MASTER IN DATA SCIENCE AND ADVANCED ANALYTICS

Machine Learning

READY TO BE DISCHARGED: EXAMINING HOSPITAL READMISSIONS

Group Project

Machine Learning 2023/2024

Professors:

Roberto Henriques
Ricardo Santos
Rafael Pereira

Team Members - Group 46

Gonçalo Cardoso - 20230588
Pedro Barão - 20201614
Rodrigo Silva - 20230536
Burcu Yesilyurt - 20230763
Guilherme Sa - 20230520

Index

Abstract - Hospital readmissions pose a considerable challenge in the healthcare sector, serving as a vital measure of care quality and contributing to increasing costs. When patients require readmission shortly after discharge, it signifies potential deficiencies in care and exacerbates the financial strain on the healthcare system. Readmissions among individuals with diabetes are of particular concern as they have been found to have a significant impact on overall costs. Consequently, the ability to predict these instances of readmission holds the potential for enhancing patient care and realizing substantial cost savings. The main hypothesis is that medical factors, medications, age, gender, weight, admission type, and specific medical treatments received can be used to predict readmissions accurately. Previous studies have suggested that by leveraging these factors, it is possible to develop effective predictive models for identifying patients at high risk of readmission. To investigate this, a dataset comprising diabetic and non-diabetic patient records and readmission information will be analyzed. The dataset was treated and used to train and optimize predictive models. In the end, one model was found to be the best for both prediction scenarios stipulated and their results were analyzed, along with the limitations of the current implementation and further improvements.

Keywords: Machine Learning, Predictive Modelling, Classification, Diabetes, Hospital, Readmission

1 - Introduction

Diabetes is a chronic medical condition characterized by high blood sugar levels. It requires ongoing management and can lead to various complications if not properly controlled. Therefore, it is crucial to address the factors contributing to readmissions and implement strategies to reduce them, improving patient outcomes and hospital efficiency.

In this project, we aim to address the challenge of hospital readmissions, specifically focusing on diabetic patients. We expect to enhance patient care, mitigate the financial burden on the healthcare system, and improve overall healthcare outcomes for patients.

The first target variable to predict is *readmitted_binary* and can either be "Yes" if the patient was readmitted in less than 30 days or "No" otherwise. Our second objective is *readmitted_multiclass*, which predicts the timeframe of a patient's readmission. Readmitted multiclass can either be "No", "<30" or ">30".

2 - Data Exploration

We initiated the project with the data exploration phase, aiming to gain a comprehensive understanding of the datasets given. This would also help us to identify any issues or inconsistencies that required corrections to do in later stages. This exploration involved studying the variables, identifying missing values and outliers, and generating visualizations.

We received two datasets, one for training and another for testing, this last one without labels for the target variables. During the data exploration phase, only the training dataset was analyzed. This approach ensures that insights gained from the exploratory data analysis are derived solely from the data used for training and model optimization. By refraining from analyzing the test dataset, we prevent the possibility of unintentional data leakage from unseen data.

The data set contains general information about the patient (Gender, Age, Race), previous diagnoses, number of medications, number of previous visits, in other words, a screening. Based on the descriptive statistics (Table 1, 2, 3), we concluded that most of our variables are categorical, making it more challenging to apply some of the models. After analyzing the variables data types, it appears that there are many missing values predominantly within categorical features, namely *weight*, *glucose_test_result*, *a1c_test_result*, *medical_specialty*, and *payer_code* represent by themselves at least 40% of NaN in the entire dataset. Strategies to address and mitigate this will be discussed in subsequent sections.

In general, the majority of numerical features exhibit non-uniform distributions, suggesting that there is an uneven grouping of elements within these features. The correlations among pairs of these variables were not high with the highest in absolute value being 0.46 for *number_of_medications* and *length_of_stay_in_hospital* (Figure 1). Both linear (Pearson) and nonlinear (Spearman) correlations were verified. The Fisher-Pearson coefficient of skewness is a measure of the asymmetry of a distribution. A normally distributed variable would have a value of 0 for this coefficient, while positive values indicate more weight in the right tail for instance. After calculating the coefficient, 3 variables showed up as highly skewed: *inpatient_visits_in_previous_year*, *emergency_visits_in_previous_year*, and *outpatient_visits_in_previous_year* (Figure 2). Upon further inspection, this skewness is justified by a large presence of 0s for these variables (more than 65% for the first and more than 85% for the other two) and a huge drop in the count of observations with values larger than 0. These same 3 features also showed a few outliers . *non_lab_procedures* showed abundant 0 values (45%) as well (Figure 3).

Categorical features don’t show uniform distributions for the most part, however, some variables contain an even distribution, one example is *change_in_meds_during_hospitalization* (Figure 4). One variable, *country*, was found to be constant, showing that all patients present are from the USA. Some categorical features present a large number of categories, such as the three diagnosis variables and *medical_specialty*. Two binary features were found to be associated, *change_in_meds_during_hospitalization* and *prescribed_diabetes_meds*. No patient who had a change in medications was not prescribed diabetes meds. Furthermore, there is a substantial number of patients who had a change in medications and were also prescribed diabetes medicines (Table 4). In an effort to identify discriminative features for the target variables, we plotted each variable against *readmitted_binary*. However, none of the features exhibited a clear ability to distinguish readmitted patients.

The distribution of the binary and multiclass target features reveals potential challenges in classification (Figure 5). Notably, the prevalence of 'No' constitutes 88.83% of negative cases, while positive cases represented by 'Yes' account for only 11.17%. Given the relationship between the binary and multiclass target variables, this issue will be addressed in their respective sections.

3 - Preprocessing

3.1 Incoherences

The initial preprocessing step involved addressing data inconsistencies. We conducted a thorough check for incoherent observations. If the discrepancies were limited in number, we opted to remove them - ultimately, this resulted in the removal of all identified inconsistencies. Our first check for inconsistencies focused on instances where patients were admitted as newborns, yet their recorded age exceeded 10 years. Given that only five such observations were identified, they were excluded from the dataset. Additionally, cases involving sick babies and extramural births were eliminated when the age information was inconsistent with the admission source. Instances where the gender variable was recorded as 'Unknown/Invalid' were also dropped, with only three occurrences identified. Following these steps, less than 2% of the observations were removed.

3.2 Outliers

As mentioned in the data exploration, outliers were found in three variables: *outpatient_visits_in_previous_year*, *emergency_visits_in_previous_year,* and *inpatient_visits_in_previous_year*. These variables were also skewed due to a large presence of 0s (Figure 3). These findings were supported by checking their kurtosis measures (Figure 6). Kurtosis is a numerical measure of how heavy or light tailed a variable distribution is when compared to a normal one. Values larger than 3 show the presence of a heavy tail and, consequently, outliers.

It was also noticed that, for each of these features, not only the number of outliers was low (less than 25 for all 3 summed up), after a certain value, they all showed the same behavior regarding readmitted binary (Figure 7). What was done to treat these observations was winsorization, but instead of choosing a threshold for the cut, due to similar behavior with the target variable and their low presence in numbers, they were assigned for each respective variable the largest value before them that also had the same attribute in the dependent variable. For instance, in *outpatient_visits_in_previous_year*, any values exceeding 23 demonstrated a consistent pattern of being associated with a negative outcome for the *readmitted_binary* variable. Given this distinctive behavior, and in alignment with the winsorization strategy previously described, all values surpassing

23 (recognized as outliers) were systematically replaced with the value 24. This strategic adjustment aimed not only to mitigate the impact of outliers but also to align with the observed trend, reinforcing the cohesiveness of the data within the context of the target variable.

3.3 Missing Values

Several variables in the dataset exhibited missing values (Figure 8), and the approach to address them varied based on each variable's characteristics. Given the generally low correlations and associations between variables, we opted not to use KNN imputation. Weight, with a substantial 96.8% of missing values, was not dropped at this stage but will be addressed in the next section. Since the missing values were predominantly associated with categorical features, most of them were filled using the mode of their respective variables. However, exceptions were made for certain variables. For instance, *payer_code*, representing the code of the insurance used, was filled considering that missing values could signify patients without insurance. Following a research on insurance codes (Reference 1), it was discovered that 'SP' denoted 'self-paid,' so missing values were appropriately filled as such. The missing values in the glucose and a1c test result variables were interpreted as indicating that no test was conducted. For variables with a low percentage of missing values (< 10%), imputation was done using the mode of each variable. This strategy applied to the three diagnosis variables, as well as *race* and *age*. However, for variables with a higher percentage of missing values, a different approach was adopted. Missing values were grouped into a new category, typically labeled as 'Not Available' or 'Other,' allowing us to preserve these variables and the information associated with these instances.

3.4 Feature Engineering

At this phase of the project, our primary objective is to generate a comprehensive set of meaningful features for predicting individuals' return to the health facility (Table 5).

To correct the skewness of the previously identified features, *outpatient_visits_in_previous_year*, *emergency_visits_in_previous_year*, and *inpatient_visits_in_previous_year*, we applied the log(1+x) transformation. Since their distribution is right-skewed and there is a large presence of 0s, the ‘1+’ factor was important to not mistakenly delete observations. We also created a new feature, *visits_in_previous_year*, which represents the total number of visits made in the previous year.

We processed the medication variable, initially a list of medications per patient, by creating individual binary features for each available medicine. These binary features indicate whether a patient took a specific medicine or not. After thorough research on a1c and glucose tests (Reference 6, 7), we identified that patients with a1c values exceeding 7 and glucose levels surpassing 200 were indicative of diabetes. Consequently, we introduced a new binary feature named *test_result*, where 1 denotes a positive test result for diabetes and 0 otherwise. Due to a substantial number of missing values in the weight variable, we generated a binary feature - *weight_bin* - where 1 signifies the availability of weight information for a person, and 0 indicates the absence of this information. Using *patient_id*, a new feature termed *regular_patient* was produced, where 1 means that the patient appeared in the dataset more than once and 0 otherwise. At this stage, variables that could no longer be used or did not provide any value were dropped, namely *country*, *patient_id*, *weight*, *medication*, *glucose_test_result*, *a1c_test_result*, and the original visits in the previous year variables.

3.5 Encoding and Grouping

Algorithms designed to receive numerical inputs will be used to predict our targets, so it's important to group and encode the categorical features into a format that these algorithms can process (Table 6, 7, 8, 9, 10, 11, 12).

Starting with the binary features, such as *gender*, where the only values after removing the inconsistencies are ‘Male’ and ‘Female’; *change_in_meds_during_hospitalization*, where the strings ‘Ch’ and ‘No’ are implicit; and *prescribed_diabetes_meds*, in which the phenomenon is the same but the strings are ‘True’ and ‘False’, these categories transformed to 1s and 0s accordingly, like the remaining binary features, created previously. As *age* is a feature that has a range of values between 0 and 100, we encoded it based on the average of each class.

For the other features, we replaced the strings with numerical values such as 0, 1, 2, 3, etc. This method was chosen over one-hot encoding to avoid the creation of an excessive number of new features, which would then require additional feature selection. An alternative approach involving one-hot encoding was explored, with categories with less than 5% of observations in the dataset grouped together. However, this strategy was discarded due to its negative impact on model performance. Additionally, we experimented with frequency encoding, but this approach was also abandoned as it yielded suboptimal scores in our evaluations. We would like to highlight the treatment of the *payer_code* feature, where we applied a threshold of 1000 observations to consolidate classes in the minority. The methodology employed for grouping the remaining medical features was informed by research on the subject (Reference 3). For instance, the three diagnosis features encompass codes representing the type of diagnosis performed, allowing us to group them using the ICD9 (International Classification of Diseases). *medical_specialty* involved the consolidation of related categories, such as grouping all pediatric specialties into one category and grouping all obstetrics and gynecology-related specialties into another. Similarly, *discharged_disposition* had its categories organized, including groupings like 'Discharged to home,' 'Transferred to another medical facility,' and 'Hospice,' among others. The approach to grouping categories in 'admission_source' followed a similar approach.

3.6 Scaling and UnderSampler

Some machine learning algorithms exhibit sensitivity to the scale of input features. It helps make sure everything is consistent, which improves how quickly the model learns and performs better during optimization. This, in turn, contributes to the development of more efficient models.

In our exploration of various scaler options(Reference 5), we decided to use the MinMaxScaler. It is worth noting that while MinMaxScaler is not inherently robust to outliers, this characteristic did not significantly impact our results. Outliers were addressed in advance, minimizing their influence on the overall scaling process and subsequent model performance. Distributions were also accounted for, as this method preserves the original distribution shape, with all skewed variables having been fixed previously.

As previously mentioned, there is a heavy class imbalance in the target variables To address this issue, under-sampling was applied to balance the categories (Figure 5). Further details on this process and its results can be found in section 10.1 of the annex.

For model evaluation in later stages, KFold Cross Validation was predominantly used. The only exception was RFE for feature selection, discussed in section 4.3, which required the hold-out method. Accordingly, a 20% split was created here to form a validation set for this purpose.

4 - Feature Selection

For feature selection, various methods were explored. PCA was tested but did not provide satisfactory results; a detailed explanation and results overview can be found in section 10.2 of the annex. Additionally, a combination of filter, embedded, and wrapper methods was employed to identify the set of features with the best predictive ability. Since each approach yielded different results, we synthesized the conclusions from all methods. The final set of selected features comprised those chosen by at least two of the methods.

4.1 Filter Methods

Filter methods(Reference 16) select features independently of the classification process, essentially assessing the input features' dependence or independence from the target variable using statistical techniques to evaluate their relationship. Initially, we checked for univariate numerical variables by examining whether any had zero variance. This ended up removing 5 medication variables (Table 13). To check the association of categorical features with the target, we applied the Chi-Square test(Reference 17). This way, a set of categorical features highly dependent with the response variable was created (Table 14).

For numerical variables, we used mutual information to initially identify the most important characteristics based on their association with the target variable (Table 15). To not be too selective in a single method, we decided to only discard one of the numerical features, having this method select the best 8 among the 10 numerical ones. We used Spearman Correlation (Figure 9) to assess multicollinearity between metric variables. *visits_in_previous_year* was correlated with the three visits variables (Spearman correlation showed 0.55, 0.63, and 0.73). As one of the values was higher than 0.7 and it was also correlated with other two variables, *visits_in_previous_year* was dropped here.

