The 30-days hospital readmission risk in diabetic patients: classification with machine learning classifiers

*Prathamesh - Melissa - Sai Prajit*

1. **Abstract**

In 2018 the CDC estimated that 34.2 million people (10.5% of the population) within the United States have diabetes.  “The estimated economic burden associated with diabetes in the US in 2012 was $245 billion in the form of higher medical costs ($176 billion) and reduced productivity ($69 billion).” (American Diabetes Association, 20018).  Diabetes is a significant comorbid condition which impacts the overall cost of healthcare for the individuals affected.  The UCI Diabetes Data Set includes treatment records for 130 US hospitals between the years of 1999 and 2008.  This data set has 49 attributes for the patients treated within these hospitals as well as the target variable, readmission within 30 days.  Our goal is to build and evaluate machine learning models that can predict if a patient with diabetes will be readmitted within 30 days of their discharge.  Identifying these patients and the attributes that predict readmission will allow healthcare providers to provide more tailored treatment during the first hospital stay and attempt to drive down the probability of readmission for these patients.

1. **Introduction**

In order to prepare the data set for modeling, all of the features were investigated and deleted, modified or retained as original. Several features were removed due several different reasons outlined in more detail below. A correlation matrix was created to determine which variables should be eliminated for redundant information. Categorical features were modified to numeric and all features were normalized. The target variable was originally multi-class: readmission less than 30 days, readmission greater than 30 days and no readmission. This was converted to a binary variable with readmission less than 30 days being the positive class and the remaining two variables combined to the negative class. This target variable was unbalanced, so oversampling was used to create a balanced data set. The final dataset after preprocessing 71k rows and 21 features.

A baseline model was established to test our future model’s effectiveness on prediction. Machine Learning Algorithms selected: Logistic Regression, K-Nearest Neighbors, Support Vector Machine and Naive Bayes. The dataset was split between training and testing data sets. Hyper parameters were selected with a 5-cross validation grid search CV. The optimal parameters were selected and the model was run on the testing data set.

Each model was evaluated using the ROC, PR, AUC and F1 Score. Ultimately the models were evaluated against each other using their Recall Scores. This score was selected as the best measure of performance due to the nature of the problem. Our target variable was unbalanced, the positive class being much smaller than the negative class. Using the Recall Score to evaluate our models we can gain better insight into the success of each model in predicting the patient’s readmission.

At the conclusion, we identified the variables that have the strongest impact on the readmission probability. Ultimately, if we are able to identify these variables, providers can check each patient during patient admission to determine if they are at risk for readmission within 30 days after discharge. The provider will then be able to offer additional treatment or more specific treatment to attempt to drive down this probability and reduce the cost to the patient and impact on their health.

