Predicting hospital readmission of patients with diabetes

Son Nguyen 658463601
Vanisa Achakulvisut 667903568
Rongchuan Guo 658216522

IDS 575: Machine Learning Statistics
UIC Business Liautaud Graduate School
University of Illinois at Chicago

December 8, 2020

Abstract

This project aims to develop the different machine learning algorithms using K Nearest Neighbor, Naïve Bayes, Elastic Net, and Support Vector Machines algorithms to predict readmission. Data preprocessing and feature engineering have been performed before building the models using 10-fold Cross-Validation to avoid overfitting problems. The hyperparameters are tuned by the grid as well as random searching. Additionally, different approaches Moving Threshold, Downsampling, and ROSE algorithms are utilized to solve the imbalanced dataset problem. Different evaluation metrics from AUC, $F_1 Score$, or False Negative Rate are used to evaluate the performance of different models. K-nearest Neighbor models perform the worst and downsampling Radial Kernel Support Vector Machine performs the best with the highest AUCs on both resampling and test.

1 Introduction

Hospital readmission is a state where a patient is admitted to a hospital in less than 30 days after being discharged from an earlier hospital stayand diabetes remain one of the greatest risk factors for increased 30-day readmissions [1]. It is costly and leads to unfavorable patient outcomes. Reducing Hospital readmissions has long been the goal of hospitals and the healthcare industry because it represents an opportunity to lower costs, improve healthcare quality, and increase patient satisfaction.

In this project, different machine learning models are developed using K-Nearest Neighbor, Naïve Bayes, Elastic Net Regression, and Support Vector Machines algorithms to identify readmission with great performance. In order to do so, we will use patient's information such as race, gender, age, weight, and medical records like the number of hospital visits, blood-tests and medication as our attributes. The goal is to prevent unnecessary hospital readmissions among diabetic patients.

2 Materials and Methods

2.1 Dataset

The data was collected on the Diabetes 130 US hospitals for years 1999-2008 Kaggle page which was published by Humberto Brandão in 2017 [2]. It was made available using the Health Facts database from the Cerner Corporation, a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. The data contains 101,766 obs. of 50 different attributes such as patient number, race, gender, age, admission type, time in the hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatients, inpatients, and emergency visits in the year before the hospitalization of patients from 1999 to 2008. The target is readmission status with three levels: greater than 30 days, less than 30 days, and non-readmitted. Only less than 30 days readmissions are interested and they are rare events. Later on, appropriate steps are required to resolve the problem of an imbalanced dataset. The original dataset has 101,766 obs. of 50 variables with the information of types, descriptions, values, and percentages of missing data is provided in Appendix 1.

2.2 Data Preprocessing and Feature Engineering

The raw data consists of inconsistent or redundant records with incomplete and high dimensionality attributes which make it difficult for building prediction models. Thus, it is treated appropriately to transform it into a suitable format.
The target variable indicating readmission is recoded into two classes: readmitted for less than 30 days readmission and non-readmitted for greater than 30 days and non-readmitted. To make sure each instance is statistically independent, only the first encounter per patient is considered. Later admissions of the same patients will be eliminated. Identifiers, one-value, and unbalanced predictors are omitted. Secondary diagnoses will also be dropped since they are additional diagnoses to diagnosis 1.
Incomplete columns is treated based on the missing proportions and the relevance of them.

Table 1: Proportion of NA

Proportion of NA	Feature name
More than 50%	weight
More than 30%	payer_code, medical_specialty
More than 0%	race, weight, payer_code, medical_specialty, diag_1, diag_2, diag_3

Attributes with many categories are collapsed into smaller and manageable classes as in the following tables:

Table 2: Admission Type ID Categories

Admission Type ID	Description	Assigned Category
1	Emergency	Emergency
2, 4, 7	Urgent, Newborn, Trauma Center	Urgent
3	Elective	Elective
5, 6, 8	Not Available, NULL, Not Mapped	NotAvailable

Table 3: Discharge Disposition ID

Discharge Disposition ID	Description	Assigned Category
1, 6, 8, 12, 16, 17	- Discharged to home - Discharged/transferred to home with home health service - Discharged/transferred to home under care of Home IV provider - Still patient or expected to return for outpatient services - Discharged/transferred/referred another institution for outpatient services - Discharged/transferred/referred to this institution for outpatient services	Home
2, 5, 23, 30, 27, 28, 29	- Discharged/transferred to another short term hospital - Discharged/transferred to another type of inpatient care institution - Discharged/transferred to a long term care hospital - Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere - Discharged/transferred to a federal health care facility - Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital - Discharged/transferred to a Critical Access Hospital (CAH)	Another Hospitals
3, 4, 10, 22, 24	- Discharged/transferred to SNF - Discharged/transferred to ICF - Neonate discharged to another hospital for neonatal aftercare - Discharged/transferred to another rehab fac including rehab units of a hospital - Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.	Rehab Facilities
4, 18, 25, 26	- Not Available - NULL - Not Mapped - Unknown/Invalid	Not Available

Table 4: Admission Source Categories

Admission Source ID	Description	Assigned Category
1, 2, 3	- Physician Referral - Clinic Referral - HMO Referral	Referral
4, 5, 6, 10, 18, 19, 22, 25	- Transfer from a hospital - Transfer from a Skilled Nursing Facility (SNF) - Transfer from another health care facility - Transfer from critical access hospital - Transfer From Another Home Health Agency - Readmission to Same Home Health Agency - Transfer from hospital inpt/ same fac result in a sep claim - Transfer from Ambulatory Surgery Center	AnotherHospitals
7, 12, 23, 24, 26	- Emergency Room - Premature Delivery - Born inside this hospital - Born outside this hospital - Transfer from Hospice	EmergencyRoom
8, 9, 11, 13, 14, 15, 17, 20, 21	- Normal Delivery - Sick Baby - Extramural Birth - Not Available, NULL, Not Mapped, Unknown/Invalid	Others/ NotAvailable