4.2 Embedded Methods - LASSO and DecisionTree

LASSO and Decision Tree(Reference 16) are techniques that integrate feature selection into the model training process. Unlike Filter and Wrapper Methods, which are separate stages of model training, integrated methods incorporate feature selection directly, aiming to choose the most relevant features for the model's performance. Lasso regularization, as an embedded method, utilizes the L1 norm of the coefficients as a penalty term in the optimization objective. This penalty term reduces the coefficients of the least important features to zero, effectively removing them from the model. Features with a coefficient equal to 0 were, in this case, eliminated. It's important to note that Lasso regression incorporated cross-validation to find the optimal regularization parameter, aiding in selecting the optimal strength of the penalty term(Table 16 and Figure 10). A decision tree used as a classifier evaluates feature importance during the model training. Feature importances represent the percentage contribution of each variable to improving data separation. In this context, variables with an importance of 0 haven't been considered in creating any nodes throughout the construction of the decision tree. Two evaluation methods were tried for the decision tree specifically: 'gini' and 'entropy' (Figure 11) They differ in how they calculate node splits, with 'gini' focusing on the probability of misclassification and 'entropy' on the concept of information gain. The important features were selected by retaining only those with an importance greater than their expected contribution, calculated as the average percentage expected from each feature. For instance, with 47 features, the expected contribution threshold was set at 0.021 (i.e., (100/47)/100). After selecting the important features based on each criterion, their intersection was used to form the final set of important features (Table 16).

4.3 Wrapper Methods - RFE

Wrapper methods(Reference 15, Reference 16) assess the importance of features based on their usefulness during machine learning model training. These methods treat feature selection as a search problem, generating and evaluating different combinations using a predictive model. We employed RFE (Recursive Feature Elimination), a technique that iteratively adjusts a model and removes the least important feature until a desired number of features is reached. The importance of features is determined by the model itself. In our case, a RandomForest model was used, and feature importance was measured by information gain. The dynamic selection process runs RFE with the number of features varying between 1 and the maximum possible. To evaluate this process, the previously mentioned split data was used here (Table 16).

5 - Binary Classification

As there were two target variables which differ in nature, binary and multiclass, some adjustments had to be made for each one.

5.1 Final Feature Selection

To finalize our feature selection, we integrated the results from the various methods described above. The Chi-Square test for categorical features identified 14 variables, including *age*, *admission_source*, and *regular_patient*, among others. MutualInfo flagged *log_outpatient_visits_in_previous_year* and *log_emergency_visits_in_previous_year* as non-useful. RFE and Decision Tree methods determined their own feature sets, which coincided with the Chi-Square results for categorical features. In particular, RFE had the best model performance with only 8 variables. Lasso removed variables whose coefficients were 0, where among the 13 features deemed unimportant, only 2 were not medication features. Since filtering methods were target-independent, the previously mentioned drops were retained for both classification problems so we have 13 final features (Table 16).

5.2 Model Selection with Repeated K-Fold Cross Validation

In order to be more efficient in optimizing models and only select a few which show promising results, we conducted an initial test on a comprehensive list of models to understand the general performance of these models prior to optimization (Table 17). This way, we could focus on trying to improve models which already showed to be fit for this problem. The full list of models experimented with is Logistic Regression, KNN, Decision Tree, Support Vector Classifier, NaiveBayes, Bagging Classifier, Random Forest, AdaBoost, GradientBoosting, and MultiLayer Perceptron (Reference 8, 11, 12, 13, 20, 21). The hyperparameters were all set to default except in the Logistic Regression, where the inverse regularization strength ‘C’ was changed to 10. This change was made due to the model classifying all observations as ‘No’ for the target variable, with this adjustment being necessary to get meaningful results. Large values of this hyperparameter make the algorithm less constrained and might lead to overfit data, but given the extremely poor results of the model, this change was necessary. The number 10 was obtained from an optimization conducted specifically for this purpose. It is also important to note that as KNN uses Euclidean distances, the features used were restrained to the final numerical ones.

To select the models with the most potential, we calculated their average performance using Repeated K-Fold Cross Validation. This method involves splitting the data set into K subsets, then training and evaluating K times, with each fold being used as the validation set once and the remaining folds as the training set. This process is then repeated N times. We decided to keep the number of folds to the default value 5 since many models were going to be tested and the computational cost increases with the number of folds and to avoid the diminishing return of large K values. Due to computational restraints as well, the number of repeats was set to 2.

After evaluating the results, we selected models with the best average F1 scores in both training and validation sets for further fine-tuning: Logistic Regression, SV Classifier, Random Forest and Gradient Boosting. These models showed the best compromise between training and validation scores, overfitting less, and good predictions (comparatively high F1 scores).

5.3 Model Optimization

With the dataset prepared and the models selected, the next step involved optimizing hyperparameters for these models. For each model, a GridSearch was conducted when feasible to identify the optimal set of hyperparameters maximizing the F1 score. In cases where a GridSearch(Reference 24) was impractical due to computational costs, as seen with the Support Vector Classifier(Reference 8, 23), a RandomizedSearch(Reference 22) was employed. To explore a larger set of hyperparameters and an understanding of how the model responded to different combinations, the Random Forest was similarly optimized using this approach. The RandomizedSearch selected values for continuous hyperparameters by setting their distributions, such as a uniform distribution between 0.1 and 10.1 for the regularization of the C parameter for SVC, and randomly selected integers for discrete ones, such as integers between 1 and 20 for the max depth for Random Forests. After confirming the selected values through iterative experimentation, ensuring they fell within reasonable bounds and were not overly extreme, we assessed the need for further search expansion.

The Logistic Regression was tested with several ‘C’ values, a hyperparameter explained in the previous section, and different penalties (which add or not penalty terms to avoid overfitting) and solver options (which optimize and find the parameters of the loss function in different ways).

Gradient Boosting is an ensemble method which uses weak decision trees sequentially to improve their results. Here, the hyperparameters tested were the number of estimators (decision trees), the learning rate which controls the contributions of each estimator to the ensemble, max depth of the trees, minimum number of samples required for both a split and a leaf, and the percentage of the train data to use in training.

As the random forest is a set of several decision trees, the hyperparameters tested were similar to the Gradient Boosting: number of trees, maximum depth, minimum of samples for split and leaf. Other 2 different ones were the maximum number of features to look for a split and if bootstrap should be done, that is repeatedly drawing random subsets with replacement from the original dataset.

Finally, SVC was an extremely expensive model to optimize. The hyperparameters optimized were the ‘C’, which controls the trade-off between a boundary that fits to the data or a more smooth and generalized one; ‘gamma’ which defines the influence of each individual record on the boundary; and the kernel used to define the type of boundary. Nonlinear options were explored and this was the main expense of the optimization.

After this assessment, new rounds of experimentation were performed until the final hyperparameters were defined for each model. This paved the way for a conclusive model comparison and selection. The final comparison of optimized models and their hyperparameters is presented in the table...

5.4 Assessment and Final Model Choice

After fine-tuning the four predictive models, it became evident that the model with the highest F1 score did not necessarily excel in both precision and recall. To address this, we opted for an ensemble approach using a voting classifier, combining the model with the best recall and the model with the best precision. To assess recall, we employed an ROC curve, offering a comprehensive evaluation of classifier performance across sensitivity and specificity trade-offs. This led to the fusion of GradientBoost, which exhibited the highest Area Under the Curve (AUC) score, with Logistic Regression, recognized for its precision (Figure 12).

In evaluating the models and their respective scores, we conducted a thorough analysis considering their performance in both cross-validation and the Kaggle competition. Notably, the Random Forest Classifier emerged as the top performer overall. It demonstrated robustness, excelling in generalization for unseen data, accuracy by providing superior solutions, and efficiency by requiring minimal time for training and optimization (Table 18).

6 - Multiclass Classification

The multiclass problem uses the same dataset as the binary problem, but with a target variable featuring three categories, as previously explained. This alteration introduces some distinctions in the methods applied, as well as the outcomes of data treatment, models utilized, and optimization. In this context, the objective is to predict whether a patient will not be readmitted, be readmitted within 30 days, or be readmitted after 30 days.

6.1 Preprocessing Changes

The preprocessing steps remained largely consistent, as the steps followed were unrelated to the target variable. The same checks for incoherences and removal of outliers were performed, identical methods of imputing missing values were applied, and the feature engineering, encoding, and grouping of categories were kept constant. Following these preprocessing steps, MinMaxScaler was once again employed. Recognizing class imbalance within the three categories of the target variable (Figure 5), RandomUnderSampler was utilized (refer to section 10.1). The split for the exclusive use of RFE was also implemented at this stage.

6.2 Final Feature Selection

In the feature selection phase, we observe the first notable changes. As mentioned earlier, filter methods did not consider the target variable, resulting in the removal of variables with zero variance (Table 19). The Chi-Square test for categorical features identified a total of 22 features, which is 8 more than in the binary problem(Table 20). MutualInfo was applied similarly to before, but this time it excluded the feature *number_lab_tests* from the set (Table 21). Due to having the same features as before, *visits_in_previous_year* was once again dropped due to multicollinearity (Figure 13). The Decision Tree, using both 'gini' and 'entropy', selected another set of 14 features that contributed more than the expected value. Lasso removed 11 features, with 10 of them being medication features.

The final set was defined again as features selected by at least 2 models, so we have 13 features after confirming with the filter methods (Table 22).

6.3 Model Selection with Repeated K-Fold Cross Validation

For model selection, we again tested several models to only use a selected few for further optimization and scoring. Using the same idea as last model selection, we have used Repeated K-Fold with the number of folds set to 5 and the number of repeats set to 2. The set of the models to be tested remained the same, although the Logistic Regression was set to multinomial in order to handle the multiclass nature of the problem. After analyzing how robust and accurate the models were, the same 4 models were chosen, with AdaBoost being a close choice (Table 23).

6.4 Model Optimization

Optimization followed the same principles as before: for computationally expensive models, a RandomizedSearch was employed, while for all others, a GridSearch was implemented. Since the same models were selected, similar choices for both search methods were explored. The same RepeatedKFold method was utilized for optimization.

The Logistic Regression had the particular change of being used with the multinomial option for the parameter *‘multi_class’* to ensure meaningful results. Following the optimization of the model, the same penalty was applied, but with a significantly higher 'C' value of 1000. This indicates that logistic regression required very low penalties for accurate predictions, necessitating a more intense capture and learning of the training data.

The GradientBoosting classifier ended up utilizing a similar number of estimators (7 more), while featuring a higher max_depth and a slightly elevated learning rate. This conclusion reinforces the findings from the logistic regression, indicating that the model requires more complex and less generalized trees, and greater learning from each estimator to improve predictions.

The RandomForest also used more estimators, with much greater depth. The minimum number of samples for a split to occur decreased.

The SVC revealed itself again to be an extremely expensive model. For this task, it was necessary to force the model to use a linear kernel (which might not be the optimal solution) in order to produce results.

6.5 Assessment and Final Model Choice

The final model selection was determined by meeting specific criteria to ensure appropriateness. These criteria included consistent scores, where significant differences between training and evaluation scores were avoided. Additionally, emphasis was placed on efficiency in terms of computation and time. These criteria proved essential, as some tested models consumed excessive time without justifiable improvements in scores.

The chosen models, specifically Gradient Boosting, not only exhibited balanced performance across various metrics but were also efficient in terms of computation, delivering results swiftly without compromising on accuracy.

7. Final Results and Discussion

The F1 score emerged as the preferred metric during model selection and performance evaluation due to its consideration of both false positives and false negatives - crucial factors in a medical setting. It is noteworthy that for binary models, the F1 score is based on the evaluation of class 1, while for multiclass models, it represents a weighted average across all classes. While the F1 score served as the primary metric, it is essential not to overlook other metrics offering additional perspectives on predictions. To obtain a comprehensive understanding, accuracy, recall, and precision were also analyzed. Examining Table 24, the accuracy, recall, precision, and F1 scores of the four final chosen models were assessed for both binary and multiclass problems using RepeatedKFold. Notably, the scores for the multiclass problem were consistently lower, underscoring the increased difficulty in accurately predicting across multiple classes. As mentioned earlier, the logistic regression model exhibited the best precision for the binary problem, whereas the Gradient Boosting classifier excelled in the multiclass scenario. These models demonstrated superior correctness in positive predictions. Regarding recall, Gradient Boosting performed exceptionally well in the binary problem, while in the multiclass scenario, it was matched by the Support Vector Classifier (SVC). The Gradient Boosting Classifier achieved the highest F1 score for both binary and multiclass problems. It is crucial to acknowledge the contribution of the Voting Classifier, which consistently delivered high scores across all four metrics. While it did not excel in every measure, its balanced performance warrants recognition.

8. Conclusion, Limitations, and Further Improvements

This project aimed to address the challenge of hospital readmissions, particularly focusing on diabetic patients. Through extensive data exploration, preprocessing, and feature engineering, we prepared the dataset for building predictive models, which were then optimized to improve the predictive power of said models. The project utilized binary and multiclass classification to predict readmissions and the timeframe of readmission, respectively.

The objective of building a predictive model that was able to predict readmissions was fulfilled, although the model might not be fail-proof. The multifaceted nature of factors influencing readmissions emphasized the inherent challenges in creating a model that can provide absolute certainty. The findings and insights from this project contribute to understanding factors influencing hospital readmissions and provide a foundation for developing strategies to prevent them.

Several limitations need to be acknowledged. A constrained timeframe and limited computational resources posed challenges during the execution of various stages. The intricate nature of medical data, coupled with the complexity of predicting readmissions accurately, added to the complexity of the task. The slow improvements observed in predictions, while consistent, revealed how difficult and arduous it was to accurately predict a target in a medical setting.

To improve the results of this project, more time and more computational capability would fulfill the requirements to search for enhanced solutions. A collaboration with medical specialists could potentially provide more information about the data at hand and a deeper understanding that goes beyond correlations and feature importance measures.

In summary, while this project represents a positive step forward, acknowledging its limitations and addressing them will likely lead to improved results.

