

# Determining early readmission of diabetes patients within 30 days of discharge

Lakshmi Gayathri Rangaraju  
*lrangar@clemson.edu*  
Graduate Student  
Clemson University

Charan Basireddy  
*cbasire@clemson.edu*  
Graduate Student  
Clemson University

Sandya Rani Prasadam  
*sandyap@clemson.edu*  
Graduate Student  
Clemson University

Venkata Subbu Sai Hiranmayee Machavolu  
*Vmachav@clemson.edu*  
Graduate Student  
Clemson University

## 1 COURSE DETAILS

**Course Name:** Applied Data Science  
**Course Number:** 6300  
**Semester:** Fall 2023

## 2 INTRODUCTION

### 2.1 What is the main question your project seeks to answer ?

Our project centers around the pivotal question: Can we construct a reliable predictive model to ascertain the likelihood of hospital readmission within a 30-day period for diabetic patients? Leveraging the extensive diabetes dataset spanning the years 1999-2008 from the UCI Machine Learning Repository, our overarching goal is twofold. Firstly, we aim to develop a robust predictive model that goes beyond a mere binary prediction, delving into the identification of the nuanced factors contributing to hospital readmissions. By scrutinizing patient data, our intention is to pinpoint the key variables influencing readmission risk. This endeavor holds immense significance as it can empower healthcare providers with the ability to proactively identify high-risk diabetic patients and tailor interventions, accordingly, potentially reducing readmission rates and improving overall patient outcomes. Additionally, our project seeks to contribute to a broader understanding of temporal trends in diabetes-related hospital admissions, unraveling patterns, and changes over the decade under study. Through these endeavors, our research aspires to not only enhance predictive capabilities but also provide actionable insights that can positively impact healthcare resource allocation and, ultimately, the quality of care for diabetic populations.

### 2.2 Provide a brief motivation for your project question. Why is this question important? What can we learn from your project ?

The question of predicting hospital readmission within a 30-day window for diabetic patients is of paramount importance due to its potential to revolutionize healthcare delivery for individuals with diabetes. Diabetes is a chronic condition with significant implications for patients' well-being and healthcare systems. By addressing this question, our project seeks to fill a critical gap in current medical practices. Accurate prediction of readmission risk allows for timely interventions, personalized care plans, and optimized resource allocation. This has the potential to enhance patient outcomes, alleviate the burden on healthcare systems, and streamline the delivery of healthcare services. Understanding the key factors influencing readmission not only facilitates precise risk assessment but also opens avenues for targeted interventions, ultimately contributing to a more effective and patient-centric approach to diabetes management. Moreover, the temporal analysis provides insights into the evolving landscape of diabetes-related hospital admissions, offering a historical perspective that can inform future healthcare strategies. In essence, our project has the potential to deliver actionable insights that can significantly improve the quality of care for diabetic populations, marking a crucial step towards a more efficient and patient-oriented healthcare system.

### 2.3 Briefly describe the data source(s) you have used in your project. Where is the data from? How big is the data in terms of data points and/or file size? If the data was not already available, how did you collect the data ?

The primary data source for our project is the diabetes dataset obtained from the UCI Machine Learning Repository. This dataset spans the years 1999-2008 and comprises information

from 130 U.S. hospitals. The dataset is publicly available and was pre-processed for research purposes. It includes a comprehensive set of 47 features related to diabetic patient admissions, covering demographic information, clinical attributes, and outcome variables such as hospital readmission status within a 30-day period. The dataset is substantial, containing a significant number of data points that capture the diversity of diabetic patient profiles and their healthcare experiences. Specifically, it consists of 101766 instances and 18,711 kb in terms of file size. The use of this dataset provides a rich and extensive foundation for our predictive modeling and analysis. As the data was already available and curated by the UCI repository, there was no need for additional data collection efforts for this particular project.

### 3 SUMMARY OF EDA

#### 3.1 What is the unit of analysis ?

Given that the goal is to determine the early readmission of patients within 30 days of discharge, the most suitable unit of analysis would likely be at the patient level. Analyzing the data at the patient level allows for a comprehensive understanding of individual patient characteristics, treatments, and outcomes, which are essential in identifying patterns and factors associated with early readmissions.