Table 5: ICD9 Categories

idc 9	Description	Assigned Category
1,2,3	- Physician Referral - Clinic Referral - HMO Referral	Referral
390–459, 785	Diseases of the circulatory system	Circulatory
460:519,786	Diseases of the respiratory system	Respiratory
520:579,787	Diseases of the digestive system	Digestive
250	Diabetes	Diabetes
800:999	Injury and poisoning	Injury
710:739	Diseases of the musculoskeletal system and connective tissue	Musculoskeletal
580:629,788	Diseases of the genitourinary system	Genitourinary
140:239	Neoplasms	Neoplasms
780, 781, 784, 783, 789, 790:799, 240:279, 680:709, 782, 001:139, 290:319, 280:289, 320:359, 630:679, 360:389, 740:759, "E", "V"	Other 3-digits codes except from no. 1-8 and "E","V"	Other

Table 6: Medical Specialty Categories

Medical Specialty	Assigned Category
Physician Not Found, Not Available, Cardiology , Cardiology-Pediatric, Cardiology, Surgeon, Surgery-Cardiovascular, Surgery-Cardiovascular/Thoracic, Surgery-Colon Rectal, Surgery-General, Surgery-Maxillofacial, Surgery-Neuro, Surgery-Pediatric, Surgery-Plastic, Surgery-Plastic with Head and Neck, -Thoracic, Surgery-Vascular, SurgicalSpecialty	Surgery
Orthopedics, Orthopedics - Reconstructive, Orthopedics,	Orthopedics
Obstetrics Gynecology-Gynecologiconco, Obstetrics, Obstetrics and Gynecology	Obstetrics
Pediatrics, Pediatrics-CriticalCare, Pediatrics-EmergencyMedicine, Pediatrics-Endocrinology, Pediatrics-Hematology-Oncology, Pediatrics-Neurology, Pediatrics-Pulmonology	Pediatrics
Psychiatry, Psychiatry-Addictive, Psychiatry-Child/Adolescent	Psychiatry
Radiologist, Radiology, Radiology, Anesthesiology, Anesthesiology-Pediatric	Anesthesiology
Resident, Family/GeneralPractice,	Family/GeneralPractice
Group smaller groups into "others": NotAvailable, InternalMedicine ,Family/GeneralPractice, Emergency/Trauma, Cardiology, Surgery, Orthopedics	Others

The variable comprising a finite set of discrete values with a ranked ordering between values is re-assigned into ordinal variables.

Table 7: Age Categories

Age	Assigned Category
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 8: Glucose serum test Categories

Glucose serum test	Assigned Category
>300	300
>200	200
Norm	100
None	0

Table 9: A1C test Categories

A1C test	Assigned Category
>8	8
>7	7
Norm	100
None	0

2.3 Exploratory Data Analysis

To explore the main characteristics of the dataset, data visualization is used.

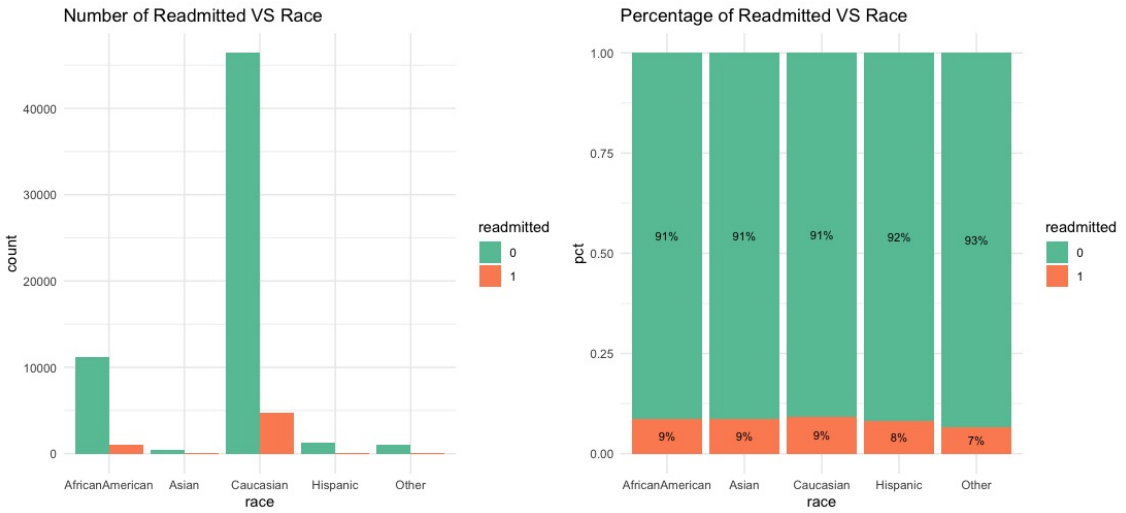

Figure 1: Percentage of Readmitted VS Race

The left-hand chart indicates the number of the readmitted of each race. The number of readmitted Caucasian is the highest, while the number of readmitted other races is relatively low. The right-hand chart shows the percentage of readmitted in each race. It outlines that the proportions of readmitted of African, Asian, Caucasian, and Hispanic are approximately the same at 8-9%, while the other group and not available group are at 7%.

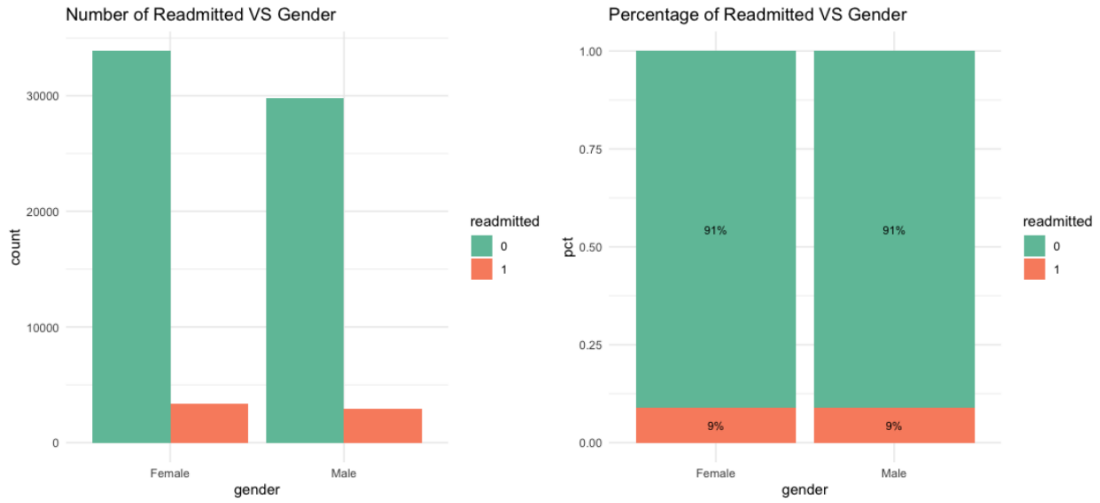

Figure 2: Percentage of Readmitted VS Gender

As can be seen from the chart, for the observed number of readmitted by gender, there are more female diabetic patients compared to male patients. However, the readmission rates are similar for

both genders.