9- References

1 - Prayer Type List - Health Safety Net https://www.mass.gov/doc/hsn-payer-code-list20130118pdf/download

2 - Diabetes 130-US hospitals for years 1999-2008 - UC Irvine https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

3 - Predictive Modelling Of Hospital Readmissions In Diabetic Patients Clusters - Anne Monteiro Mendes de Senna https://run.unl.pt/bitstream/10362/145706/1/TGI1636.pdf

4 - Metrics For Multi-Class Classification - Margherita Grandini, Enrico Bagli, Giorgio Visani https://arxiv.org/pdf/2008.05756.pdf,

5 - The choice of scaling technique matters for classification performance - Lucas B.V. de Amorim, George D.C. Cavalcanti, Rafael M.O. Cruz https://arxiv.org/pdf/2212.12343.pdf

6 - Understanding Your A1C Test - American Diabets Association https://professional.diabetes.org/sites/default/files/media/ada-factsheet-understandingyoura1ctest.pdf

7 - The A1C Test and Diabetes - National Institute Of Diabetes And Digestive And Kidney Diseases https://www.govinfo.gov/content/pkg/GOVPUB-HE20-PURL-gpo22893/pdf/GOVPUB-HE20-PURL-gpo22893.pdf

8 - Survey on Multiclass Classification Methods - Mohamed Aly https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a546f2c88c588a2a46c054f67b39a3ebefdae694

9 - Foundations of Data Imbalance and Solutions for a Data Democracy - Ajay Kulkarni, Feras A. Batarseh, and Deri Chong https://arxiv.org/ftp/arxiv/papers/2108/2108.00071.pdf

10 - Comparision Of Undersampling Methods For Prediction Of Casting Defects Based On Process Parameters - Simon Lööv https://www.diva-portal.org/smash/get/diva2:1597599/FULLTEXT01.pdf

11 - Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning - Sebastian Raschka University of Wisconsin–Madison Department of Statistics

https://arxiv.org/pdf/1811.12808.pdf

12 - Model Selection Techniques - Jie Ding, Vahid Tarokh, and Yuhong Yang

https://arxiv.org/pdf/1810.09583.pdf

13 - Model Selection - Arturo Perez Adroher

https://essay.utwente.nl/76580/7/Master_Thesis_Perez_openbaar-12.pdf

14 - Feature Selection - CMU School of Computer Science

https://www.cs.cmu.edu/~kdeng/thesis/feature.pdf

15 - Journal of Machine Learning Research - Isabelle Guyon, Andre Elisseeff

https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

16 - Feature Selection for Classification: A Review - Jiliang Tang, Salem Alelyani and Huan Liu

https://www.cse.msu.edu/~tangjili/publication/feature_selection_for_classification.pdf

17 - The Chi-Square Test - Diana Mindrila, Ph.D. Phoebe Balentyne, M.Ed.

https://www.westga.edu/academics/research/vrc/assets/docs/ChiSquareTest_LectureNotes.pdf

18 - Principal Components Analysis https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf

19 - PRINCIPAL COMPONENTS ANALYSIS (PCA) - *Steven M. Holand Department of Geology, University of Georgia, Athens, GA 30602-2501* http://stratigrafia.org/software/pdf/pcaTutorial.pdf

20 - MULTI-LAYER PERCEPTRON https://egyankosh.ac.in/bitstream/123456789/12687/1/Unit-6.pdf

21 - Boosting Algorithms: A Review of Methods, Theory, and Applications - Artur Ferreira and Mario Figueiredo https://repositorio.ipl.pt/bitstream/10400.21/1853/4/Boosting_AFerreira.pdf

22 - Random Search for Hyper-Parameter Optimization - James Bergstra, Yoshua Bengio

https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

23 - Support Vector Machines: Hype or Hallelujah? - Kristin P. Bennet, Colin Campbell

https://kdd.org/exploration_files/bennett.pdf

24 - Machine Learning Model Optimization with Hyper Parameter Tuning Approach - Md Riyad Hossain & Dr. Douglas Timmer https://core.ac.uk/download/pdf/539593628.pdf

25-The project was developed with the support of the Machine Learning class contents, made available on the online platform Moodle

10 - Annexes

10.1 RandomUnderSampler

Undersampling is a technique used in machine learning to address the class imbalance issue in a dataset. For our project, we used the Random Undersampling method to create a more balanced dataset. Random Undersampling is one specific technique among various under-sampling methods used to address the issue of class imbalance in machine learning datasets. The main difference between Random Undersampling and other under-sampling techniques lies in how they select which instances to remove from the majority class. Random under-sampling randomly selects instances from the majority class without considering their specific characteristics or their relationship with other instances. Before Undersampling, class '1': 7950 instances, class '0': 62134 instances. The shape of the training set after random under-sampling is (15900, 51), indicating that the number of instances has been reduced to 15900, and there are 51 features in each instance. Class '1' and class '0' now have 7950 instances, balancing the class distribution. After applying random under-sampling, the class distribution is now balanced, with equal representation from both classes.
This balanced training set can be used to train machine learning models that are less biased towards the majority class, potentially improving the model's ability to generalize to both classes.

10.2 PCA

PCA is a dimensionality reduction technique designed to diminish the number of variables by creating composite indices. These indices are linear combinations of the original variables, with the objective of capturing the maximum variance in the data using a smaller set of indices. After standardizing the dataset, the covariance matrix is computed, revealing how different variables change together. Next, eigenvectors and eigenvalues of this matrix are calculated. The eigenvectors represent directions of maximum variance, while the eigenvalues quantify the magnitude of variance along these directions. The eigenvectors are then sorted based on their corresponding eigenvalues, with the highest eigenvalue indicating the direction of maximum variance. By selecting a specific number of these eigenvectors, termed principal components, a new basis for the data is established. This transformation matrix is then used to project the original data onto the space defined by the selected principal components. The outcome is a reduced-dimensional representation of the data, preserving the most significant information. In essence, PCA identifies a set of uncorrelated axes, called principal components, where the first component accounts for the maximum variance, and subsequent components capture the residual variance in descending order.

The primary motivation for applying PCA is to address the curse of dimensionality by creating a reduced set of features that convey the dataset's information based on the original variables. This, in turn, enhances the speed of training and optimizing models. To ensure quality results, the StandardScaler was used instead of the MinMaxScaler to ensure unit variance and zero mean.

To determine the number of components to retain, two well-established methods were employed:

- Scree plot method: This involves plotting the percentage of variance explained by each principal component and identifying an "elbow" point in the graph.
- Pearson's criteria: Retaining every principal component until a specified threshold, in this case, until 80% of the variance is explained.

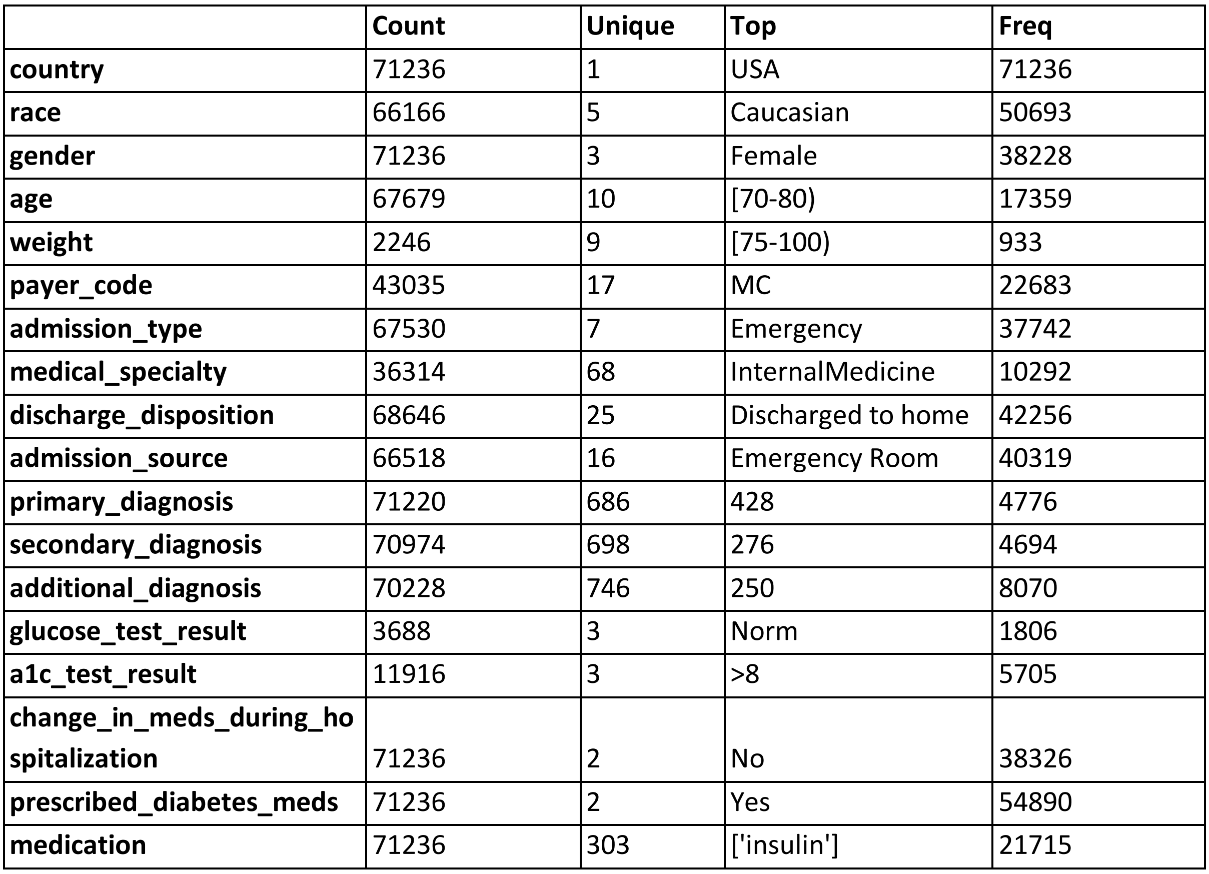
Analysis of the Scree Plot revealed an inflection point at 15 principal components, a finding supported by Pearson's criteria (Figure 14). The first principal component explains 18% of the variance in the original data, while each subsequent component starting from the 5th, contributes less than 5% (Figure 15).

Due to the substantial number of principal components needed to reach a significant amount of variance explained and their limited ability to capture variance effectively, PCA was not utilized in the predictive models. Being an unsupervised technique, the decision was final for both the binary and multiclass problems. Instead, the analysis proceeded with feature selection.

10.3 Figures and Tables

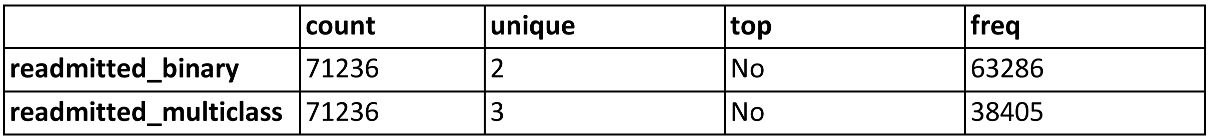
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
average_pulse_bpm	71236	99.611	23.041	60	80	100	119	139
length_of_stay_in_hospital	71236	4.391	2.989	1	2	4	6	14
number_lab_tests	71236	43.096	19.643	1	31	44	57	121
non_lab_procedures	71236	1.341	1.707	0	0	1	2	6
number_of_medications	71236	15.995	8.122	1	10	15	20	75
number_diagnoses	71236	7.421	1.938	1	6	8	9	16
outpatient_visits_in_previous_year	71236	0.37	1.287	0	0	0	0	42
emergency_visits_in_previous_year	71236	0.196	0.911	0	0	0	0	76
inpatient_visits_in_previous_year	71236	0.64	1.267	0	0	0	1	21

Table 1 - Numerical Descriptive Statistics



	Count	Unique	Top	Freq
country	71236	1	USA	71236
race	66166	5	Caucasian	50693
gender	71236	3	Female	38228
age	67679	10	[70-80)	17359
weight	2246	9	[75-100)	933
payer_code	43035	17	MC	22683
admission_type	67530	7	Emergency	37742
medical_specialty	36314	68	InternalMedicine	10292
discharge_disposition	68646	25	Discharged to home	42256
admission_source	66518	16	Emergency Room	40319
primary_diagnosis	71220	686	428	4776
secondary_diagnosis	70974	698	276	4694
additional_diagnosis	70228	746	250	8070
glucose_test_result	3688	3	Norm	1806
a1c_test_result	11916	3	>8	5705
change_in_meds_during_hospitalization	71236	2	No	38326
prescribed_diabetes_meds	71236	2	Yes	54890
medication	71236	303	['insulin']	21715

Table 2 - Categorical Descriptive Statistics



	count	unique	top	freq
readmitted_binary	71236	2	No	63286
readmitted_multiclass	71236	3	No	38405