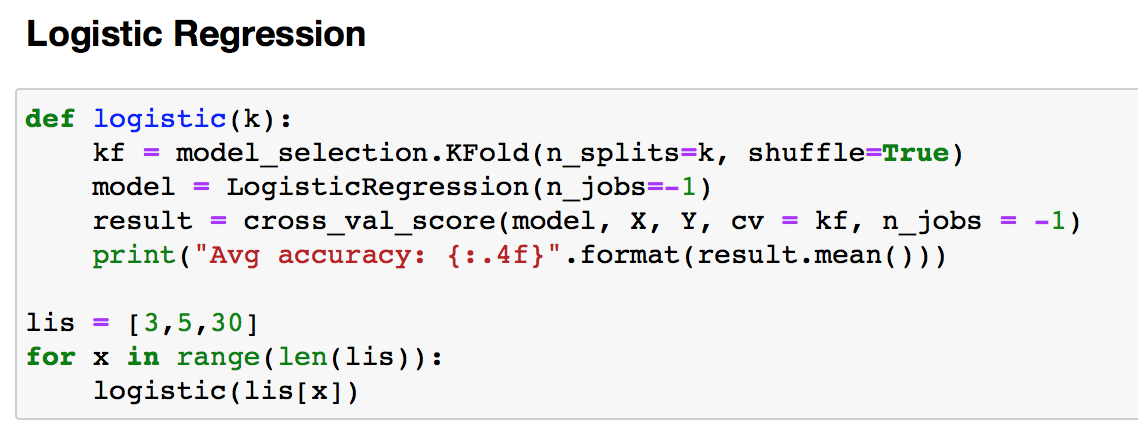
1. **Related Work**

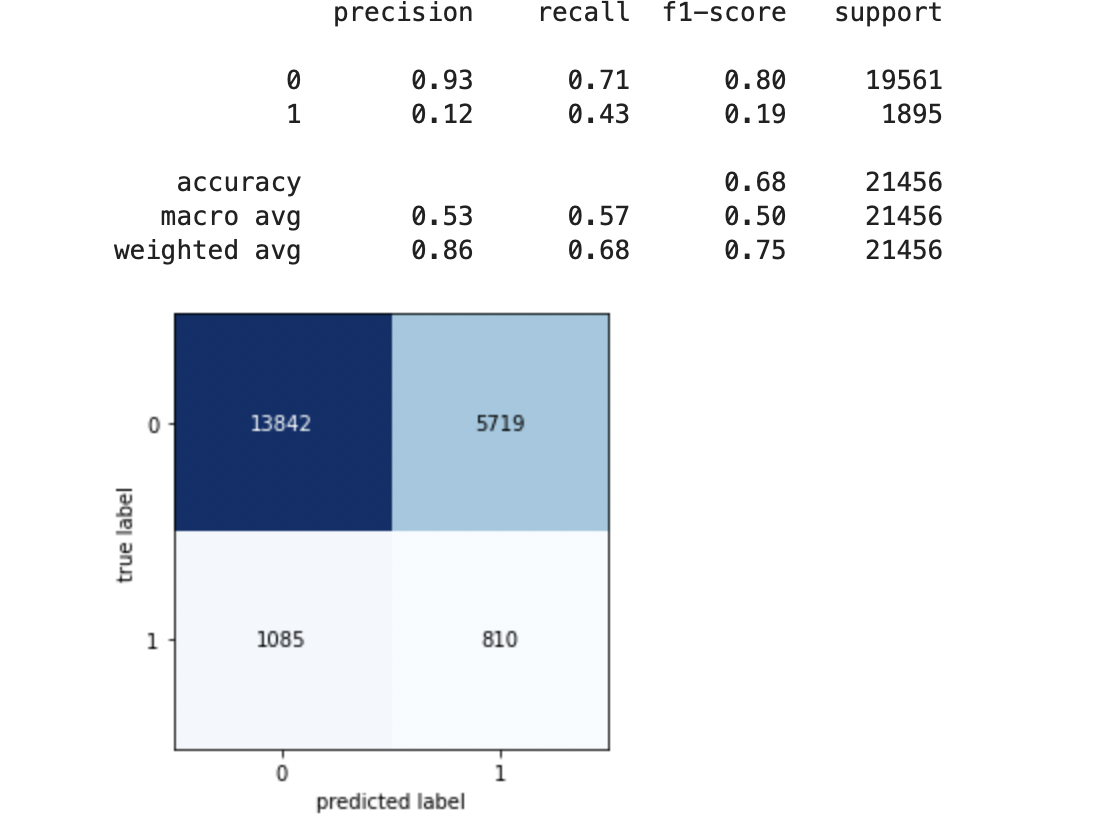
While working on this problem, we reviewed other notable work on this dataset.

Two of the papers reviewed are summarized below:

In 2014, Strack Et. Al. modeled the impact of the patient’s HbA1c levels on the patient’s readmission status using this same data set and multivariable logistic regression.  A patient’s HbA1c can be measured through an A1C test.  This test will provide a patient’s average blood sugar for the three months leading up to the test.  The results of this test (given in a percentage) will dictate if the patient is normal (less than 5.7%), prediabetic (between 5.7% and 6.5%) or diabetic (greater than 6.5%).  Strack Et. Al. found that this test was only ordered in 18.4% of inpatients and when this test was ordered on patients diagnosed with diabetes, they were readmitted less frequently, regardless of their A1C result.  Through their analysis they ultimately found that patients with a circulatory diagnosis had the highest rate of readmission.

In 2021, Shang Et. Al. used Random Forrest, Naive Bayes and Tree Ensamble to discover that “the number of hospitalizations, age, length of hospital stay and sex were the main features that determined the probability of accidently readmission”. They further discovered that “length of stay longer than 5 days was associated with a greater than 87% risk of readmission compared with a stay shorter than or equal to 2 days.” When evaluating diagnosis codes they found that the second diagnosis code was a more accurate predictor than the first diagnosis code. Shang Et. Al. hypothesized that “the subsequent diagnosis in a patient’s EHR could more accurately reflect the patient’s condition”.

1. **Model**
2. ***Logistic Regression***

Logistic regression is a statistical method that analyzes the data, in which there are one or more independent variables that determine the outcome where the outcome has only two possible values.

For example, classifying a boy or girl, binary digits 0 or 1 etc.

Binary logistic regression identifies the influence of numerous independent variables provided at the same time to predict membership in one of two dependent variable categories. We can't predict a numerical value for the dependent variable using logistic regression because it's dichotomous, so the standard regression least squares deviations criteria for best fit approach of minimizing error around the line of best fit is ineffective (it's impossible to calculate deviations with binary variables!). Instead, logistic regression uses binomial probability theory, in which only two values are predicted: that the probability (p) is 1 rather than 0, indicating that the event/person belongs to one of two groups.