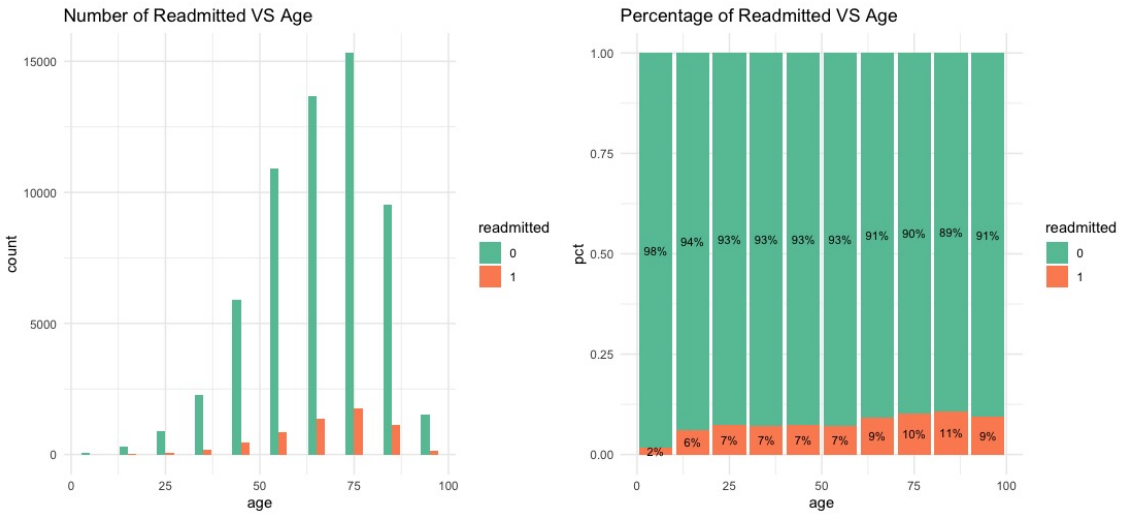

Figure 3: Percentage of Readmitted VS Age

The number of patients increases as the age gets higher. Older and elderly people are more likely to be readmitted to the hospital compared to younger patients. The percentage of readmitted diagrams demonstrates that 10% of elderly patients within the age range of 60 -100 tend to be readmitted to the hospital. 7% of patients ages between 20 to 60 are likely to return to the hospital and only 4% of patients younger than 20 years old are likely to return to the hospital.

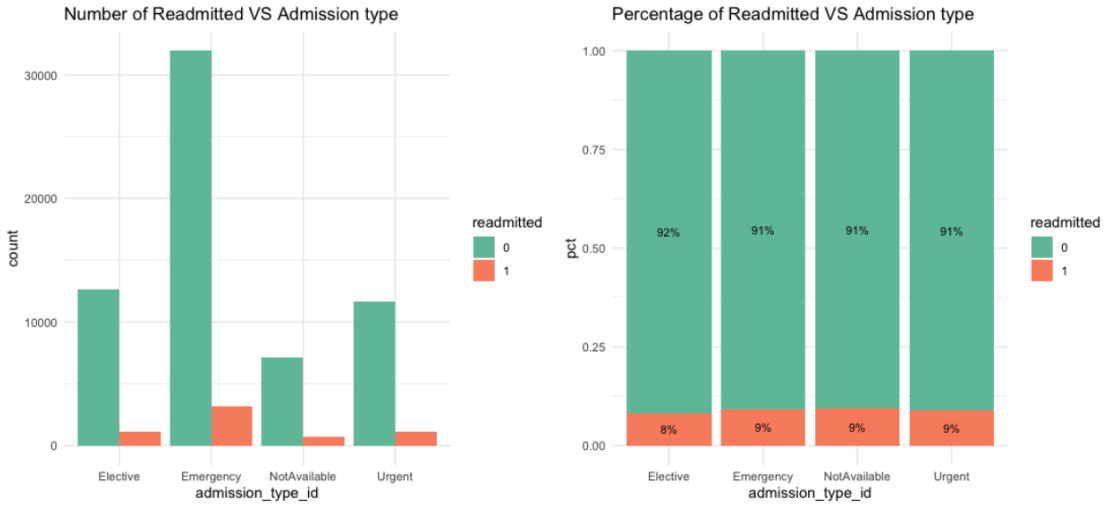

Figure 4: Percentage of Readmitted VS Admission type

There is not much difference in percentages of readmitted in each admission type. The rates are similar at around 9%.