Table 3 - Targets Descriptive Statistics

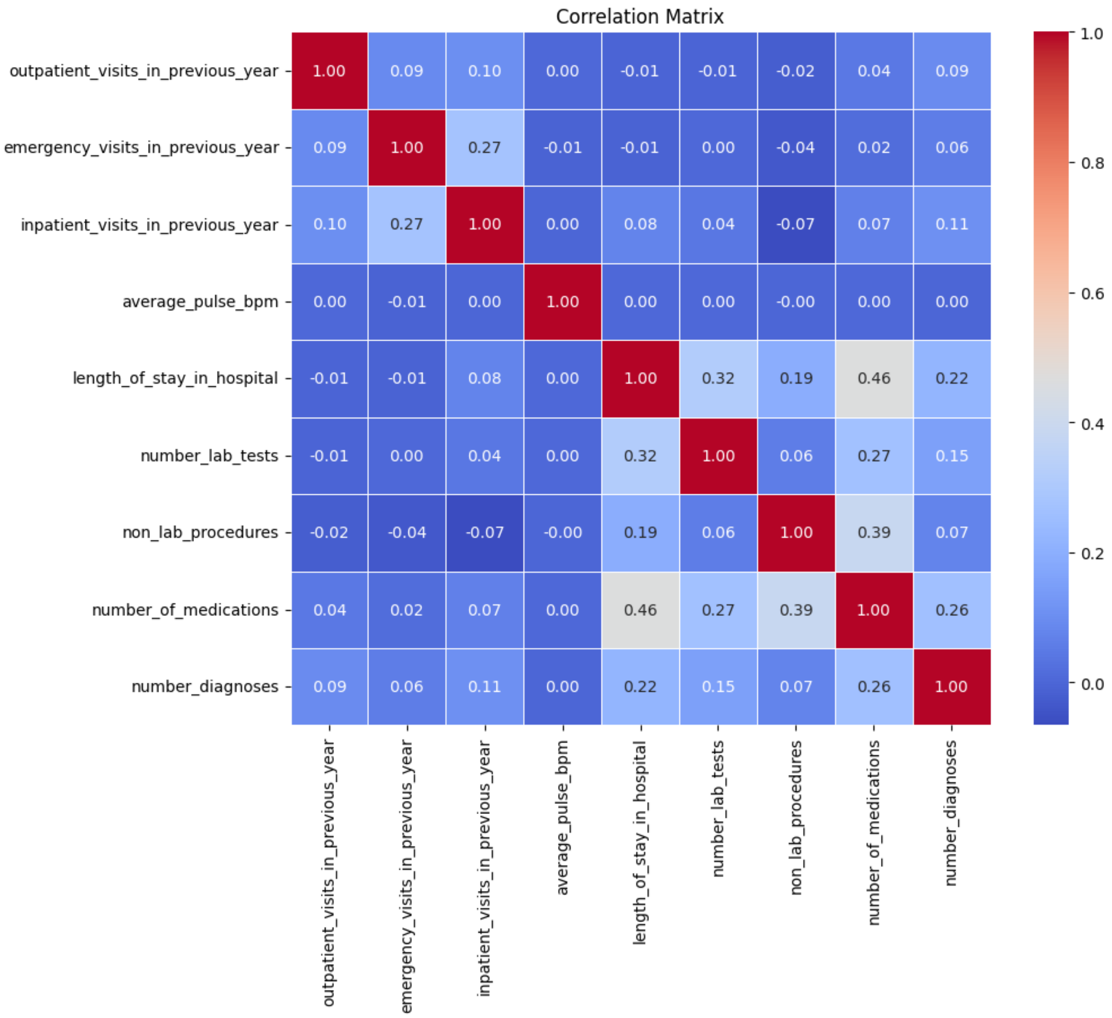
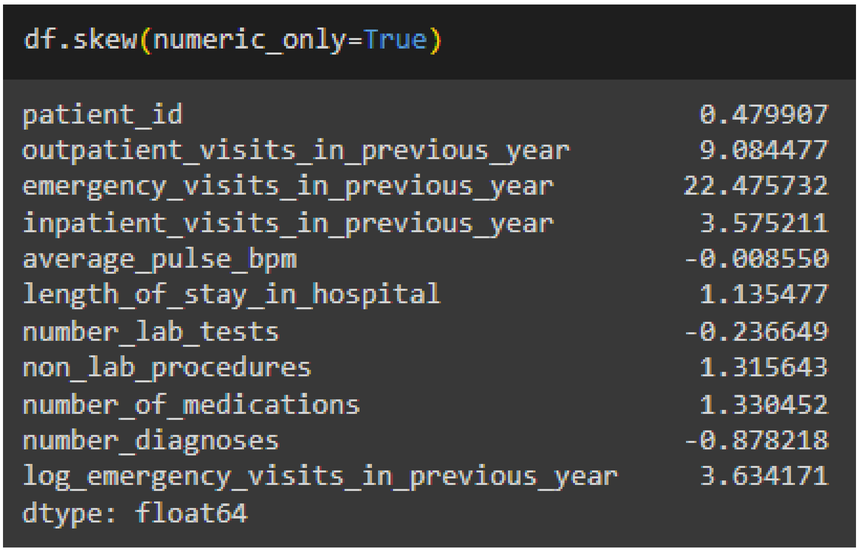


Figure 1 - Numarical Features Correlation Matrix



```
df.skew(numeric_only=True)
```

```
patient_id                              0.479907
outpatient_visits_in_previous_year      9.084477
emergency_visits_in_previous_year      22.475732
inpatient_visits_in_previous_year       3.575211
average_pulse_bpm                      -0.008550
length_of_stay_in_hospital              1.135477
number_lab_tests                       -0.236649
non_lab_procedures                      1.315643
number_of_medications                   1.330452
number_diagnoses                       -0.878218
log_emergency_visits_in_previous_year   3.634171
dtype: float64
```

Figure 2 - Numerical Features Skewness

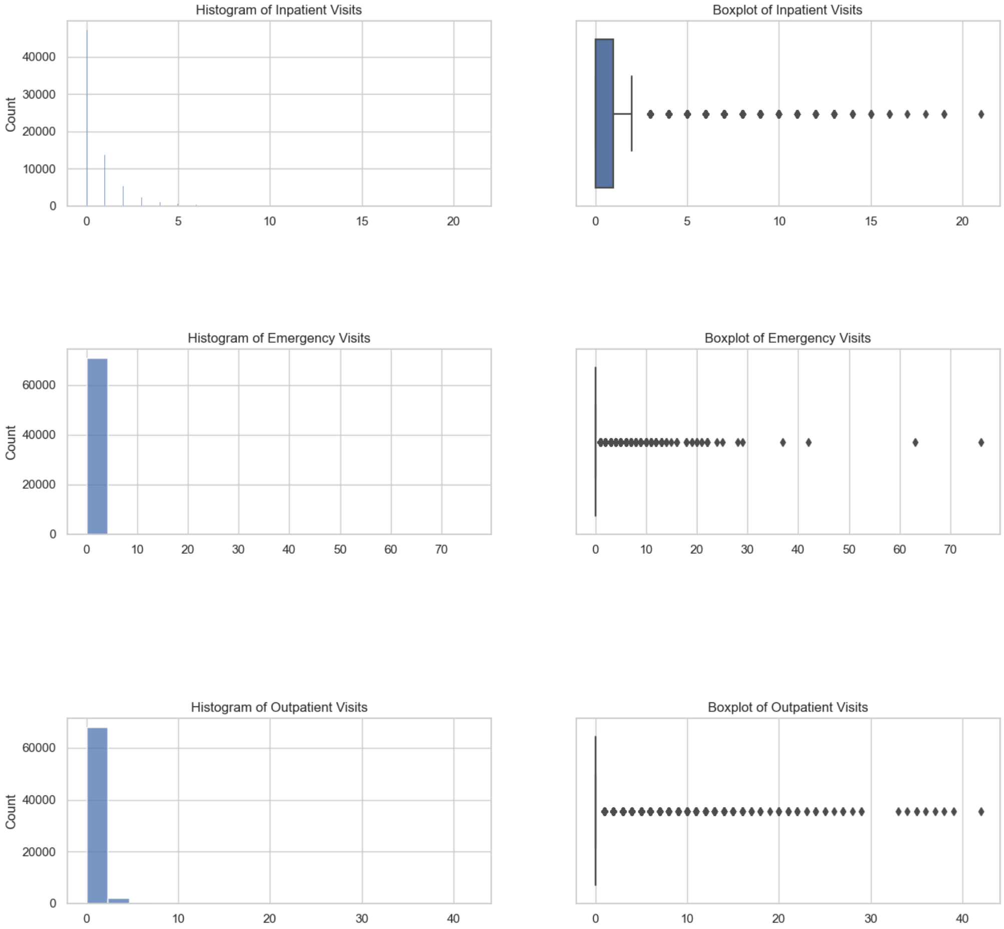


Figure 3 - Histograms and Box plots for different type of visits in previous year

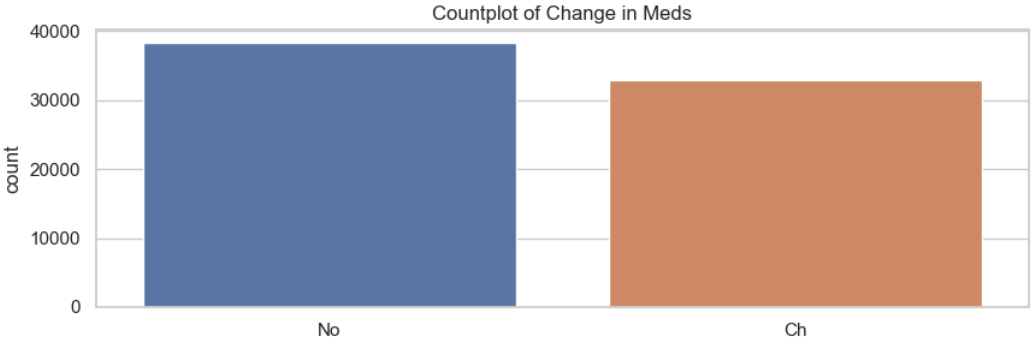


Figure 4- Count plot for Change in Meds

prescribed_diabetes_meds	No	Yes	Total

change_in_meds_during_hospitalization			
Ch	0	32910	32910
No	16346	21980	38326
Total	16346	54890	71236

Table 4 - Relation between *prescribed_diabetes_meds* and *change_in_meds_during_hospitalization*

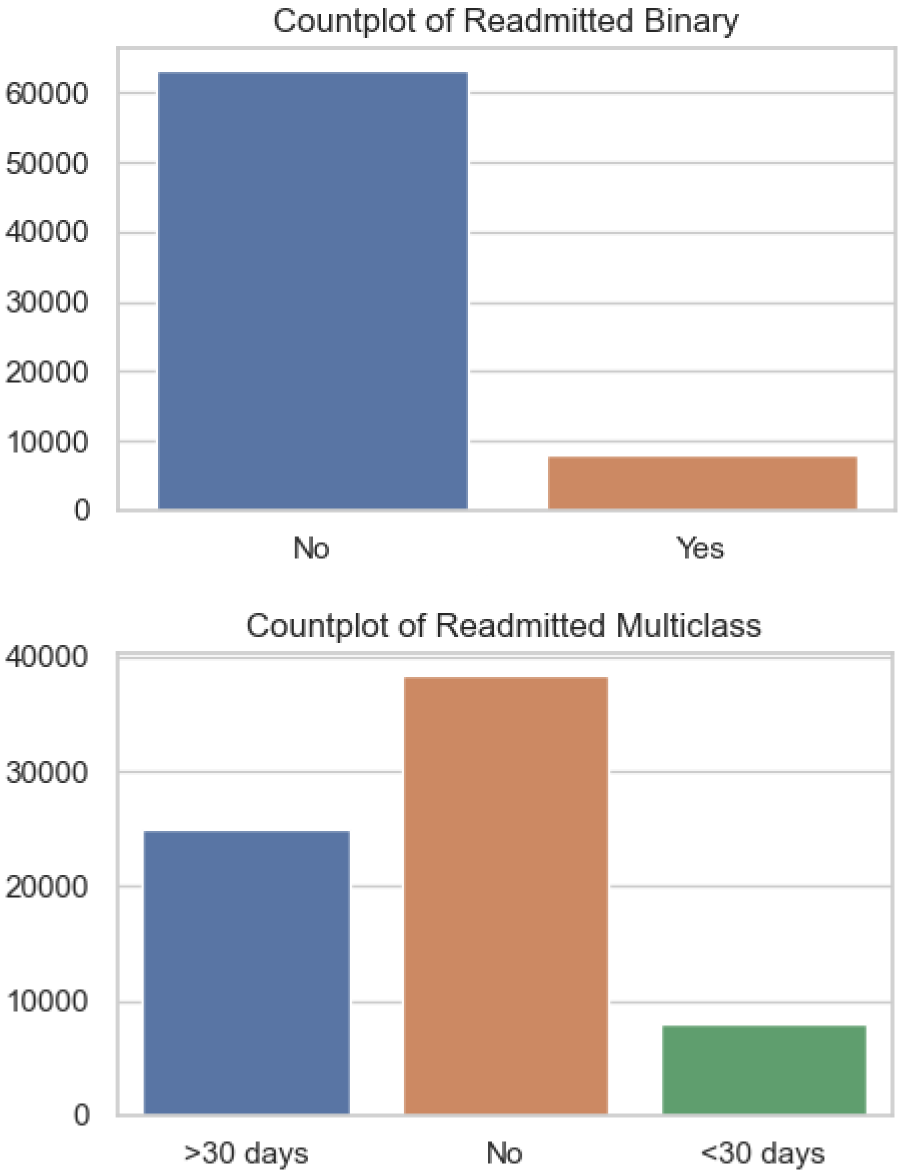
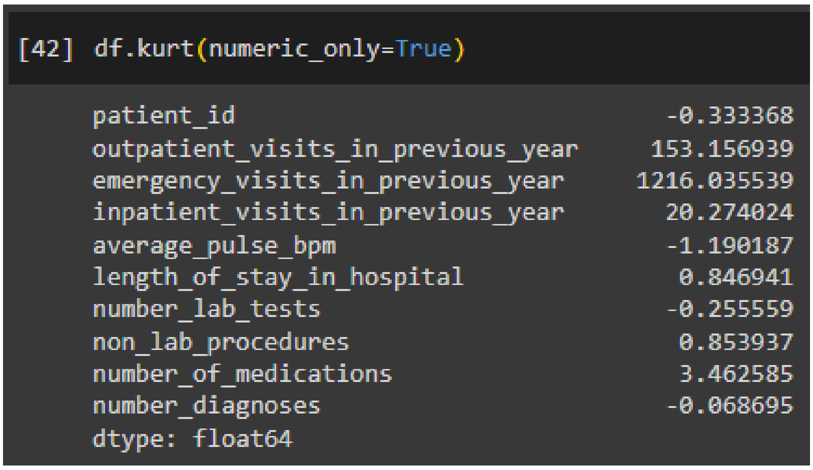


Figure 5 - Distribution of the Binary and Multiclass target features



```
[42]  df.kurt(numeric_only=True)

      patient_id                            -0.333368
      outpatient_visits_in_previous_year   153.156939
      emergency_visits_in_previous_year   1216.035539
      inpatient_visits_in_previous_year     20.274024
      average_pulse_bpm                     -1.190187
      length_of_stay_in_hospital             0.846941
      number_lab_tests                      -0.255559
      non_lab_procedures                     0.853937
      number_of_medications                  3.462585
      number_diagnoses                      -0.068695
      dtype: float64
```