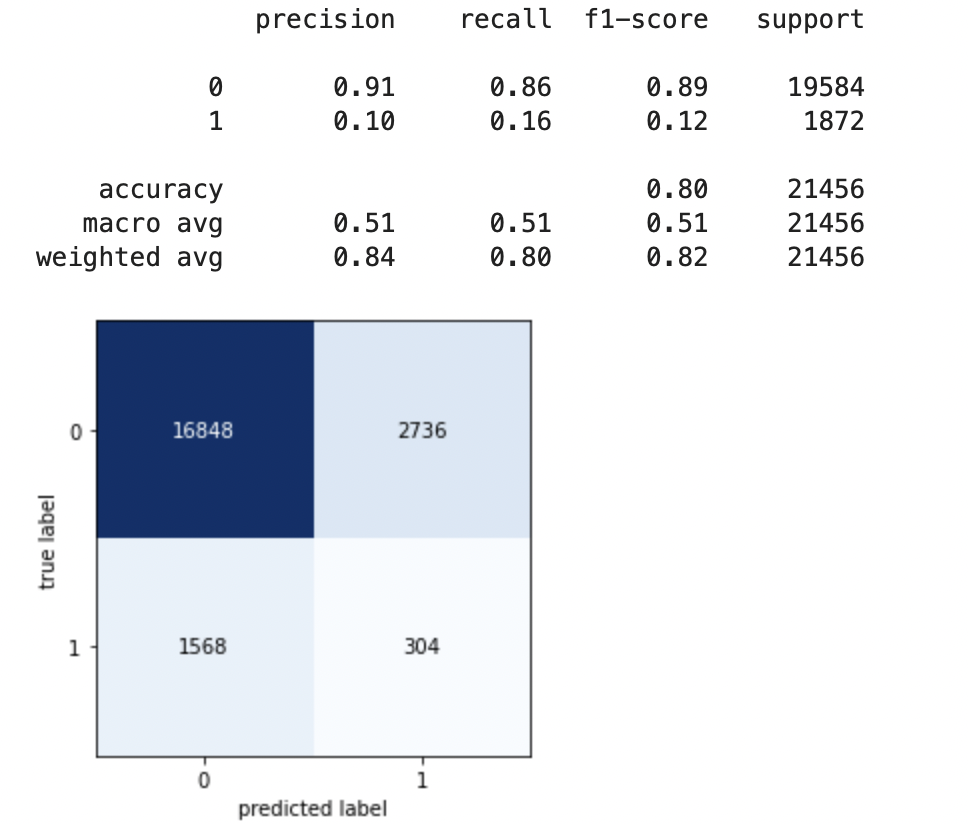


**Hyper Parameters:**

‘C’: [0.001, 0.01, 0.1, 1, 10, 100, 1000]

1. ***K-Nearest Neighbors (KNN)***

The k-nearest neighbors (KNN) algorithm is a data classification approach that estimates the likelihood that a data point will belong to one of two groups based on the data points closest to it. Unlike artificial neural network classification, k-nearest neighbors classification is simple to understand and implement. It's suitable for scenarios with well-defined or non-linear data points. To put it in other words, KNN uses a voting method to decide the class of an unknown observation. This signifies that the data point's class will be determined by the class with the most votes. If the value of K is 1, we’ll only utilize the nearest neighbor to identify the data point’s class. If K equals 10, we’ll use 10 nearest neighbors, and so on. There is no specific approach for determining the best K value, or the number of neighbors in KNN. This means you may need to experiment with a few different values before selecting which one to use.



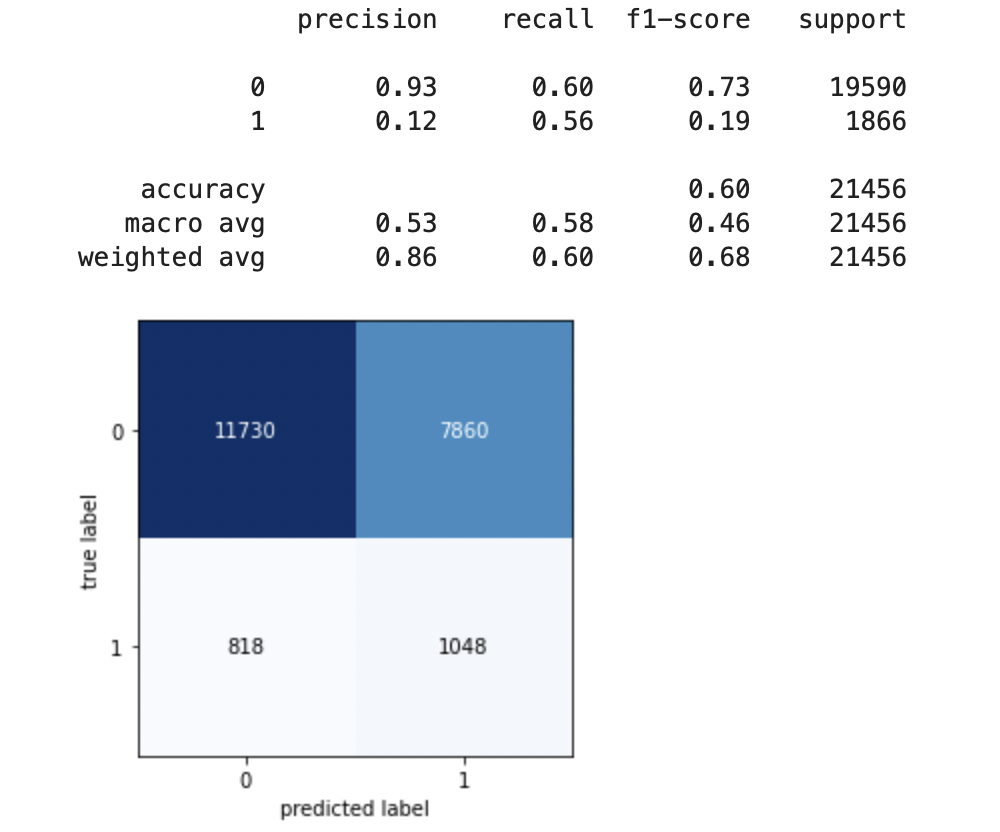
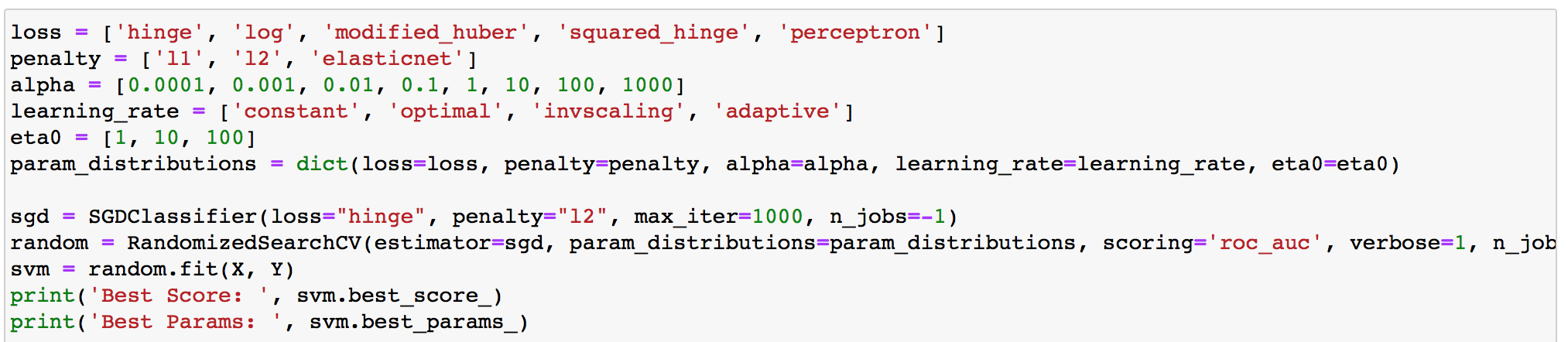
**Hyper Parameters:**