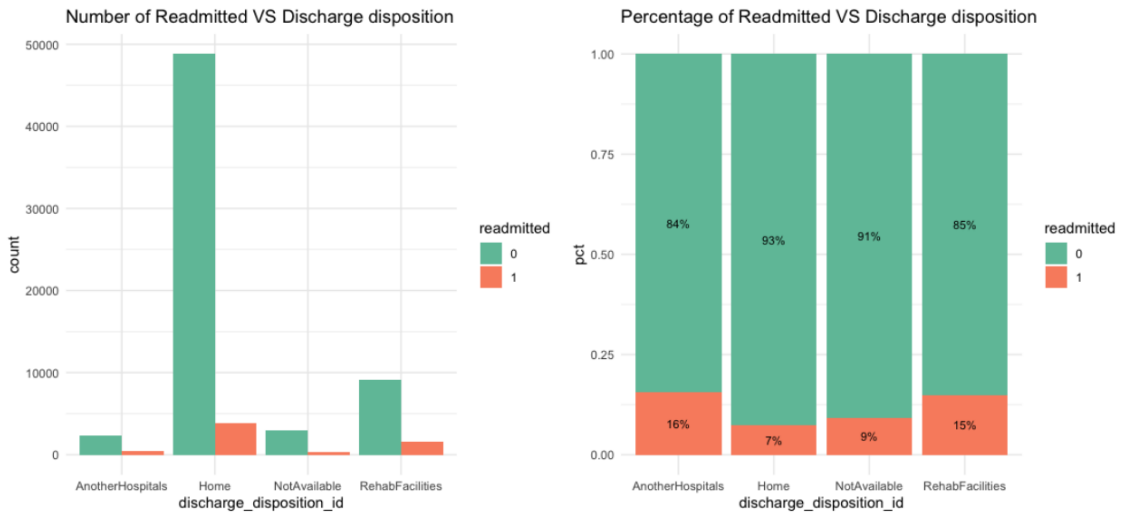

Figure 5: Percentage of Readmitted VS Discharged disposition

Most of the patients in the observation are discharged to their homes. Fewer patients get discharged to rehab facilities and other hospitals presented in the observation. Patients who are discharged to home have the lowest percentage of readmitted at 7%. Whereas, the patients discharged to the rehab facilities and another hospital have twice the readmitted rate around 15% compared to patients discharged to home.

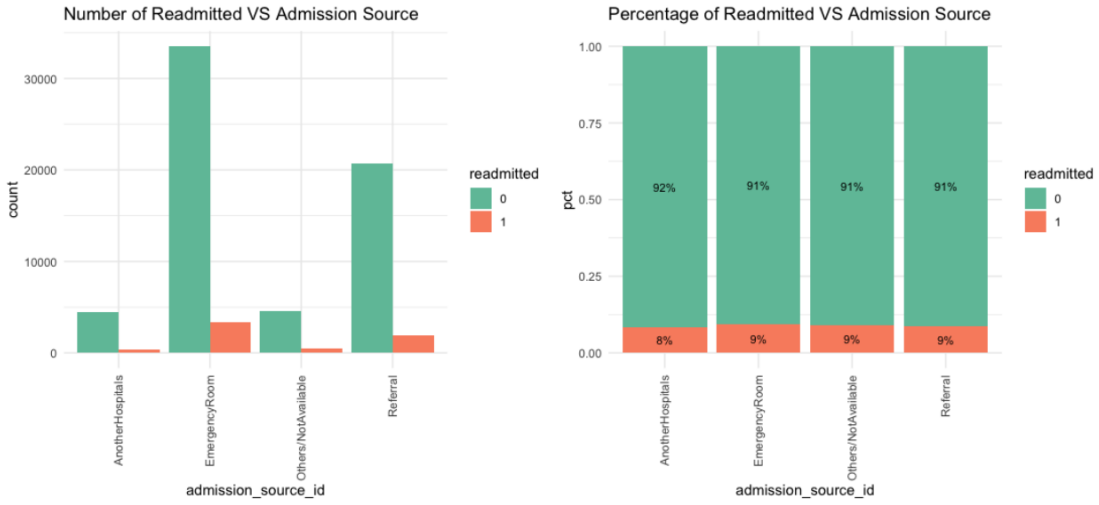

Figure 6: Percentage of Readmitted VS Admission source id

The percentages are approximately 8-9% which are very similar to each type of admission source id.

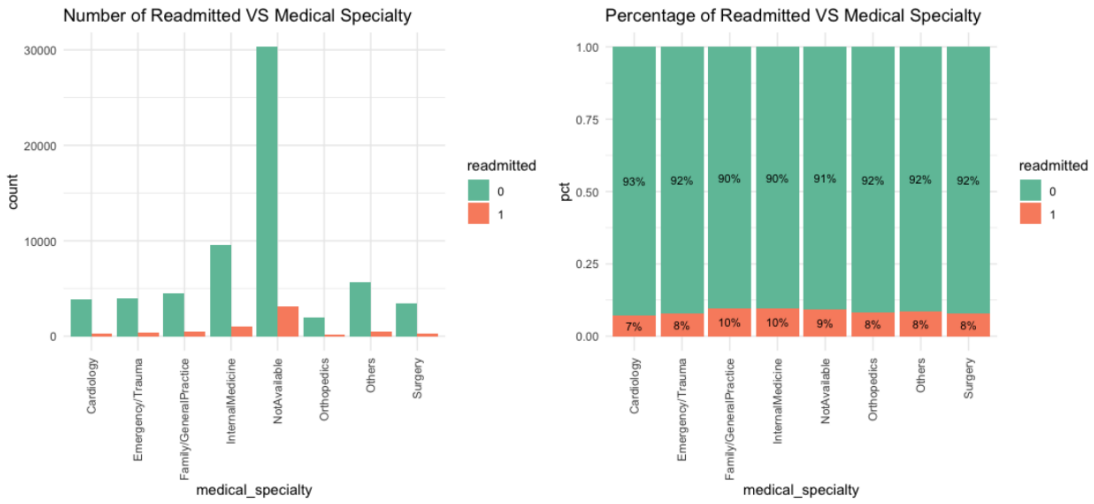

Figure 7: Percentage of Readmitted VS Medical Specialty

The least percentage of readmitted is at 7% presented in the Cardiology group. The second-lowest is at 8% presented in Emergency/Trauma, Surgery and other groups. The Family/ General practice and Internal Medicine are at 10% which are likely to be the group that will be readmitted in the future. Departments that are specialized are able to achieve lower readmitted rates than ones that are general and not available.