Figure 6 - Numerical Features Kurtosis

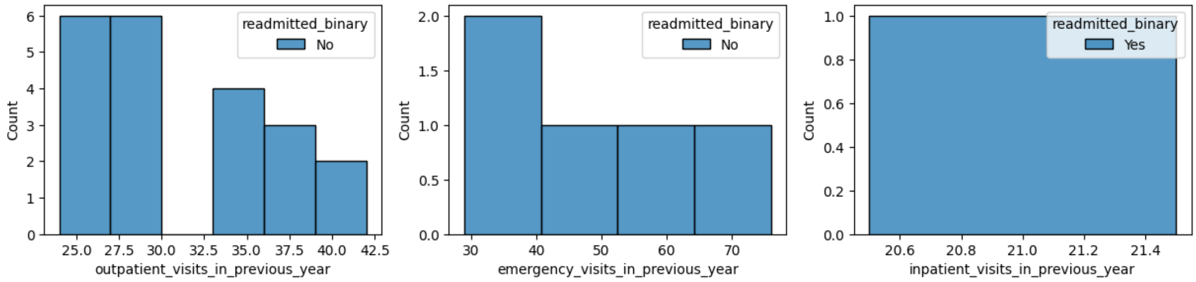
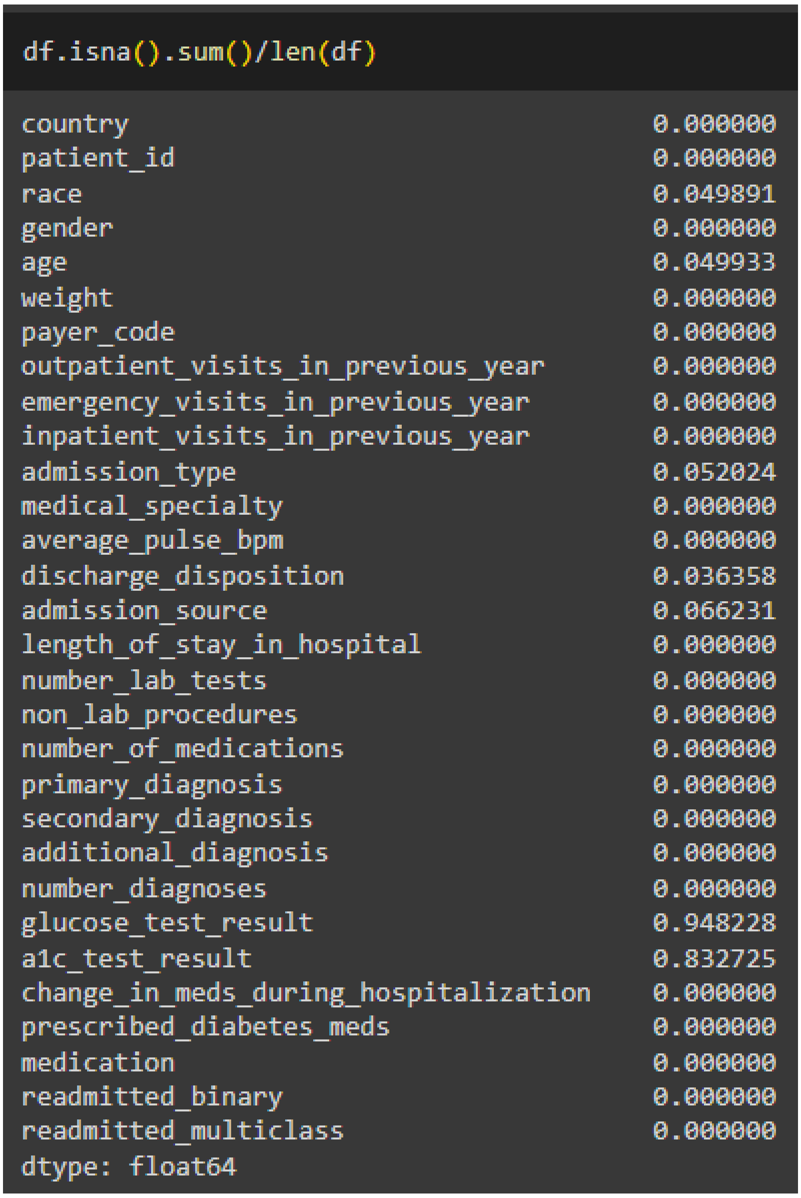


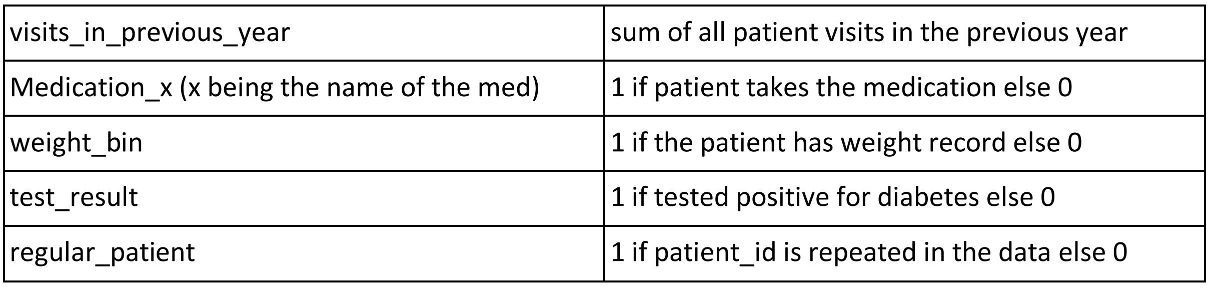
Figure 7 - Distribution from patients that do not return on visitis in previous year features



```
df.isna().sum()/len(df)
```

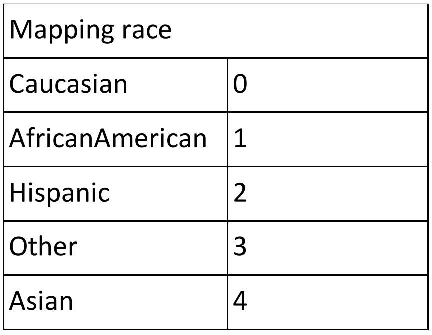
```
country                                 0.000000
patient_id                              0.000000
race                                    0.049891
gender                                  0.000000
age                                     0.049933
weight                                  0.000000
payer_code                              0.000000
outpatient_visits_in_previous_year      0.000000
emergency_visits_in_previous_year       0.000000
inpatient_visits_in_previous_year       0.000000
admission_type                          0.052024
medical_specialty                       0.000000
average_pulse_bpm                       0.000000
discharge_disposition                   0.036358
admission_source                        0.066231
length_of_stay_in_hospital              0.000000
number_lab_tests                        0.000000
non_lab_procedures                      0.000000
number_of_medications                   0.000000
primary_diagnosis                       0.000000
secondary_diagnosis                     0.000000
additional_diagnosis                    0.000000
number_diagnoses                        0.000000
glucose_test_result                     0.948228
a1c_test_result                         0.832725
change_in_meds_during_hospitalization   0.000000
prescribed_diabetes_meds                0.000000
medication                              0.000000
readmitted_binary                       0.000000
readmitted_multiclass                   0.000000
dtype: float64
```

Figure 8 - Percentage of missing values



visits_in_previous_year	sum of all patient visits in the previous year
Medication_x (x being the name of the med)	1 if patient takes the medication else 0
weight_bin	1 if the patient has weight record else 0
test_result	1 if tested positive for diabetes else 0
regular_patient	1 if patient_id is repeated in the data else 0

Table 5 - New Features

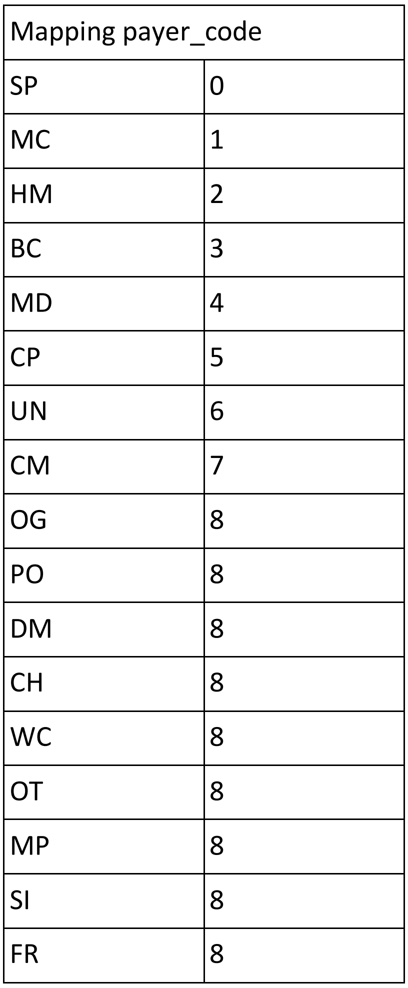


Mapping race	
Caucasian	0
AfricanAmerican	1
Hispanic	2
Other	3
Asian	4

Table 6 - *race* Mapping

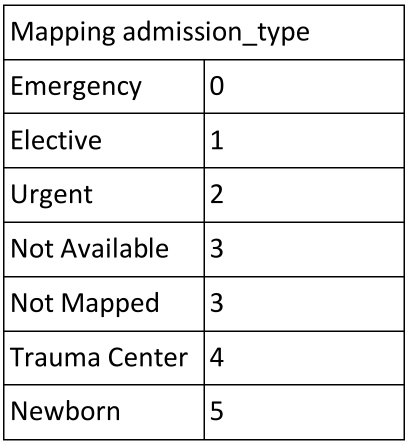
Mapping age	
[0-10)	5
[10-20)	15
[20-30)	25
[30-40)	35
[40-50)	45
[50-60)	55
[60-70)	65
[70-80)	75
[80-90)	85
[90-100)	95

Table 7 - *age* Mapping



Mapping payer_code	
SP	0
MC	1
HM	2
BC	3
MD	4
CP	5
UN	6
CM	7
OG	8
PO	8
DM	8
CH	8
WC	8
OT	8
MP	8
SI	8
FR	8

Table 8 - *payer_code* Mapping



Mapping admission_type	
Emergency	0
Elective	1
Urgent	2
Not Available	3
Not Mapped	3
Trauma Center	4
Newborn	5

Table 9 - *admission_type* Mapping

SPECIALITY	MEDICAL SPECIALITY	NUMERICAL VALUE
Allergy and Immunology	AllergyandImmunology	0
Anatomic and Clinical Pathology	Pathology	1
Anesthesiology	Anesthesiology	2
	Anesthesiology-Pediatric	2
Dermatology	Dermatology	3
Diagnostic Radiology	Radiologist	4
	Radiology	4
Emergency Medicine	Emergency/Trauma	5
Family, GeneralPractice	Family/GeneralPractice	6
General Surgery	Surgeon	7
	Surgery-Cardiovascular	7
	Surgery-Cardiovascular/Thoracic	7
	Surgery-Colon&Rectal	7
	Surgery-General	7
	Surgery-Maxillofacial	7
	Surgery-Neuro	7
	Surgery-Pediatric	7
	Surgery-Plastic	7
	Surgery-Thoracic	7
	Surgery-Vascular	7
	SurgicalSpecialty	7
	Surgery-PlasticwithinHeadandNeck	7
Internal medicine	Cardiology	8
	DCPTEAM	8
	Endocrinology	8
	Endocrinology-Metabolism	8
	Gastroenterology	8

	Hematology	8
	Hematology/Oncology	8
	Hospitalist	8
	InfectiousDiseases	8
	InternalMedicine	8
	Nephrology	8
	Neurophysiology	8
	Oncology	8
	Proctology	8
	Pulmonology	8
	Rheumatology	8
	SportsMedicine	8
	Urology	8
Neurology	Neurology	9
Obstetrics and Gynecology	Gynecology	10
	Obsterics&Gynecology-GynecologicOnco	10
	Obstetrics	10
	ObstetricsandGynecology	10
Ophthalmology	Ophthalmology	11
Orthopaedic Surgery	Orthopedics	12
	Orthopedics-Reconstructive	12
Osteopathic Neuromusculoskeletal Medicine	Osteopath	13
Other Healthcare Practitioners	Dentistry	14
	Podiatry	14
	Psychology	14
	Resident	14

	Speech	14
Otolaryngology-Head and Neck Surgery	Otolaryngology	15
Pediatrics	Cardiology-Pediatric	16
	Pediatrics	16
	Pediatrics-AllergyandImmunology	16
	Pediatrics-CriticalCare	16
	Pediatrics-EmergencyMedicine	16
	Pediatrics-Endocrinology	16
	Pediatrics-Hematology-Oncology	16
	Pediatrics-InfectiousDiseases	16
	Pediatrics-Neurology	16
	Pediatrics-Pulmonology	16
	Perinatology	16
	Psychiatry-Child/Adolescent	16
Physical Medicine and Rehabilitation	PhysicalMedicineandRehabilitation	17
	PhysicianNotFound	17
Psychiatry	Psychiatry	18
	Psychiatry-Addictive	18
Other	Other	19
	OutreachServices	19

Table 10 - *medical_specialty* Mapping

	Discharge Disposition	**NUMERICAL VALUE**
Discharged to home	Discharged to home	0
	Discharged/transferred to home with home health service	0
	Discharged/transferred to home under care of Home IV provider	0
Transferred to another medical facility	Discharged/transferred to a federal health care facility	1
	Discharged/transferred to SNF	1
	Discharged/transferred to another short term hospital	1
	Discharged/transferred to ICF	1
	Discharged/transferred to another rehab fac including rehab units of a hospital	1
	Discharged/transferred to a long term care hospital	1
	Discharged/transferred to another type of inpatient care institution	1
	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital	1
	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare	1
	Discharged/transferred/referred another institution for outpatient services	1
	Neonate discharged to another hospital for neonatal aftercare	1
Left AMA(Against Medical Advice.)	Left AMA	2
Still patient/referred to this institution	Admitted as an inpatient to this hospital	3
	Still patient or expected to return for outpatient services	3
	Discharged/transferred within this institution to Medicare approved swing bed	3
	Discharged/transferred/referred to this institution for outpatient services	3
Not Available	Not Mapped	4

	Not Available	4
Hospice	Hospice / home	5
	Hospice / medical facility	5
Expired	Expired	6
	Expired at home. Medicaid only, hospice	6
	Expired in a medical facility. Medicaid only, hospice	6

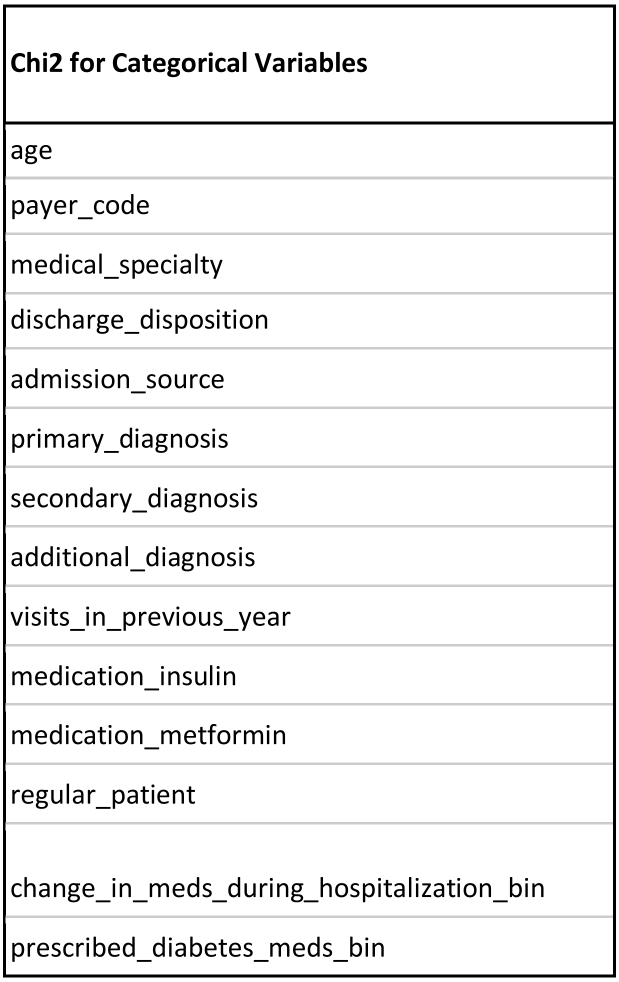
Table 11 - *discharge_disposition* Mapping