‘n\_neighbors’: [1,2,3,50]

1. ***Support Vector Machine (SVM):***

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. SVM models can categorize new text after being given sets of labeled training data for each category. They have two key advantages over newer algorithms like neural networks: greater speed and better performance with a limited number of samples (in the thousands). This makes the approach particularly well suited to text classification issues, where it's common to only have access to a few thousand tagged samples.

We chose to use SGD regression over Linear SVM because of the size of the dataset. Minibatch (online/out-of-core) learning is possible with SGD. As a result, SGD makes sense for large-scale situations where it is highly efficient. Because the minimum of the Logistic Regression cost function cannot be calculated directly, we use Stochastic Gradient Gradient Descent, to try to minimize it. For each training observation that we encounter, we decrease the cost function towards its minimum. Another reason to use SGD Classifier is that if you can't maintain the record in RAM, SVM or logistic regression won't function. SGD Classifier, on the other hand, continues to function.

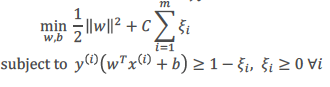


**Hyper Parameters:**

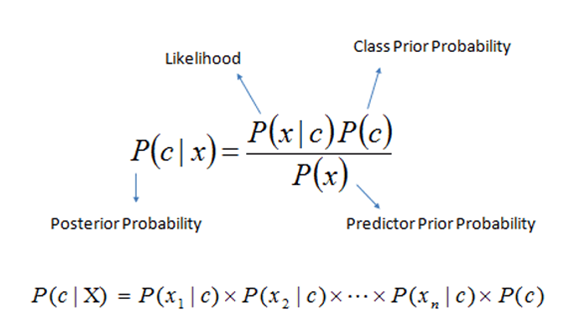
loss = ['hinge', 'log', 'modified\_huber', 'squared\_hinge', 'perceptron']

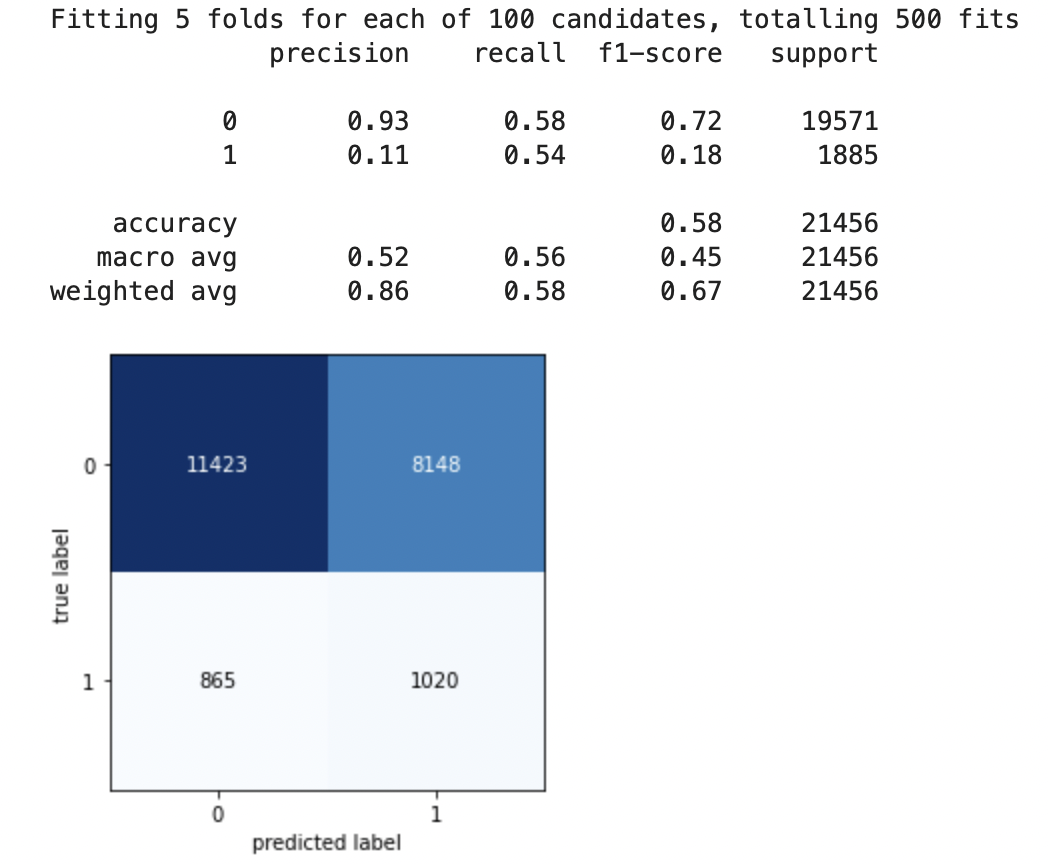
penalty = ['l1', 'l2', 'elasticnet'] alpha = [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]

