payer_code
	gender
	race
	patient_id

Table 2: ***Missing Values Preprocessing***

	Missings values	% Missing values
weight	47678	96.800260
medical_specialty	24194	49.120884
admission_type	4912	9.972794
race	3466	7.036992
admission_source	3371	6.844114
age	2481	5.037154
discharge_disposition	1783	3.620011
Additional_Diagnosis	714	1.449628
Second_Diagnosis	187	0.379665
First_Diagnosis	12	0.024364
gender	2	0.004061

Table 3: ***Chi-Square Test Binary – Insignificant Variables***

	Variable	p_value
67	pioglitazone	1.000000
24	Second_Diagnosis_diseases of the circulatory system	0.997917
38	First_Diagnosis_diseases of the respiratory system	0.982661
1	admission_type_Urgent	0.820409
13	First_Diagnosis_diseases of the genitourinary system	0.798917
18	medical_specialty_Surgery-General	0.774739
45	discharge_disposition_Other reason	0.771070
8	glipizide	0.764321
22	Second_Diagnosis_diseases of the respiratory system	0.748540
12	admission_source_Clinic Referral	0.695916
7	race_Asian	0.659518
6	Additional_Diagnosis_symptoms, signs, and ill-defined conditions	0.648644
72	Second_Diagnosis_diseases of the digestive system	0.624657
56	race_Hispanic	0.476918
37	race_AfricanAmerican	0.393906
10	glyburide	0.370269
39	gender_Female	0.369692
29	gender_Male	0.369692
62	Second_Diagnosis_external causes of injury and supplemental classification	0.364788
57	Additional_Diagnosis_injury and poisoning	0.358099
55	Second_Diagnosis_diseases of the genitourinary system	0.314822
36	rosiglitazone	0.271103
11	admission_source_ Transfer from another health care facility	0.269834
53	Additional_Diagnosis_diseases of the digestive system	0.258961
31	medical_specialty_InternalMedicine	0.258221
77	medical_specialty_Orthopedics	0.257073
58	medical_specialty_Psychiatry	0.242441
60	Second_Diagnosis_injury and poisoning	0.201901
48	Second_Diagnosis_symptoms, signs, and ill-defined conditions	0.171229
49	Additional_Diagnosis_diseases of the circulatory system	0.140302
74	medical_specialty_Family/GeneralPractice	0.127164
42	admission_source_Transfer from a hospital	0.100439
73	First_Diagnosis_diseases of the circulatory system	0.088105
59	Additional_Diagnosis_other diseases	0.072185
19	glimepiride	0.064290
66	glucose_test_result	0.061878
20	race_Caucasian	0.060038
47	Second_Diagnosis_endocrine, nutritional and metabolic diseases, and immunity disorders	0.054813

Table 4: ***Chi-Square Test Multiclass – Insignificant Variables***

	Variable	p_value
30	admission_source_ Emergency Room	0.977047
10	glyburide	0.964875
67	pioglitazone	0.958755
17	Additional_Diagnosis_diseases of the genitourinary system	0.957644
45	discharge_disposition_Other reason	0.952345
23	First_Diagnosis_injury and poisoning	0.941140
31	medical_specialty_InternalMedicine	0.922572
15	Additional_Diagnosis_external causes of injury and supplemental classification	0.912896
42	admission_source_Transfer from a hospital	0.908125
62	Second_Diagnosis_external causes of injury and supplemental classification	0.849143
4	admission_type_Elective	0.841843
46	discharge_disposition_Expired	0.832808
76	insulin	0.829758
35	medical_specialty_Other	0.812619
27	race_Other	0.786309
3	Additional_Diagnosis_diseases of the respiratory system	0.774909
7	race_Asian	0.772896
21	admission_type_Emergency	0.764250
52	insurance	0.754344
43	discharge_disposition_Discharged/transferred to another type of inpatient care institution	0.751505
65	medical_specialty_Cardiology	0.734780
51	discharge_disposition_Discharged to home	0.726297
68	no medication prescribed	0.717788
26	prescribed_diabetes_meds_No	0.717788
54	prescribed_diabetes_meds_Yes	0.717788
18	medical_specialty_Surgery-General	0.685776
28	a1c_test_result	0.680568
77	medical_specialty_Orthopedics	0.676817
6	Additional_Diagnosis_symptoms, signs, and ill-defined conditions	0.668062
64	First_Diagnosis_endocrine, nutritional and metabolic diseases, and immunity disorders	0.660545
1	admission_type_Urgent	0.639187
74	medical_specialty_Family/GeneralPractice	0.634681
61	First_Diagnosis_symptoms, signs, and ill-defined conditions	0.621277
11	admission_source_ Transfer from another health care facility	0.611345
25	admission_source_ Physician Referral	0.601466
20	race_Caucasian	0.598880
36	rosiglitazone	0.565867
70	medical_specialty_Orthopedics-Reconstructive	0.528961
50	Second_Diagnosis_other diseases	0.521665
55	Second_Diagnosis_diseases of the genitourinary system	0.513287
34	medical_specialty_Emergency/Trauma	0.512990
49	Additional_Diagnosis_diseases of the circulatory system	0.512370
14	discharge_disposition_Discharged/transferred to SNF	0.500308
24	Second_Diagnosis_diseases of the circulatory system	0.470152
0	change_in_meds_during_hospitalization_No	0.469235
78	change_in_meds_during_hospitalization_Ch	0.469235
53	Additional_Diagnosis_diseases of the digestive system	0.420030
75	medical_specialty_Nephrology	0.399539
5	age	0.398937
69	Additional_Diagnosis_endocrine, nutritional and metabolic diseases, and immunity disorders	0.397398
12	admission_source_Clinic Referral	0.382877
8	glipizide	0.345988
13	First_Diagnosis_diseases of the genitourinary system	0.343587
59	Additional_Diagnosis_other diseases	0.330053
72	Second_Diagnosis_diseases of the digestive system	0.323674
57	Additional_Diagnosis_injury and poisoning	0.309652
73	First_Diagnosis_diseases of the circulatory system	0.299989
2	medical_specialty_Radiologist	0.298124
47	Second_Diagnosis_endocrine, nutritional and metabolic diseases, and immunity disorders	0.285836
58	medical_specialty_Psychiatry	0.222912
48	Second_Diagnosis_symptoms, signs, and ill-defined conditions	0.178766
9	discharge_disposition_Discharged/transferred to another short term hospital	0.163282
63	admission_source_Other	0.147468
60	Second_Diagnosis_injury and poisoning	0.144401
44	discharge_disposition_Discharged/transferred to home with home health service	0.136783
38	First_Diagnosis_diseases of the respiratory system	0.129167
37	race_AfricanAmerican	0.126003
66	glucose_test_result	0.116713
19	glimepiride	0.109557
16	First_Diagnosis_diseases of the digestive system	0.104590