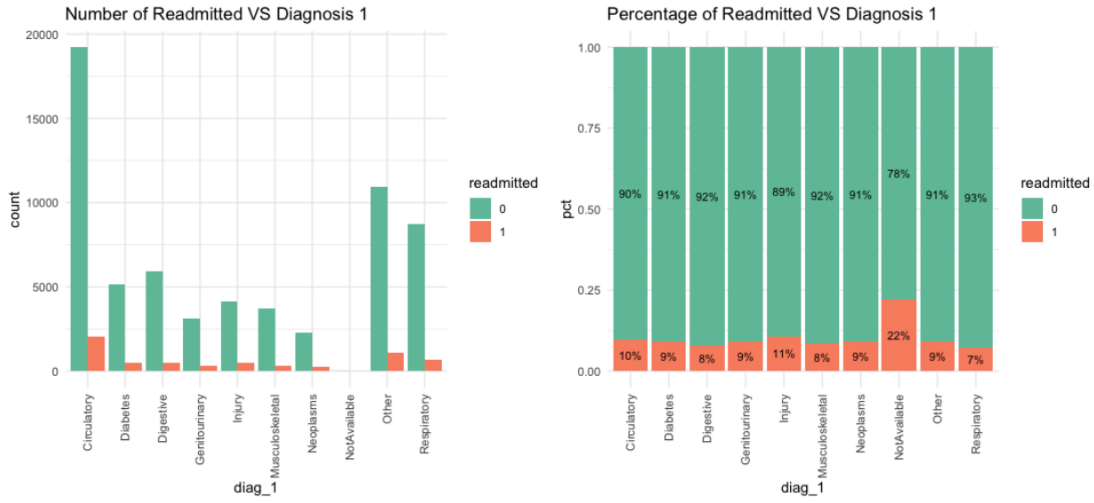

Figure 8: Percentage of Readmitted VS Diagnosis 1

22% of readmitted presents in the NotAvailable data, this information could be negligible because the actual number of readmitted is relatively low compared to other categories. Injury and circulatory have higher readmitted rates at 11% and 10% separately. Respiratory, digestive, and musculoskeletal have the lowest readmitted rates.

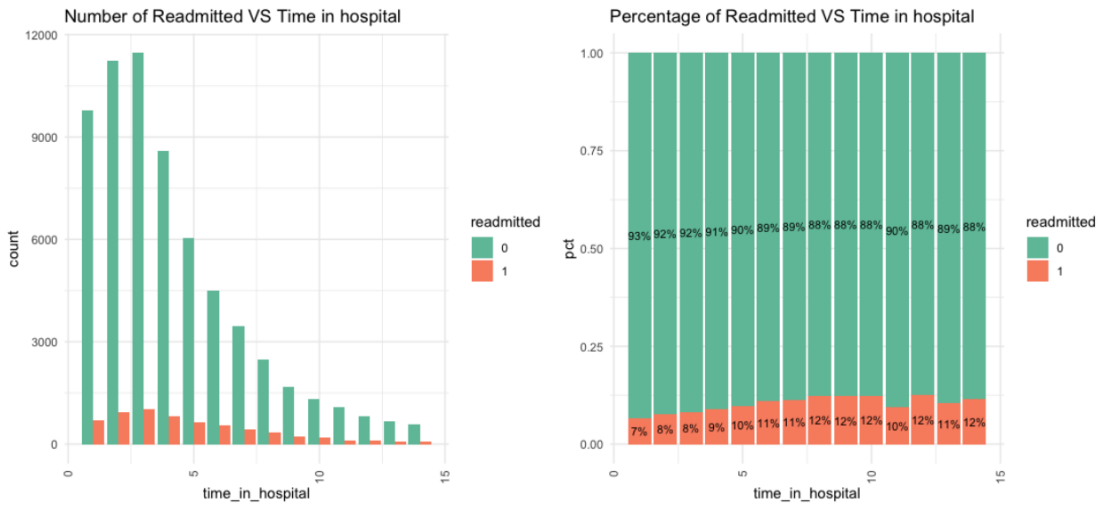

Figure 9: Percentage of Readmitted VS Age

The more days the patients stay in the hospital, the higher chance they will get readmitted to the hospital. The percentage of readmitted of the patients staying in the hospital for less than 4 days is up to 9%. On the contrary, the percentage of readmitted of the patients staying in the hospital for over 4 days is more than 10%.

3 Modeling

3.1 Train-Test Split

The size of dataset after cleaning is 67,598 observations of 29 variables. The dataset is then splitting so that 70% of it is what the models are trained from and 30% is unseen data for them to evaluate.

3.2 Feature Selection

Feature selection using Random Forest are used on the training data to reduce the number of input variables to only ones that are most useful to a model so training can be done faster.

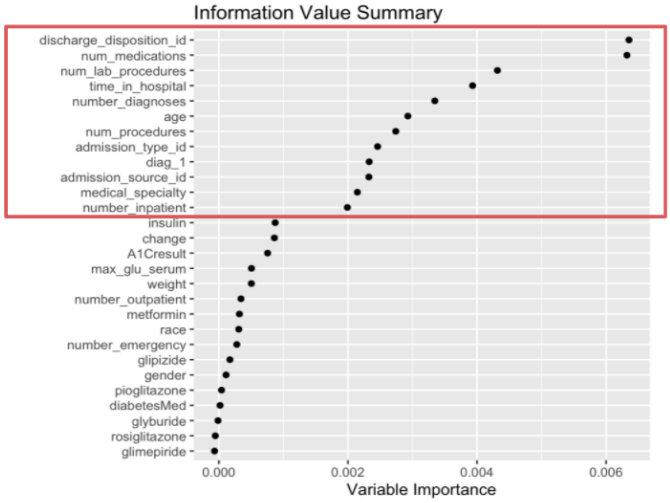

Figure 10: Variable importance information summary

Using Random Forest algorithm, it can compute how much each feature decreases the impurity. The more a feature decreases the impurity, the more important the feature is. In random forests, the impurity decrease from each feature can be averaged across trees to determine the final importance of the variable.
After performing the random forest, the variable importance is shown in Fig.11 which "age, admission_type_id, discharge_disposition_id, admission_source_id, time_in_hospital, medical_specialty, num_lab_procedures, num_procedures, num_medications, diag_1, number_diagnoses" are interested variables in this project.

3.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a supervised machine learning algorithm that classifies each new observation based on how similar to its neighboring data points. The similarity is measured using distance metrics, and Euclidean distance is the most common way to measure it. The equation for Euclidean distance is as below [3].

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{1}$$

It is a non-parametric and lazy learning algorithm because it does not have any assumption of data distribution and it must be run at prediction time. It uses Euclidean distance: therefore, normalization is strongly recommended. Also, KNN is extremely sensitive to the dimensionality of data and outliers. It is not suitable for a large number of features and outliers should be treated.