learning\_rate = ['constant', 'optimal', 'invscaling', 'adaptive'] eta0=[1, 10, 100]



1. ***Naïve Bayes:***

The probabilistic model of naive Bayes classifiers is based on Bayes’ theorem. Naïve Bayes classifiers are linear classifiers which are known to be simple and efficient models. Naïve Bayes is one of the better models when text classification is concerned.



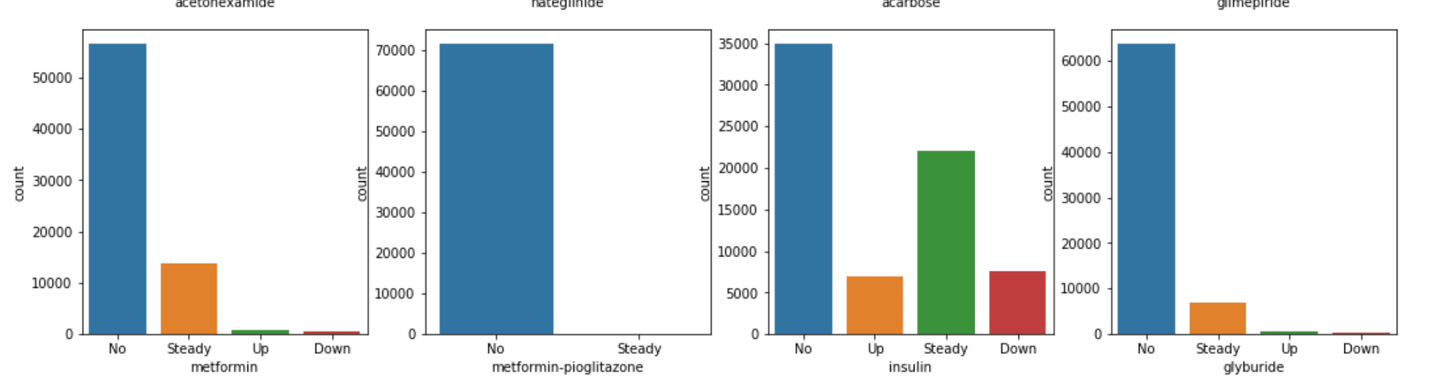
**Hyper-Parameter** 

param\_distributions = {'var\_smoothing': np.logspace(0,-9, num=100)}

1. **Experimental Results**

***1. Data Preparation***

After importing the data set, all of the 49 features were reviewed, deleted or modified to create a final data set for modeling. The target variable, readmission status, has a much smaller volume of positive class than negative class. Oversampling was utilized to develop a balanced data set for the final model. Features with large volumes of missing values were deleted (weight, medical specialization, payer code, A1C test and Ma Glue Serum). Features that only provide identifying information were also deleted (Encounter ID, Patient NBR). Drug features with a disproportionate number of “No” values were also removed. Some categorical features had an extremely high volume of distinct values (700+) and were also removed (Diagnosis 1, Diagnosis 3, Diagnosis 3). Patients who had multiple visits, who were readmitted several times, the subsequent visits were removed so each patient was only captured one time in the data set. Categorical features were converted to numeric features. All features were normalized. The final size of the data set before modeling: 71k rows and 21 features.



**2. Correlation Matrix**:

Correlation is a statistical method for determining the strength of a relationship between two variables. It can be either a positive, negative, or zero value. Any value between +1 and -1can be used for the correlation coefficient. A correlation matrix is a table that displays the coefficients of correlation between sets of variables.

***3. Baseline Model****:*

While performing the baseline model, we received naïve accuracy of 0.91 – non-admitted: total patients

***4. Model Preparation****:*

The final data set was split between training (70%) and testing (30%). As mentioned above, oversampling was used to balance the data in the training set. Each initial model was evaluated through a 5-Cross Validation with Grid Search CV to find the best set of parameters. Once the optimal parameters were found they were run on the testing set and the performance was evaluated through a Precision Recall Score, AUC Score and F1 Score. The model’s performance was evaluated against each other by using the Recall Values.