	ADMISSION SOURCE	**NUMERICAL VALUE**
Emergency Room	Emergency Room	0
Referral	Clinic Referral	1
	HMO Referral	1
	Physician Referral	1
From Hospital	Transfer from another health care facility	2
	Transfer from a hospital	2
	Transfer from a Skilled Nursing Facility (SNF)	2
	Transfer from hospital inpt/same fac reslt in a sep claim	2
Court/Law Enforcement	Court/Law Enforcement	3
Not Mapped	Not Available	4
	Not Mapped	4
From Critical Acess	Transfer from critial access hospital	5
	Transfer from Ambulatory Surgery Center	5
Sick baby	Sick Baby	6
	Extramural Birth	6
Normal Delivery	Normal Delivery	7

Table 12 - *admission_source* Mapping

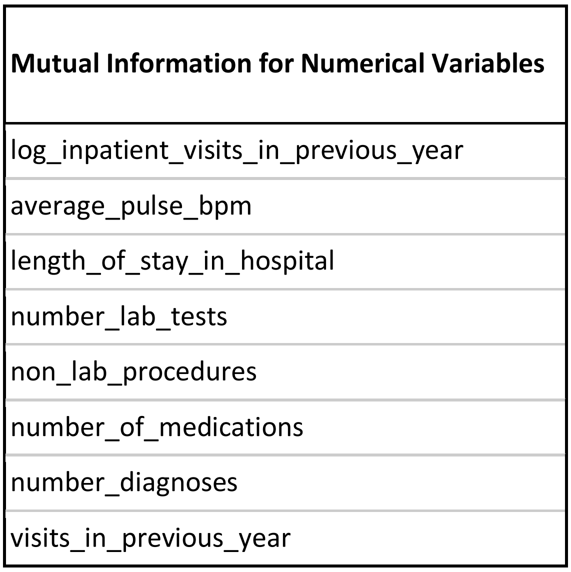
Univariate variables
medication_acetohexamide
medication_glimepiride-pioglitazone
medication_metformin-pioglitazone
medication_metformin-rosiglitazone
medication_troglitazone

Table 13 -



Chi2 for Categorical Variables
age
payer_code
medical_specialty
discharge_disposition
admission_source
primary_diagnosis
secondary_diagnosis
additional_diagnosis
visits_in_previous_year
medication_insulin
medication_metformin
regular_patient
change_in_meds_during_hospitalization_bin
prescribed_diabetes_meds_bin

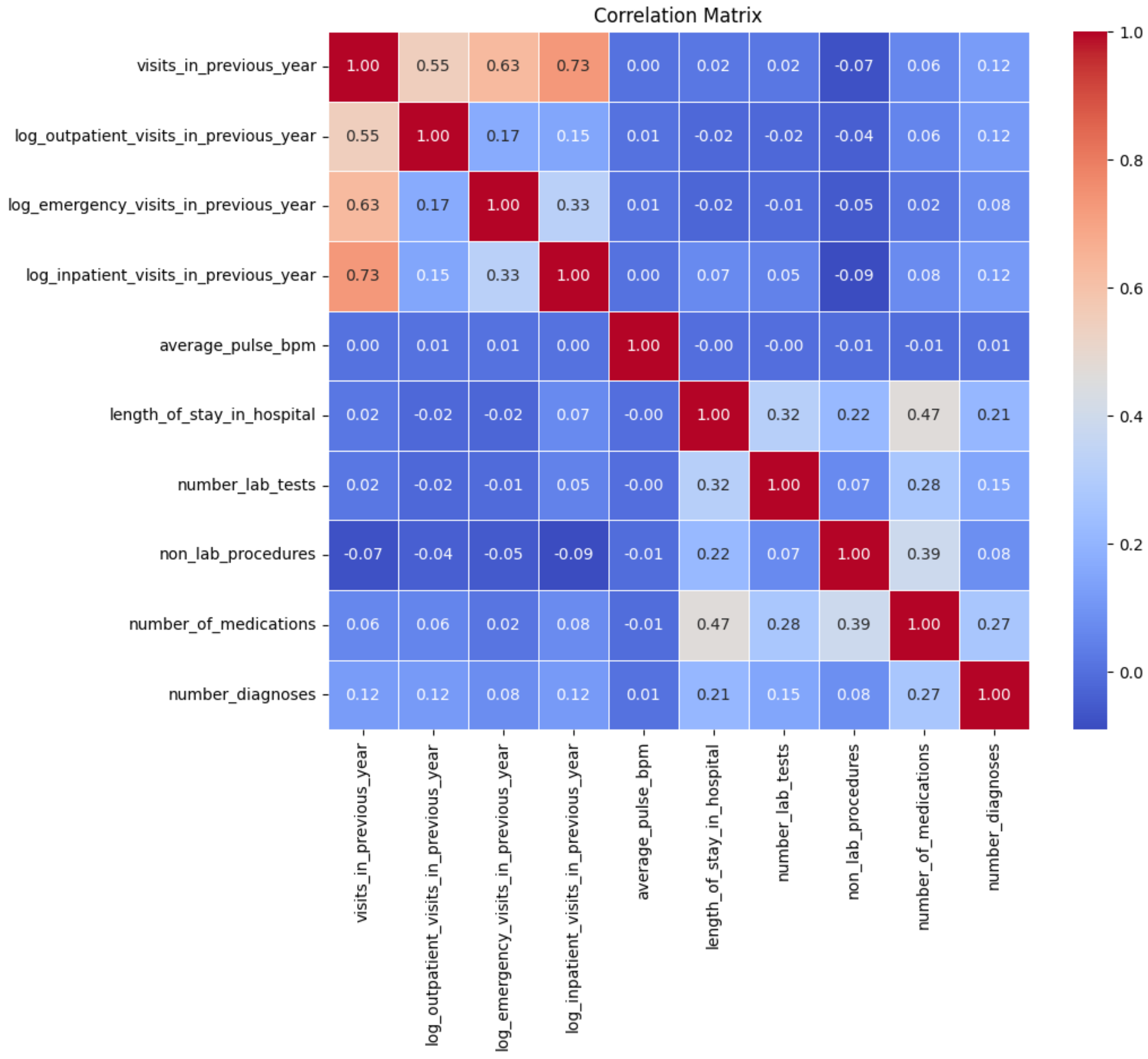
Table 14 -



Mutual Information for Numerical Variables
log_inpatient_visits_in_previous_year
average_pulse_bpm
length_of_stay_in_hospital
number_lab_tests
non_lab_procedures
number_of_medications
number_diagnoses
visits_in_previous_year

Table 15 -

Spearman Correlation



	visits_in_previous_year	log_outpatient_visits_in_previous_year	log_emergency_visits_in_previous_year	log_inpatient_visits_in_previous_year	average_pulse_bpm	length_of_stay_in_hospital	number_lab_tests	non_lab_procedures	number_of_medications	number_diagnoses
visits_in_previous_year	1.00	0.55	0.63	0.73	0.00	0.02	0.02	-0.07	0.06	0.12
log_outpatient_visits_in_previous_year	0.55	1.00	0.17	0.15	0.01	-0.02	-0.02	-0.04	0.06	0.12
log_emergency_visits_in_previous_year	0.63	0.17	1.00	0.33	0.01	-0.02	-0.01	-0.05	0.02	0.08
log_inpatient_visits_in_previous_year	0.73	0.15	0.33	1.00	0.00	0.07	0.05	-0.09	0.08	0.12
average_pulse_bpm	0.00	0.01	0.01	0.00	1.00	-0.00	-0.00	-0.01	-0.01	0.01
length_of_stay_in_hospital	0.02	-0.02	-0.02	0.07	-0.00	1.00	0.32	0.22	0.47	0.21
number_lab_tests	0.02	-0.02	-0.01	0.05	-0.00	0.32	1.00	0.07	0.28	0.15
non_lab_procedures	-0.07	-0.04	-0.05	-0.09	-0.01	0.22	0.07	1.00	0.39	0.08
number_of_medications	0.06	0.06	0.02	0.08	-0.01	0.47	0.28	0.39	1.00	0.27
number_diagnoses	0.12	0.12	0.08	0.12	0.01	0.21	0.15	0.08	0.27	1.00

Figure 9 -

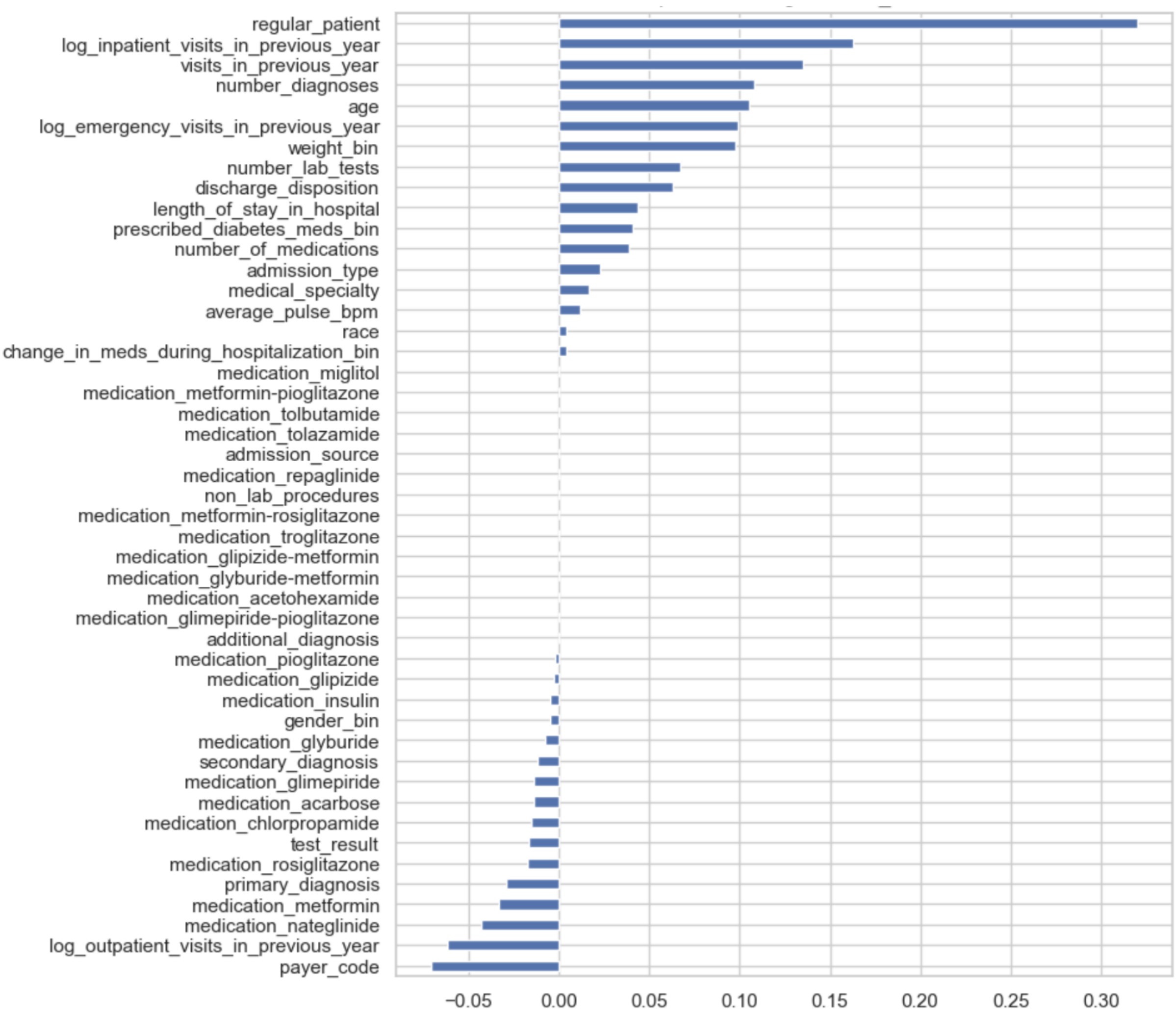


Figure 10 -

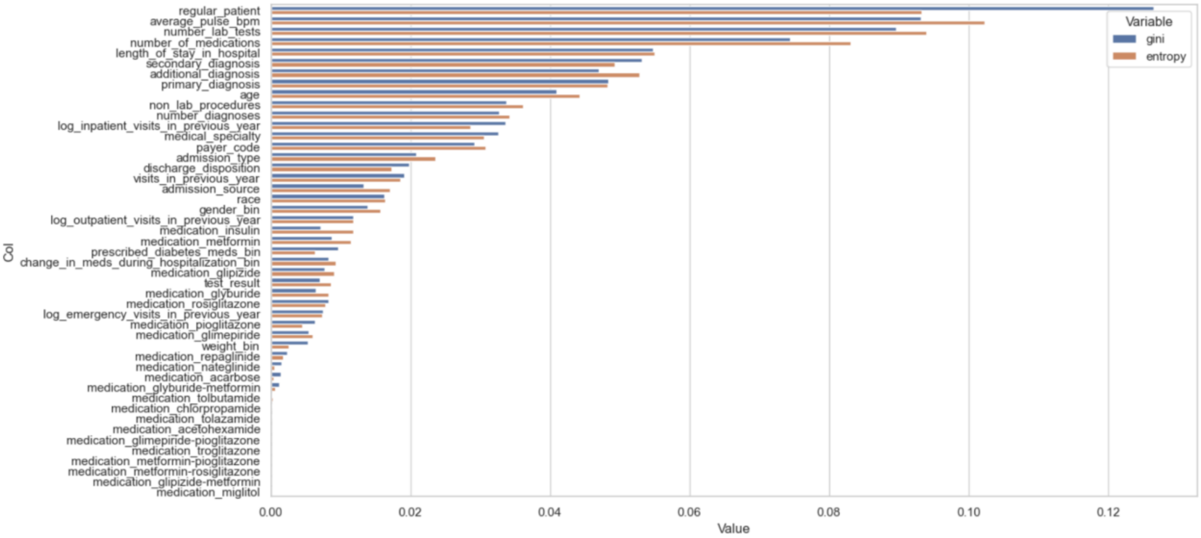
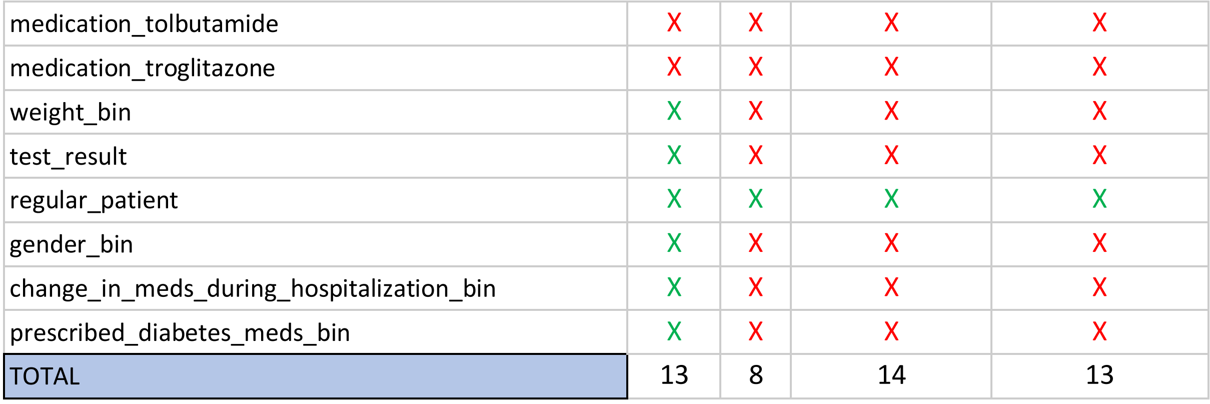