3.4 Naïve Bayes

The Naïve Bayes is a probabilistic classifier based on Bayes theorem [4].

$$P(C|X) = \frac{P(C|X)P(C)}{P(X)} \tag{2}$$

P(C) is called class prior probability. It is the probability of the class C historically.
P(X) is called predictor prior probability. It is the probability of the predictor variables historically.
P(X|C) is called likelihood or conditional probability, which is the probability of X occurring under the premise that C has occurred. Basically, it is the probability of seeing the predictor values for the response variable's class.
P(C|X) is called posterior probability. After observing information mentioned above, we can get the posterior probability of an observation having the class C.

Additionally, the assumption of Naïve Bayes classifier is that the predictor variables are conditionally independent from one another when the response value is given. This assumption simplifies computation: thus, it is named as Naïve. Based on the assumption, the calculation can be simplified as follow [4]:

$$P(C|X) = \prod_{i=1}^{n} P(X|C) \tag{3}$$

The posterior probability P(C|X) is simplified as the product of the probability distribution for each variable with the condition on the response category.

Naïve Bayes is not only fast, scales well with a large amount of data, but also accurate [4]. however, it has the problem of zero probability, the situation of unobserved events, which can be fixed using the Laplace Smoothing technique. Also, gaussian density estimation or kernel density estimation can be used to approximate the true probability for continuous variables.

3.5 Elastic-net Logistic Regression

Logistic regression works similar to linear regression, but the outcome is the binary response. It turns linear regression to work with dichotomous data by passing it through a Sigmoid function.

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}} \tag{4}$$

The loss function for logistic regression is [5]:

$$L_{log} = -ln(L) = -\sum_{i=1}^{n} [-ln(1 + e^{(\beta_0 + \beta_i x_i)}) + y_i(\beta_0 + \beta_i x_i)] \tag{5}$$

The problem of Logistic Regression is it is very flexible if there are too many features, and this leads to overfitting. Therefore, regularization is very important for logistic regression to dampen model complexity. Most logistic regression models use L1 regularization or Lasso; and, L2 regularization or Ridge. λ is the regularization penalty term that penalizes the effect of coefficients in Ridge

regression and Lasso regression. Therefore, Lasso regression which has absolute term is able to penalize towards zero. In contrast to Lasso regression, Ridge regression is able to penalize the cost function to nearly zero.
Lasso regression loss function is defined as follows [5]:

$$J = L_{log} + \lambda \sum_{j=1}^{m} |\beta_j| \tag{6}$$

Ridge regression loss function is defined as follows [5]:

$$J = L_{log} + \lambda \sum_{j=1}^{m} \beta_j^2 \tag{7}$$

Elastic Net is the combination of Lasso regression and Ridge regression. α is the mixing parameter between ridge (α=0) and lasso (α=1). Elastic Net regression loss function is defined as follows [5]:

$$J = L_{log} + \lambda \sum_{j=1}^{m} (\alpha \beta_j^2 + (1 - \alpha)|\beta_j|) \tag{8}$$

3.6 Support Vector Machine

Support Vector Machines (SVM) is a binary classification model. The target is to separate data points using a hyperplane with maximizing the margin.

For Non Linear SVM, the primal function is [6]:

$$L_p = min_{w,b}\frac{1}{2}w \cdot w - C\sum_{j=1}^{m}\xi_j \tag{9}$$

$$(w \cdot x_j + b)y_j \geq 1 - \xi_j, \forall_j$$
$$\xi_j \geq 0, \forall_j$$

Then solve for w, b, α:

$$w = \textstyle\sum_{i=1}^{n} \alpha_i y_i x_j$$
$$\mathrm{b} = \mathrm{y}_k - w \cdot x_k$$

Therefore, the dual function is [6]:

$$L_d = max_\alpha \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n,m}\alpha_i\alpha_j y_i y_j (x_i, x_j) \tag{10}$$

$$\textstyle\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\mathrm{C} \geq a_i \geq 0$$

Different Kernels can also be used to help for computation which otherwise would involve computations in higher dimensional space. The dual then now is [6]:

$$L_d = max_\alpha \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n,m}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \tag{11}$$

$$\textstyle\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\mathrm{C} \geq a_i \geq 0$$

Where kernel such as gaussian radial basis function is used :

$$K(x_i, x_j) = exp\frac{\|x_i - x_j\|^2}{2\sigma^2} \tag{12}$$

3.7 Evaluation Metrics

The result is evaluated by a confusion matrix as illustrated in the table as follow:

Table 10: Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

According to the confusion matrix,
TN is the number of negative samples correctly classified.
FP is the number of negative samples incorrectly classified as positive.
FN is the number of positive samples incorrectly classified as negative.
TP is the number of positive samples correctly. classified.

The overall accuracy is obtained by

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

True Negative Rate (TNR) is obtained by

$$TNR = \frac{(TN)}{(N)} = \frac{(TN)}{(TN+FP)}$$

True Positive Rate (TPR) or Recall is obtained by

$$TPR = Recall = \frac{(TP)}{(P)} = \frac{(TP)}{(TP+FN)}$$

False Negative Rate (FNR) is obtained by

$$FNR = \frac{(FN)}{(P)} = \frac{(FN)}{(FN+TP)}$$

False Positive Rate (FPR) is obtained by

$$FPR = \frac{(FP)}{(N)} = \frac{(FP)}{(FP+TN)}$$

Precision is obtained by

$$Precision = \frac{(TP)}{(TP+FP)}$$

F_1 is the harmonic mean of precision and recall and F_1 is obtained by

$$F_1 = 2 \cdot \frac{(precision \cdot recall)}{(precision+recall)} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

The ROC curve can be plotted using TPR and FPR at different thresholds. AUC measures the area underneath the ROC curve. AUC and $F_1 Score$ are used as our main measurements to evaluate the models for this project.

3.8 Baseline Model

First, the baseline model is developed using K-Nearest Neighbor with default hyperparameters using Euclidean distance and k = 7 to obtain the performance reference so more complex solutions can be developed and improved on.
The model achieves great accuracy of 0.887 while the $F_1 Score$ is only 0.067021 at the default threshold of 0.5. This is caused by the problem of the imbalanced dataset as discussed above.