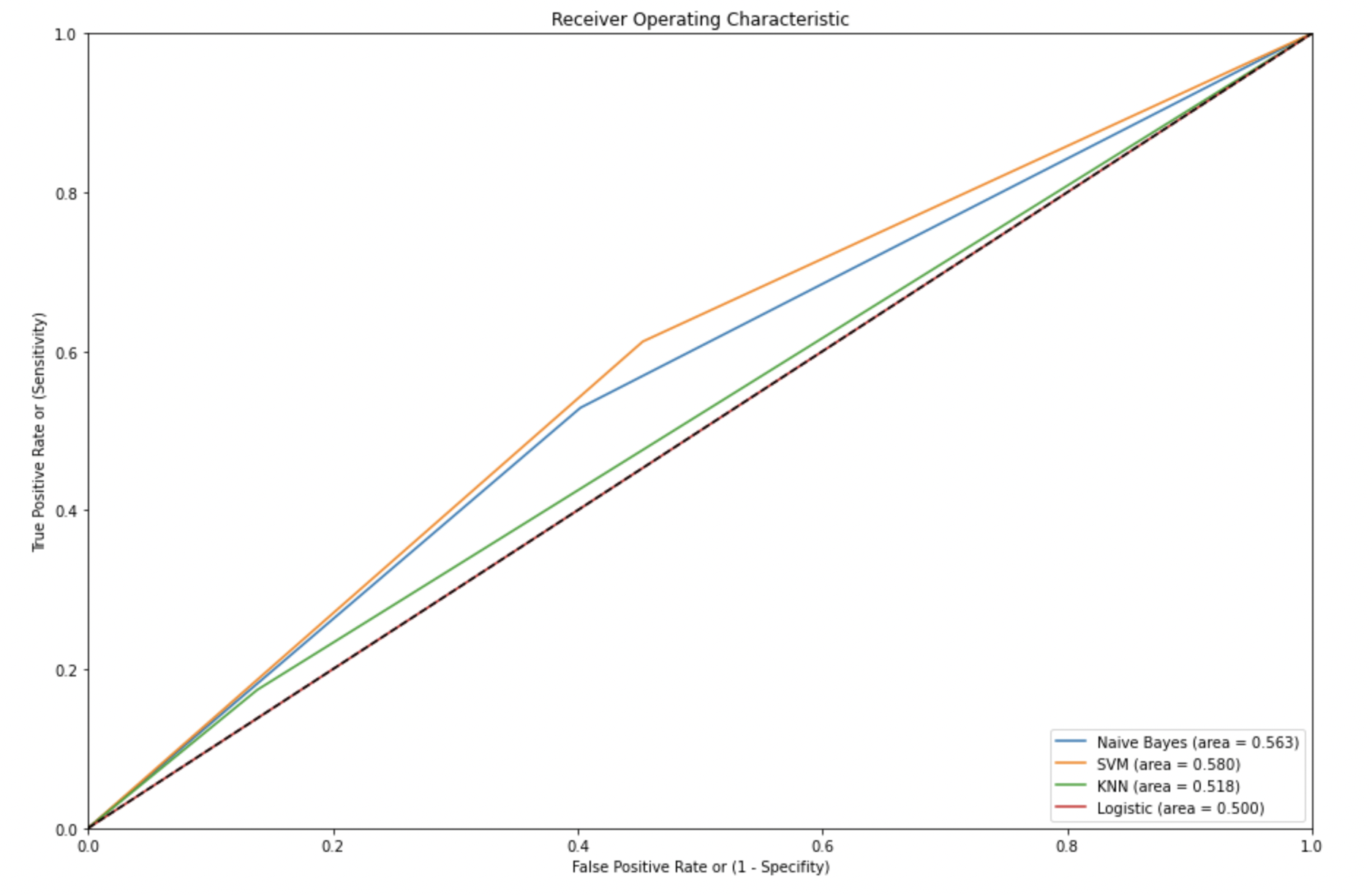
***5. Cross Validation****:*

The k-fold cross-validation procedure is a standard method for estimating the performance of a machine learning algorithm or configuration on a dataset. Repeated k-fold cross-validation is a model for improving a machine learning model's estimated performance. Simply repeating the cross-validation procedure multiple times and reporting the mean result across all folds from all runs is all that is required. This mean result is expected to be a more accurate estimate of the model's true unknown underlying mean performance on the dataset, as calculated with the standard error.