Table 5: ***Combined Result for Binary Target***

	Feature	Logistics	Random Forest	Mutual Information	Feature Importance using Lasso Model	Total
1	number_diagnoses	True	True	True	True	4
2	emergency_visits_in_previous_year	True	True	True	True	4
3	T_num_visits	True	True	True	True	4
4	number_of_medications	False	True	True	True	3
5	length_of_stay_in_hospital	False	True	True	True	3
6	insurance	False	True	True	True	3
7	inpatient_visits_in_previous_year	False	True	True	True	3
8	discharge_disposition_Expired	True	False	True	True	3
9	discharge_disposition_Discharged/transferred to another type of inpatient care institution	True	False	True	True	3
10	discharge_disposition_Discharged/transferred to another short term hospital	True	False	True	True	3
11	discharge_disposition_Discharged/transferred to another rehab fac including rehab units of a hospital .	True	False	True	True	3
12	age	False	True	True	True	3
13	Additional_Diagnosis_endocrine, nutritional and metabolic diseases, and immunity disorders	False	True	True	True	3
14	prescribed_diabetes_meds_Yes	False	False	True	True	2
15	prescribed_diabetes_meds_No	False	False	True	True	2
16	no medication prescribed	False	False	True	True	2
17	medical_specialty_Other	False	False	True	True	2
18	medical_specialty_Nephrology	False	False	True	True	2
19	medical_specialty_Emergency/Trauma	False	False	True	True	2
20	medical_specialty_Cardiology	False	False	True	True	2
21	insulin	False	False	True	True	2
22	discharge_disposition_Discharged/transferred to home with home health service	False	False	True	True	2
23	discharge_disposition_Discharged/transferred to SNF	False	False	True	True	2
24	discharge_disposition_Discharged to home	False	False	True	True	2
25	change_in_meds_during_hospitalization_No	False	False	True	True	2
26	change_in_meds_during_hospitalization_Ch	False	False	True	True	2
27	admission_type_Emergency	False	False	True	True	2
28	admission_type_Elective	False	False	True	True	2
29	admission_source_Other	False	False	True	True	2
30	admission_source_ Physician Referral	False	False	True	True	2
31	admission_source_ Emergency Room	False	False	True	True	2
32	First_Diagnosis_external causes of injury and supplemental classification	False	False	True	True	2
33	First_Diagnosis_endocrine, nutritional and metabolic diseases, and immunity disorders	False	False	True	True	2
34	a1c_test_result	False	True	False	False	1
35	race_Other	False	False	False	False	0
36	metformin	False	False	False	False	0
37	medical_specialty_Radiologist	False	False	False	False	0
38	medical_specialty_Orthopedics Reconstructive	False	False	False	False	0
39	admission_type_Other	False	False	False	False	0
40	Second_Diagnosis_other diseases	False	False	False	False	0
41	First_Diagnosis_symptoms, signs, and ill-defined conditions	False	False	False	False	0
42	First_Diagnosis_other diseases	False	False	False	False	0
43	First_Diagnosis_injury and poisoning	False	False	False	False	0
44	First_Diagnosis_diseases of the digestive system	False	False	False	False	0
45	Additional_Diagnosis_external causes of injury and supplemental classification	False	False	False	False	0
46	Additional_Diagnosis_diseases of the respiratory system	False	False	False	False	0
47	Additional_Diagnosis_diseases of the genitourinary system	False	False	False	False	0

Table 6: ***Combined Result for Multiclass Target***

	Feature	Logistics	Random Forest	Mutual Information	Feature Importance using Lasso Model	Total
1	T_num_visits	True	True	True	True	4
2	number_lab_tests	False	True	True	True	3
3	number_diagnoses	False	True	True	True	3
4	non_lab_procedures	False	True	True	True	3
5	length_of_stay_in_hospital	False	True	True	True	3
6	admission_type_Other	True	False	True	True	3
7	metformin	False	False	True	True	2
8	gender_Male	False	False	True	True	2
9	gender_Female	False	False	True	True	2
10	Second_Diagnosis_diseases of the respiratory system	False	False	True	True	2
11	First_Diagnosis_external causes of injury and supplemental classification	False	False	True	True	2
12	race_Hispanic	True	False	False	False	1
13	outpatient_visits_in_previous_year	True	False	False	False	1
14	emergency_visits_in_previous_year	True	False	False	False	1
15	discharge_disposition_Discharged/transferred to another rehab fac including rehab units of a hospital .	True	False	False	False	1
16	average_pulse_bpm	False	True	False	False	1
17	inpatient_visits_in_previous_year	False	False	False	False	0
18	First_Diagnosis_other diseases	False	False	False	False	0