Table 11: Confusion Matrix

	Yes	No
Yes	82	466
No	1817	17914

3.9 Imbalanced Dataset Solution

Two approaches are presented to deal with the problem. First, the models can be trained using the imbalanced training set and the best threshold can be chosen. The other approach is to subsampling using Downsampling and ROSE algorithms within each fold of the 10-fold Cross-Validation of the imbalanced training set [7]. Upsampling is not used in this case because of computational constraints.

3.10 Preprocessing

The final preprocessing part is done in this stage within each fold of the 10-fold Cross-Validation. Yeo-Johnson Transformation, Centering, and Scaling will be applied at each training set within the 10-fold cross-validation for K-Nearest Neighbor, Naïve Bayes, and Support Vector Machine. For Elastic-net Logistic Regression, only Yeo-Johnson Transformation is performed.

3.11 Hyperparameters Tuning

Each of the models will be trained with 10-fold Cross-Validation to optimized model hyperparameters and to prevent overfitting. Random searching as well as grid search are used for this part. As mentioned above, the final steps of preprocessing and subsampling are performed within the cross-validation. Here are the optimized parameters based on the AUC score of the models. Imbalanced and ROSE SVM are not developed because of the size of the dataset. Only Downsampling is used for SVM to reduce processing time.

Table 12: Opitmal Hyperparameters

Model	KNN	Naive Bayes	Elastic Net	SVM
Imbalanced	Best kernel: gaussian Best k: 27	fL: 0 Usekernel: TRUE Adjust: 1	alpha: 0.1 lambda: 0.009703	N/A
Downsampling	Best kernel: gaussian Best k: 50	fL: 2 Usekernel: TRUE Adjust: 6	alpha: 1 lambda: 0.03617	sigma: 0.009785 C: 1.139
ROSE	Best kernel: gaussian Best k: 18	fL: 11 Usekernel: TRUE Adjust: 4	alpha: 0.4623 lambda: 0.001012	N/A

4 Model Evaluation

In Table 13, the AUC results of different models are presented.

Table 13: Resampling and Test AUCs

Models	Resampling	Test
Baseline Model	N/A	0.5516
original_knn	0.5942	0.6006
ROSE_knn	0.5879	0.6032
down_knn	0.6064	0.6237
original_nb	0.6192	0.6444
ROSE_nb	0.6262	0.6478
down_nb	0.6252	0.6487
original_glmnet	0.6310	0.6548
ROSE_glmnet	0.4496	0.6022
down_glmnet	0.6300	0.6480
down_radialsvm	0.6314	0.6639

Resampling AUCs and Test AUCs of the models are comparable meaning that they do not have an overfitting problem. Overall, K-nearest Neighbor models perform the worst. However, the hyperparameter optimized models improve compared to the baseline models. Downsampling KNN performs the best out of KNN models with the Resampling AUC of 0.606 and Test AUC of 0.624. Naive Bayes models, Elastic Net Logistic Regression, and Support Vector Machine models perform closely to each other. Naive Bayes models have slightly lower performance compared to Elastic Net Logistic Regression on resampling but perform well on the test. Finally, Downsampling Radial Kernel Support Vector Machine Performs the best with the highest Resampling AUCs on the train as well as test.

Predictions of the best model Downsampling Radial Kernel Support Vector Machine is obtained in the figure below at threshold 0.5.

Table 14: Downsampling Radial Kernel SVM Confusion Matrix on Test

Metric	Observations	Rate	Percentage Total Observations
Correct	14046	0.6926	0.6926
Missclassified	6233	0.3074	0.3074
True Positive	951	0.5008	0.0469
True Negative	13095	0.7125	0.6457
False Negative	948	0.4992	0.0468
False Positive	5285	0.2875	0.2606

Table 15: $F_1 Score$ Summary

Metric	Value
Precision	0.1525
Recall	0.5008
$F_1 Score$	0.2338

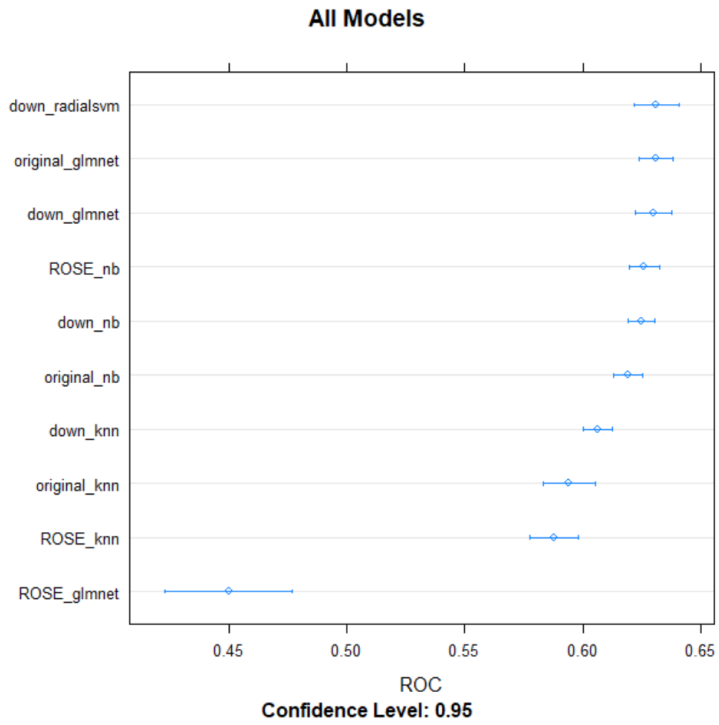

Figure 11: Resampling AUCs Comparison

The accuracy of the model at threshold 0.5 is 69%. $F_1 Score$ is the highest at 0.2338 compare to all of the models developed. However, the false negative rate is at 50% which is still very high. That means if one would like to predict who would be readmitted, he or she would be misclassified 50% at a time. To improve this, another threshold can also be set. There are different ways that one can obtain this threshold, such as using domain knowledge of what is the False Positive Rate is allowed to sacrifice to improve the False Negative Rate. In this project, the False Positive Rate chosen is 35%, meaning that this model would misclassify 3 to 4 non-readmitted patients out of 10 readmitted patients. By looking at different thresholds of the training set's evaluation metrics, we can see that the threshold of 0.4650 give us the False Positive Rate of 0.35264122 while the False Positive Rate is 0.4279215. Table 16 presents the results of test as threshold 0.4650.