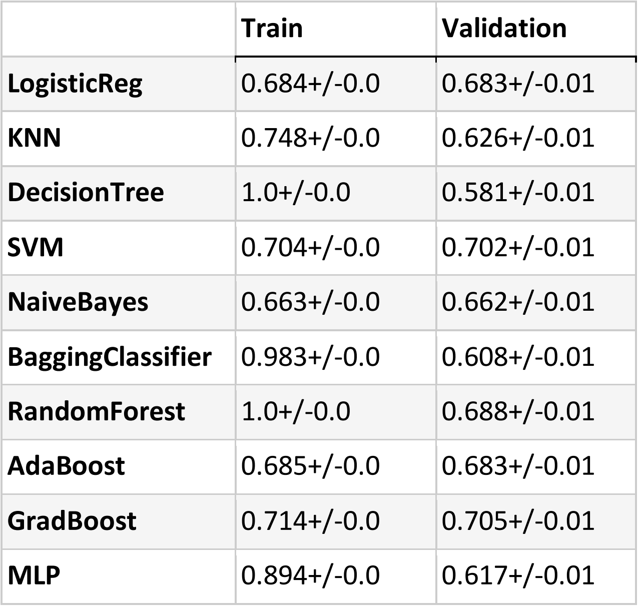
Figure 11 -

	AFTER UNDERSAMPLING			
Features	Lasso	RFE	Decision Tree	Final Selection
race	X	X	X	X
age	X	X	X	X
payer_code	X	X	X	X
log_outpatient_visits_in_previous_year	X	X	X	X
log_emergency_visits_in_previous_year	X	X	X	X
log_inpatient_visits_in_previous_year	X	X	X	X
admission_type	X	X	X	X
medical_specialty	X	X	X	X
average_pulse_bpm	X	X	X	X
discharge_disposition	X	X	X	X
admission_source	X	X	X	X
length_of_stay_in_hospital	X	X	X	X
number_lab_tests	X	X	X	X
non_lab_procedures	X	X	X	X
number_of_medications	X	X	X	X
primary_diagnosis	X	X	X	X
secondary_diagnosis	X	X	X	X
additional_diagnosis	X	X	X	X
number_diagnoses	X	X	X	X
visits_in_previous_year	X	X	X	X
medication_acarbose	X	X	X	X
medication_acetohexamide	X	X	X	X
medication_chlorpropamide	X	X	X	X
medication_glimepiride	X	X	X	X
medication_glimepiride-pioglitazone	X	X	X	X
medication_glipizide	X	X	X	X
medication_glipizide-metformin	X	X	X	X
medication_glyburide	X	X	X	X
medication_glyburide-metformin	X	X	X	X
medication_insulin	X	X	X	X
medication_metformin	X	X	X	X
medication_metformin-pioglitazone	X	X	X	X
medication_metformin-rosiglitazone	X	X	X	X
medication_miglitol	X	X	X	X
medication_nateglinide	X	X	X	X
medication_pioglitazone	X	X	X	X
medication_repaglinide	X	X	X	X
medication_rosiglitazone	X	X	X	X
medication_tolazamide	X	X	X	X



medication_tolbutamide	X	X	X	X
medication_troglitazone	X	X	X	X
weight_bin	X	X	X	X
test_result	X	X	X	X
regular_patient	X	X	X	X
gender_bin	X	X	X	X
change_in_meds_during_hospitalization_bin	X	X	X	X
prescribed_diabetes_meds_bin	X	X	X	X
TOTAL	13	8	14	13

Table 16 -



	Train	**Validation**
LogisticReg	0.684+/-0.0	0.683+/-0.01
KNN	0.748+/-0.0	0.626+/-0.01
DecisionTree	1.0+/-0.0	0.581+/-0.01
SVM	0.704+/-0.0	0.702+/-0.01
NaiveBayes	0.663+/-0.0	0.662+/-0.01
BaggingClassifier	0.983+/-0.0	0.608+/-0.01
RandomForest	1.0+/-0.0	0.688+/-0.01
AdaBoost	0.685+/-0.0	0.683+/-0.01
GradBoost	0.714+/-0.0	0.705+/-0.01
MLP	0.894+/-0.0	0.617+/-0.01

Table 17 -

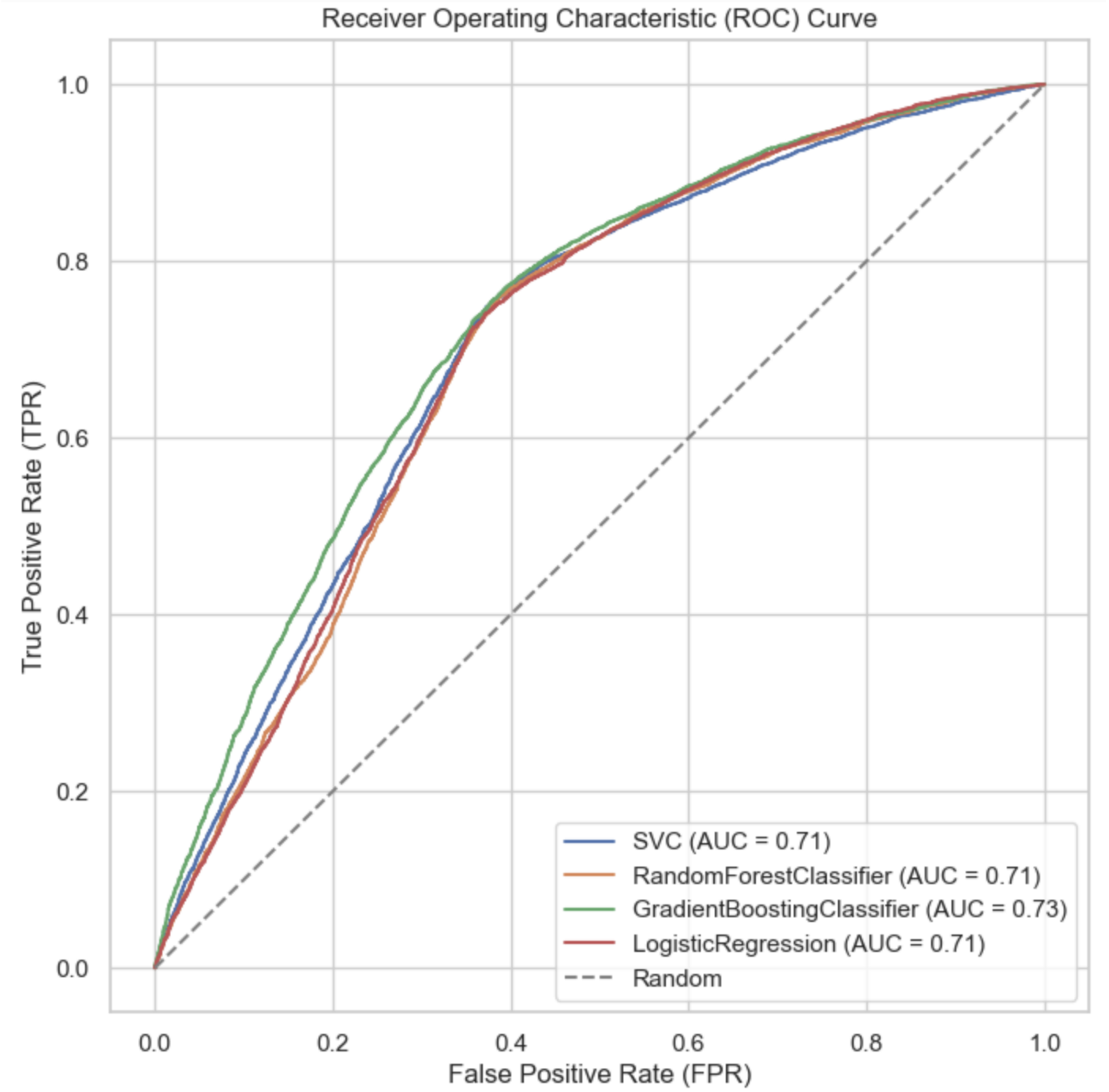
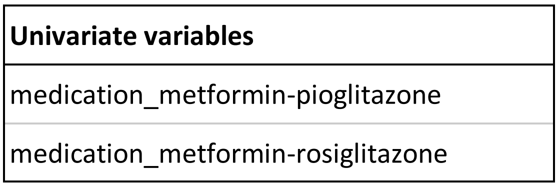


Figure 12 -

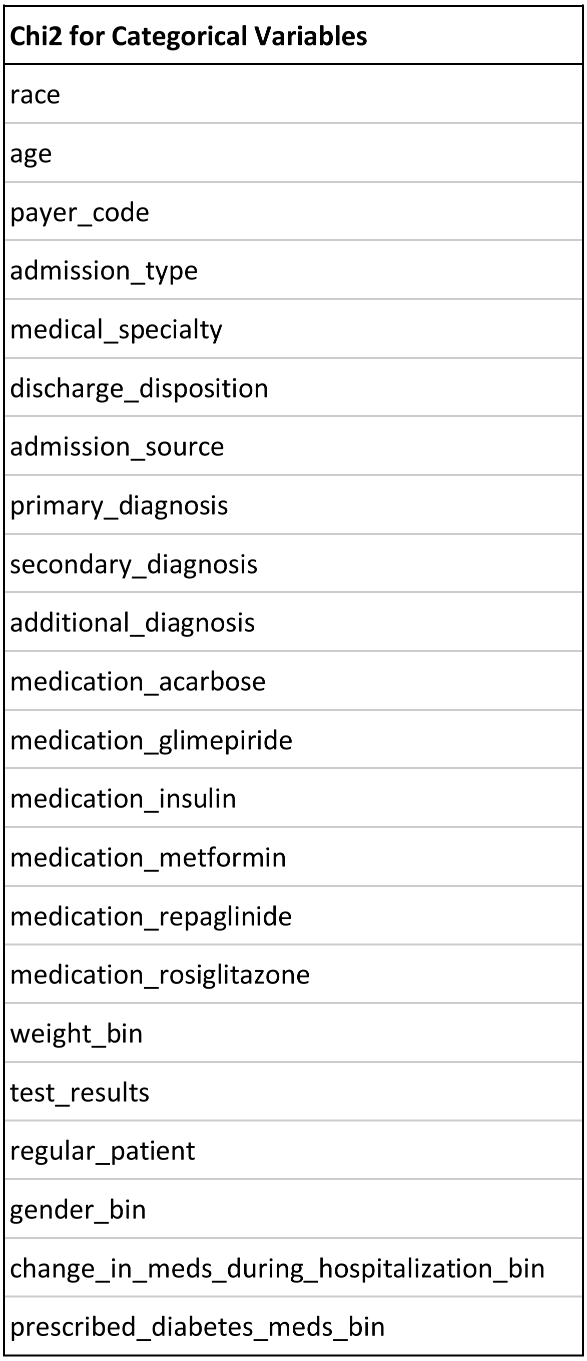
	accuracy	recall	precision	f1
Logistic Regression	0.6765	0.6993	0.6689	0.6837
Gradient Boosting Classifier	0.6866	0.7519	0.665	0.7078
Random Forest Classifier	0.6802	0.7168	0.6679	0.6914
SVC	0.6868	0.7448	0.6674	0.704
Voting Classifier	0.6845	0.7304	0.669	0.6983

Table 18 -



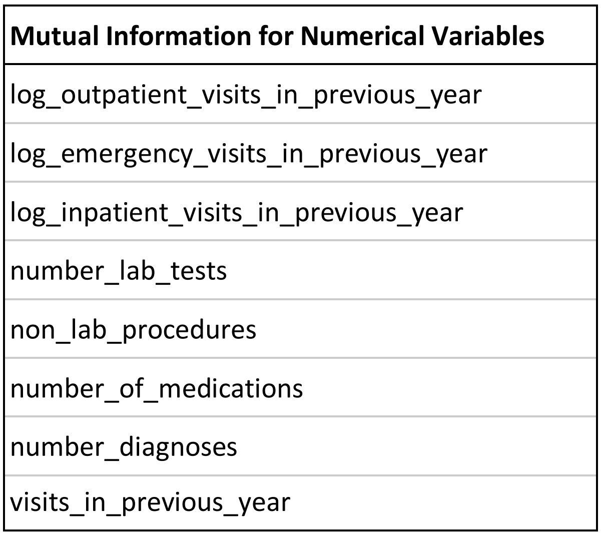
Univariate variables
medication_metformin-pioglitazone
medication_metformin-rosiglitazone

Table 19 -



Chi2 for Categorical Variables
race
age
payer_code
admission_type
medical_specialty
discharge_disposition
admission_source
primary_diagnosis
secondary_diagnosis
additional_diagnosis
medication_acarbose
medication_glimepiride
medication_insulin
medication_metformin
medication_repaglinide
medication_rosiglitazone
weight_bin
test_results
regular_patient
gender_bin
change_in_meds_during_hospitalization_bin
prescribed_diabetes_meds_bin