***6. Receiver Operating Characteristic (ROC) Curve:***

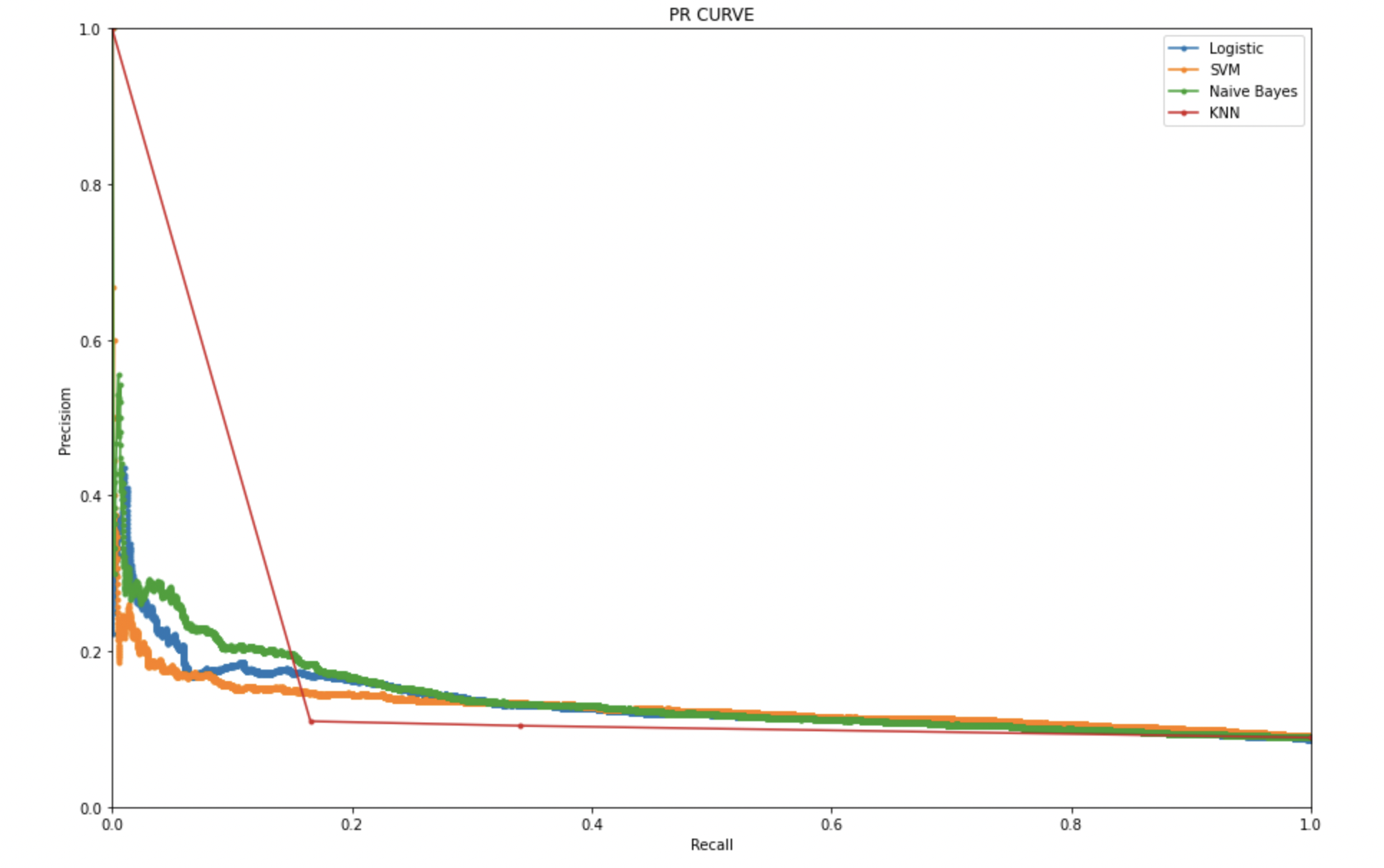
The ROC Curve plots the results of the False Positive Results (FPR) against the results of True Positive Results (TPR). The greater the Area Under the Curve (AUC) the greater the performance. This plot also offers the benefit of comparing these two measures against each other and selecting for the measure that is more important to the problem. In this case we would like to drive down FPR results and select the best AUC that also reduces FPR. A drawback of using this curve is that it can be insensitive to changes in label distribution. The ROC results are shown below.

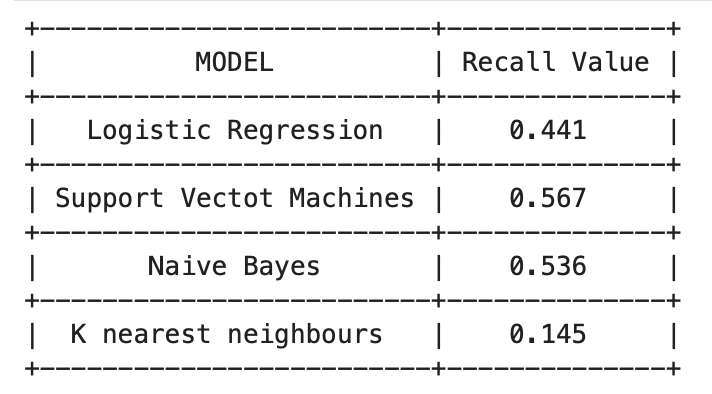




***7. Precision Recall (PR) Curve:***

The PR Curve plots the results of the Precision (% of positive examples from all actual positives) against the results of Recall (% of positive examples from all predicted positives). One of the benefits of using the PR Curve is that it can handle a data set with label imbalances that contain a greater volume of negative labels than positive ones. This describes the UCI Diabetes Data Set. Again the measure to review is the Area Under the Curve or AUC. But a drawback in using this curve is that the false positives are more important than the true positives. As a result, the models’ performance was measured based on their Recall values.

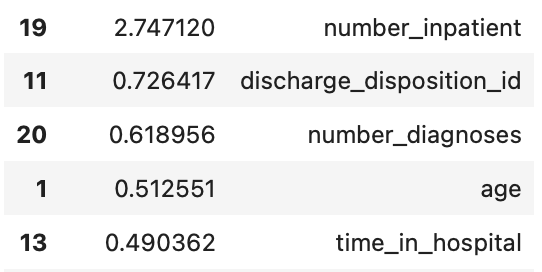
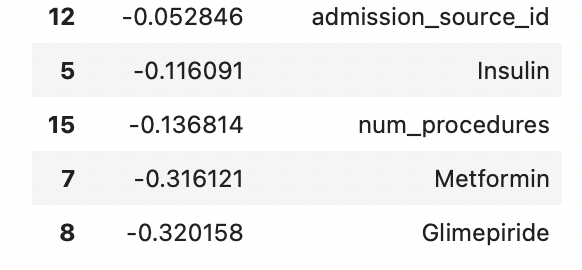


***8. Recall Measure:*** 

The percentage of positive examples from a set of all positive predictions. In this measure the Support Vector Machine (SVM) model performed the best, followed closely by the Naive Bayes Model.