Table 16: Downsampling Radial Kernel SVM Confusion Matrix on Test at 0.465 Thershold

Metric	Observations	Rate	Percentage Total Observations
Correct	13096	0.6458	0.64579
Missclassified	7183	0.3542	0.35421
True Positive	1093	0.5756	0.05390
True Negative	12003	0.6530	0.59189
False Negative	806	0.4244	0.03975
False Positive	6377	0.3470	0.31446

The False Positive Rate on test increases to 0.35 as expected while the False Negative Rate decreases to 0.42. That's a 8% increase compared to the 0.5 threshold model. In other words, if one use this model to predict readmission, he or she would falsely predict about 4 patients as non-readmitted out of 10 readmitted patient. This result means that this model is still not good enough in production

Table 17: F_1Score Summary

Metric	Value
Precision	0.1463
Recall	0.5756
F_1Score	0.2333

but it is the best out of all models we developed.

5 Conclusion and Discussion

Intensive data prepossessing, exploration, and modeling are done to find the best model predicting readmission. Different methods are used to treat imbalanced datasets and overfitting problems, two big issues of machine learning classification. Hyperparameters tuning using grid searching as well as random searching are utilized for all models except the baseline to explore their best performance in different subsampling scenarios. Parallel processing also helped to improve training time. At the end, Downsampling Radial Kernel SVM is the best model which is a great improvement over the baseline model.

The performance of the best model still doesn't meet the team's expectations. The team would like to improve more one predicting readmitted cases. In the future, spending more time on feature engineering is needed to get more useful predictors. Also, the size of the dataset is large; thus, imbalanced dataset, other subsampling methods, other SVM kernels, as well as more advanced tree-based algorithms such as XGBoost cannot be used because they need substantially more training time. Therefore, if time is not the constraint, more models would be developed. After that, Stacking is a good experiment to combine all algorithms to obtain a better prediction.

Appendix

Table 18: Dataset Information

Feature Name	Type	Description	%Missing
encounter_id	Numeric	Unique identifier of an encounter	0%
patient_nbr	Numeric	Unique identifier of a patient	0%
race	Nominal	Caucasian, Asian, African, American, Hispanic, and other	2%
gender	Nominal	male, female, and unknown/ invalid	0%
age	Nominal	10-year intervals: [0, 10), [10, 20),..., [90, 100)	0%
weight	Numeric	Weight in pounds	97%
admission_ source_id	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
disposition_id	Nominal	Discharge disposition Nominal Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
admission _source_id	Nominal	Admission source Nominal Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital 0%	
time_in _hospital	Numeric	Time in hospital Numeric Integer number of days between admission and discharge	0%
payer_code	Nominal	Payer code Nominal Integer identifier corresponding to 23 distinct values, for example, Blue CrossÄBlue Shield, Medicare, and self-pay	52%
medical_ specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, familyÄgeneral practice, and surgeon	53%
num_lab_ procedures	Numeric	Number of lab tests performed during the encounter	0%
num_procedures	Numeric	Number of lab tests performed during the encounter	0%
num_medications	Numeric	Number of distinct generic names administered during the encounter	0%
number_ outpatient	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
number_ emergency	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
number_ inpatient	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%

Table 19: Dataset Information (Cont.)

Feature Name	Type	Description	%Missing
diag_1	Nominal	Diagnosis 1 Nominal The primary diagnosis (coded as first three digits of ICD9): 848 distinct values	0%
diag_2	Nominal	Diagnosis 2 Nominal Secondary diagnosis (coded as first three digits of ICD9): 923 distinct values	0%
diag_3	Nominal	Diagnosis 3 Nominal Additional secondary diagnosis (coded as first three digits of ICD9): 954 distinct values	1%
number	Numeric	Number of diagnoses Numeric Number of diagnoses entered to the system	0%
max	Nominal	Glucose serum test result Nominal Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,”	0%
A1Cresult	Nominal	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	0%
metformin repaglinide nateglinide chlorpropamide glimepiride acetohexamide glipizide glyburide tolbutamide pioglitazone rosiglitazone acarbose miglitol troglitazone tolazamide examide citoglipton insulin glyburide- metformin glipizide-metformin glimepiride- pioglitazone metformin- rosiglitazone metformin- pioglitazone	Nominal	Diabetes medications Nominal Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” 024 features for medications. For the generic names: metformin, repaglinide, nateglinide, chlorpropamide,glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone,rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin,glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone,metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the drug was not prescribed	0%

Table 20: Dataset Information (Cont.)

Feature Name	Type	Description	%Missing
change	Nominal	Change of medications Nominal Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
diabetesMed	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
readmitted	Nominal	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.	0%

References

[1]Chang, Huan. “Reducing Preventable Readmissions for Patients with Diabetes on the Parkland Inpatient Hospitalist Units,” n.d. http://www.ihi.org/education/IHIOpenSchool/blogs/Documents/Tim%20(Huan%20Ting)%20Chang.pdf.
[2]Brandão, Humberto. “Diabetes 130 US Hospitals for Years 1999-2008,” October 31, 2017. https://www.kaggle.com/brandao/diabetes.
[3]Navlani, Avinash. “KNN Classification Using Scikit-Learn.” DataCamp Community, August 2, 2018. https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn.
[4]UC Business Analytics R Programming Guide. “Naïve Bayes Classifier.” Naïve Bayes Classifier · UC Business Analytics R Programming Guide. Accessed December 8, 2020. https://uc-r.github.io/naive_bayes.
[5]Akalin, A. (2020, September 30). Computational Genomics with R. Retrieved December 08, 2020, from https://compgenomr.github.io/book/logistic-regression-and-regularization.html
[6]Sontag, David. “Support Vector Machines amp; Kernels Lecture 6 ,” n.d. http://people.csail.mit.edu/dsontag/courses/ml13/slides/lecture6.pdf.
[7]Kuhn, Max. “The Caret Package.” 11 Subsampling For Class Imbalances, March 27, 2019. https://topepo.github. io/caret/subsampling-for-class-imbalances.html.