Table 20 -



Mutual Information for Numerical Variables
log_outpatient_visits_in_previous_year
log_emergency_visits_in_previous_year
log_inpatient_visits_in_previous_year
number_lab_tests
non_lab_procedures
number_of_medications
number_diagnoses
visits_in_previous_year

Table 21 -

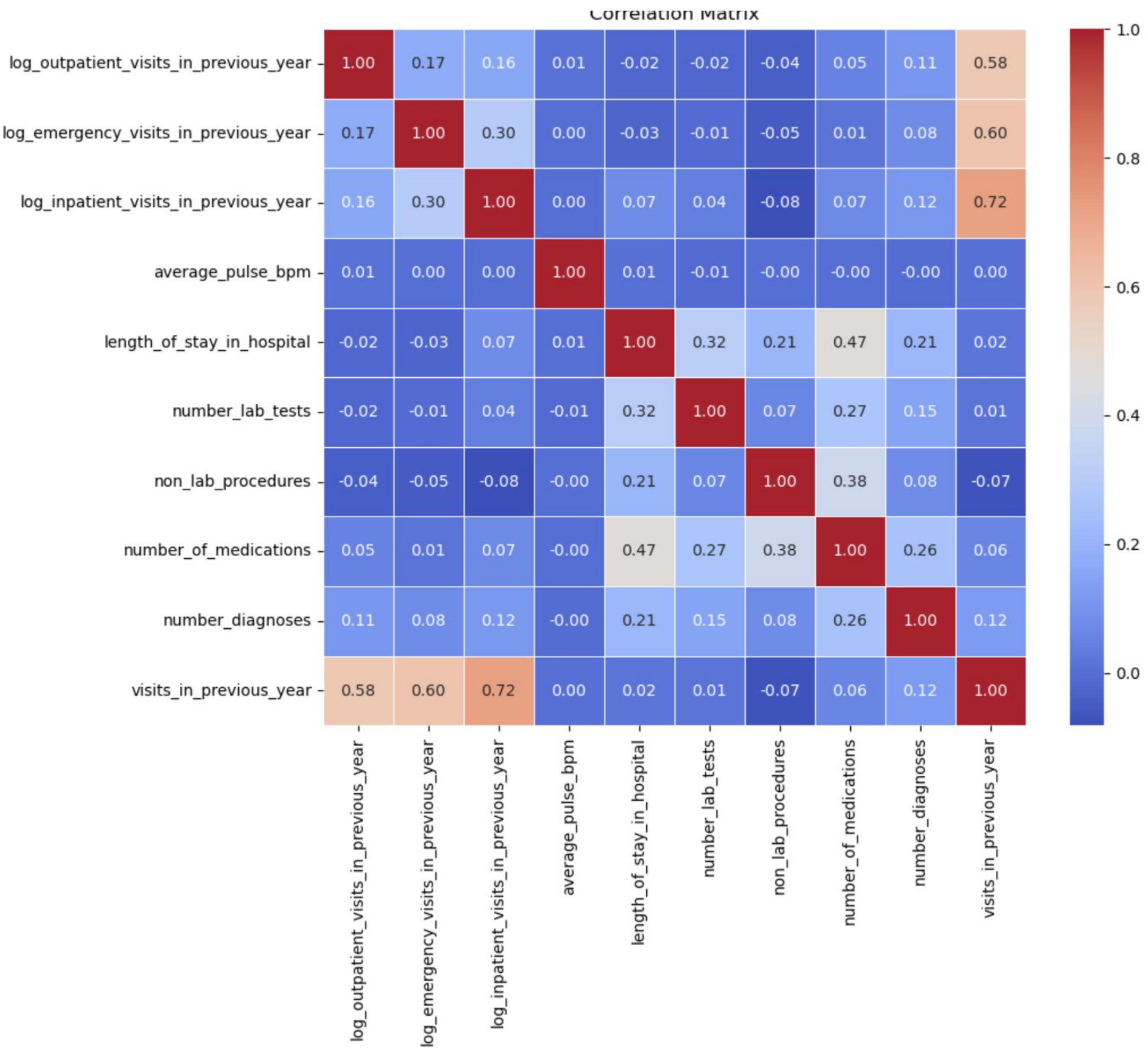
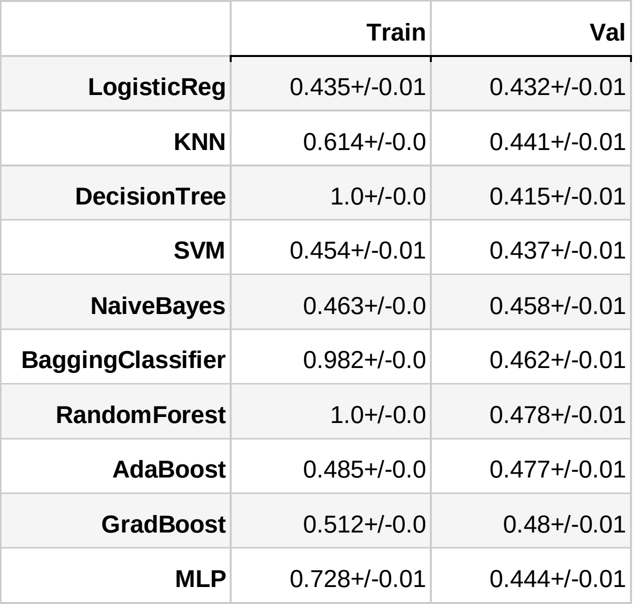


Figure 13 -

	AFTER UNDERSAMPLING			
Features	Lasso	RFE	Decision Tree	Final Selection
race	X	X	X	X
age	X	X	X	X
payer_code	X	X	X	X
log_outpatient_visits_in_previous_year	X	X	X	X
log_emergency_visits_in_previous_year	X	X	X	X
log_inpatient_visits_in_previous_year	X	X	X	X
admission_type	X	X	X	X
medical_specialty	X	X	X	X
average_pulse_bpm	X	X	X	X
discharge_disposition	X	X	X	X
admission_source	X	X	X	X

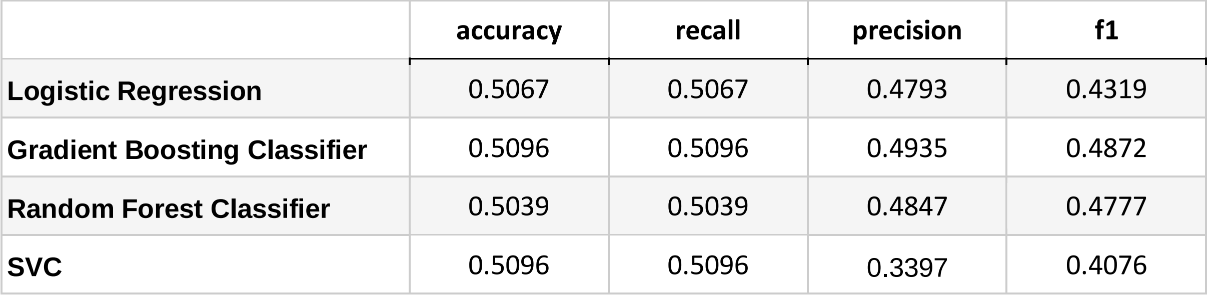
length_of_stay_in_hospital	X	X	X	X
number_lab_tests	X	X	X	X
non_lab_procedures	X	X	X	X
number_of_medications	X	X	X	X
primary_diagnosis	X	X	X	X
secondary_diagnosis	X	X	X	X
additional_diagnosis	X	X	X	X
number_diagnoses	X	X	X	X
visits_in_previous_year	X	X	X	X
medication_acarbose	X	X	X	X
medication_acetohexamide	X	X	X	X
medication_chlorpropamide	X	X	X	X
medication_glimepiride	X	X	X	X
medication_glimepiride-pioglitazone	X	X	X	X
medication_glipizide	X	X	X	X
medication_glipizide-metformin	X	X	X	X
medication_glyburide	X	X	X	X
medication_glyburide-metformin	X	X	X	X
medication_insulin	X	X	X	X
medication_metformin	X	X	X	X
medication_metformin-pioglitazone	X	X	X	X
medication_metformin-rosiglitazone	X	X	X	X
medication_miglitol	X	X	X	X
medication_nateglinide	X	X	X	X
medication_pioglitazone	X	X	X	X
medication_repaglinide	X	X	X	X
medication_rosiglitazone	X	X	X	X
medication_tolazamide	X	X	X	X
medication_tolbutamide	X	X	X	X
medication_troglitazone	X	X	X	X
weight_bin	X	X	X	X
test_result	X	X	X	X
regular_patient	X	X	X	X
gender_bin	X	X	X	X
change_in_meds_during_hospitalization_bin	X	X	X	X
prescribed_diabetes_meds_bin	X	X	X	X
TOTAL	36	8	14	13

Table 22 -



	Train	Val
LogisticReg	0.435+/-0.01	0.432+/-0.01
KNN	0.614+/-0.0	0.441+/-0.01
DecisionTree	1.0+/-0.0	0.415+/-0.01
SVM	0.454+/-0.01	0.437+/-0.01
NaiveBayes	0.463+/-0.0	0.458+/-0.01
BaggingClassifier	0.982+/-0.0	0.462+/-0.01
RandomForest	1.0+/-0.0	0.478+/-0.01
AdaBoost	0.485+/-0.0	0.477+/-0.01
GradBoost	0.512+/-0.0	0.48+/-0.01
MLP	0.728+/-0.01	0.444+/-0.01

Table 23 -



	accuracy	recall	precision	f1
Logistic Regression	0.5067	0.5067	0.4793	0.4319
Gradient Boosting Classifier	0.5096	0.5096	0.4935	0.4872
Random Forest Classifier	0.5039	0.5039	0.4847	0.4777
SVC	0.5096	0.5096	0.3397	0.4076

Table 24 -

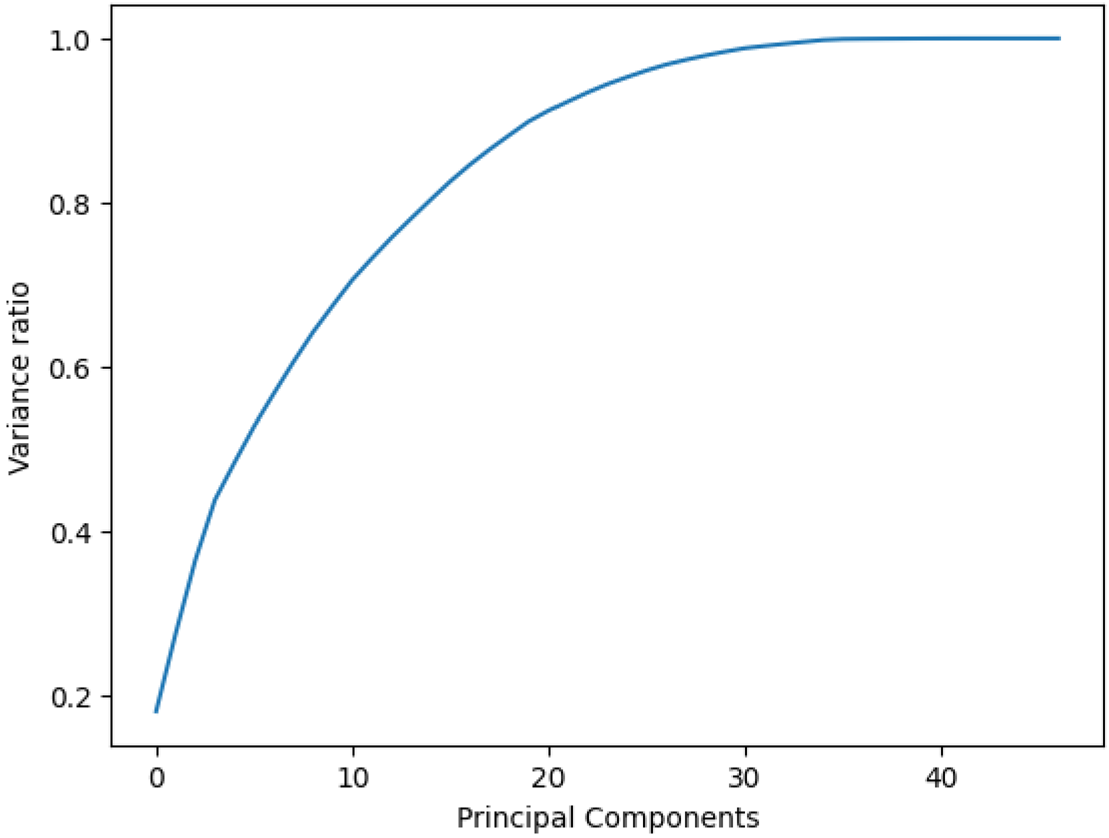
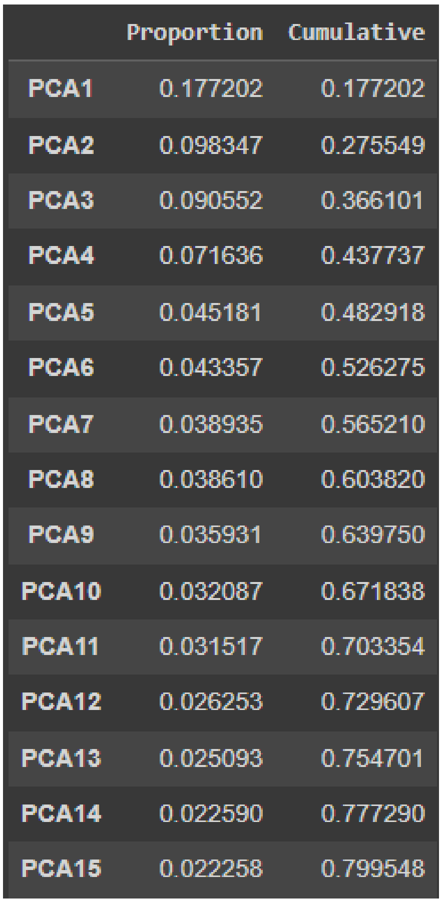


Figure 14 -



	Proportion	Cumulative
PCA1	0.177202	0.177202
PCA2	0.098347	0.275549
PCA3	0.090552	0.366101
PCA4	0.071636	0.437737
PCA5	0.045181	0.482918
PCA6	0.043357	0.526275
PCA7	0.038935	0.565210
PCA8	0.038610	0.603820
PCA9	0.035931	0.639750
PCA10	0.032087	0.671838
PCA11	0.031517	0.703354
PCA12	0.026253	0.729607
PCA13	0.025093	0.754701
PCA14	0.022590	0.777290
PCA15	0.022258	0.799548

Figure 15 -