***9. Feature Importance:***

Ultimately the goal of this Machine Learning Project is to identify the features that should be flagged to health care providers so they may tailor their treatment to patients who are admitted with Diabetes in an effort to drive down readmission with these patients. In order to identify the most important and least important features, all of the features were ranked, using the feature coefficient in SVM.

****

***Least important features Most important features***

***10. Features Identified:***

The number of inpatient visits in the prior year, discharge disposition ID, number of diagnoses, age and the time in hospital were identified as the features with the greatest impact on readmission within 30 days. Hospitals may wish to create an intake system for diabetic patients that captures the patient’s age, number of inpatient visits in the prior year and the number of diagnoses. Then utilize these features to create a system to flag the treating doctor to identify and focus treatment for patients with the greatest risk. Given that the discharge disposition and time in hospital will not be available until the treatment is completed and the patient is discharged, these features may not provide much benefit. Though, the time in hospital feature can be triggered when the patient reaches a date threshold of high risk.

1. ***Discussion***

In this study, Logistic regression, Naive Bayes, KNN and Logistic Regression were used to construct a 30-day hospital risk model. Data were used to train and validate the model on a 5 cross validation. The SVM and Naive Bayes algorithm showed good predictive performance for the positive class. The complete model design process shown here involves the selection of an algorithm, which will reference the value of other similar designs for similar predictions in the future.

According to the model results, the number of diagnosis, age, time in hospital stay, and sex were the main factors that determined the chances of accidental rehabilitation to predict reversal cases. Previous research has shown that admission times, age, and sex are important in many hospitals. “Duration longer than 5 days is associated with a higher risk of recurrence of 87% compared to duration.”

Additional sleep periods were a major cohort of patients who re-studied frequently, mainly in older patients with more severe conditions, and the duration of bed rest was much longer than in normal patients. In addition, routine tests revealed that these patients were more likely to have diabetes-related complications. In addition, diag\_2 was significantly more important than diag\_1 among the three diagnostic codes, indicating that subsequent diagnosis in a patient's EHR could accurately reflect a patient's condition. Therefore, health care providers should provide health education and follow-up to prevent complications in hospitalized patients, especially older patients, who were consistent with those in Kampala [4] that patient education, medication preparation, and discharge planning were significantly reduced. Recurrence episodes also reduced hospital stay due to recurrent hypoglycemia. By considering dispersed diagnostic codes that do not comply with the inclusion of diagnostic features, we have improved the accuracy of the predictive model and reduced the processing time by reducing the size of the data in the pre-processing phase. Next, we will examine which diagnoses have the most impact on 30-day study levels.

There are many old ML algorithms for dividing medical data. The RF algorithm can generate a DT algorithm across multiple databases, suggesting that it could be a way to calculate the value of a feature [7]. The majority of patients in the database did not have a recovery record (53.69%), while only 11.22% were returned within 30 days (<30), while the remaining patients (35.09%) were returned after 30 days (> 30). after the first exit. In fact, re-reading after> 30 days was difficult to quantify because there was not much difference between admissions on day 30 and those on day 31. Decreasing the total separation accuracy is a key goal of standard ML algorithms. The larger category gains greater attention to the classification system when data inequality occurs, and the performance of small sample identification decreases [9]. However, the target category that needs to predict the smallest portion of total value in medical data. The conflict between sensitivity and specificity was significantly reduced when the training set was limited. The DM data used in this study is an international public data set with unqualified quality control, resulting in significant research limitations.

1. ***Conclusion***

The machine learning models can help health care providers identify those patients who are at high risk to short-term readmission and may reduce their chances of being reinstated within 30 days by altering risk factors.



1. ***References***
2. American Diabetes Association. (2018). Economic costs of diabetes in the US in 2017. *Diabetes care*, *41*(5), 917-928.
3. Centers for Disease Control and Prevention. (2020). National diabetes statistics report, 2020. *Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services*, 12-15.
4. Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., ... & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers. *BMC medical informatics and decision making*, *21*(2), 1-11.
5. Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, *2014*.
6. Ramentol E. SMOTE-RS: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl Inf Syst. 2012;33(2):245–65.
7. Wikipedia: The Free Encyclopedia. Wikimedia Foundation. List of icd-9 codes. Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. Web, 2014.
8. Koola JD, Ho SB, Cao A, Chen G, Perkins AM, Davis SE, Matheny ME. Pre- dicting 30-day hospital readmission risk in a national cohort of patients with cirrhosis. Dig Dis Sci. 2020;65(4):1003–31.
9. Jiang HJ, Andrews R, Stryer D, Friedman B. Multiple hospitalizations for patients with diabetes. Diabetes Care. 2003;26(5):1421–6.
10. Enomoto LM, Shrestha DP, Rosenthal MB, Hollenbeak CS, Gabbay RA. Risk factors associated with 30-day readmission and length of stay in patients with type 2 diabetes. J Diabetes Complicat. 2017;31(1):122–7.
11. Kampan P. Effects of counseling and implementation of clinical pathways on diabetic patients hospitalized with hypoglycemia. J Med Assoc Thai. 2006;89(5):619–25.