By focusing on the patient level, you can examine various patient-specific variables, such as demographics, medical history, treatment procedures, and clinical measurements, to identify potential risk factors or predictors that contribute to early readmissions for patients with diabetes. This approach enables you to assess the impact of different factors on the likelihood of early readmission, helping you identify strategies for improving patient care and reducing readmission rates.

Analyzing the dataset at the patient level would involve exploring patient-specific features, such as demographic information, clinical indicators, treatment details, and outcomes, to conduct a comprehensive assessment of the factors contributing to early readmissions within the 30-day post-discharge period. This approach can provide valuable insights into the specific patient characteristics and clinical factors that influence the likelihood of early readmissions, facilitating the development of targeted interventions and strategies to improve patient outcomes and reduce readmission rates for individuals with diabetes.

#### 3.2 How many observations in total are in the data set ?

The entire data set consists of 101766 observations in total, which will be split into 90% training data, 10% test data. Each row in the data set concerns hospital records of patients diag-

nosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days.

#### 3.3 How many unique observations are in the data set ?

The number of unique observations are calculated by considering all the features. The number of unique observations in the entire data set is 2826.

#### 3.4 What time period is covered ?

The time period which is covered in the data set is from 1999 to 2008, which is a ten years of clinical care data at 130 US hospitals and integrated delivery networks.

#### 3.5 Briefly summarize any data cleaning steps you have performed

The target value “readmitted” which consists of values “<30”, “>30” and “no” are converted to 0 and 1. The classes “>30” and “no” are merged into 0 category as our goal is to predict if the patient gets admitted to the hospital within 30 days of discharge i.e. “<30” value. So finally we only had two classes to predict which is 0 or 1 which means if the given patient record has a high chance of getting readmitted within 30 days of discharge or not.

Considering the features, first we deleted the columns(‘encounter\_id’, ‘patient\_nbr’, ‘weight’, ‘payer\_code’, ‘medical\_specialty’) that we thought were irrelevant to the response variable. And then also deleted the columns(‘examide’, ‘citoglipton’) that had the same values across all the rows.

Moreover, There were several features that could not be treated directly since they had a high percentage of missing values. These features were weight (97% values missing), payer code (40%), and medical specialty (47%). Weight attribute was considered to be too sparse and it was not included in further analysis. Payer code was removed since it had a high percentage of missing values and it was not considered relevant to the outcome. Medical specialty attribute was maintained, adding the value “missing” in order to account for missing values.

Replaced all the instances which have ‘?’ with NAN values such that it is easy to apply the pandas inbuilt methods to get a count for number of NAN values or to replace the NAN values with some arbitrary value.

We have 5 different races value, these are - Caucasian, AfricanAmerican, Hispanic, Asian, Other. In the data set, there are 73% of Caucasians. And other 22 percent is divided into African Americans, Hispanics, Asians and Others. Hence, we decided to divide the race feature values into 3 groups like Caucasian, African American and Other as we have high number of Caucasian race samples, and African

American values. For the "Gender" feature, we dropped a single instance which had "Unknown/Invalid" response.

Age features has the values in the format of "[70-80)". To get rid of this parenthesis notation and make the Age variable a numeric value, we changed it according to the following rule: Replacing the value "[70-80)" with 75, the mean of the interval. Likewise, all other interval values are replaced with their mean values.

Admission Type ID has integer identifier corresponding to 9 distinct values - Emergency : 1, Urgent : 2, Elective : 3, Newborn : 4, Not Available : 5, NULL : 6, Trauma Center : 7, Not Mapped : 8. Here, "null", "not available", and "not mapped" are replaced with "NAN" value, and "urgent" value is replaced with "Emergency". For "discharge\_disposition\_id", all other values which include "home" word are grouped into "home" category, otherwise they are replaced with "Other".

Likewise all other categorical features have been grouped into simpler and understandable categories. And the numeric data is not altered much. Data augmentation is performed on the data points which belong to "readmitted" category. And as our goal is to accurately predict if a patient will be readmitted within 30 days of discharge, we need more samples of this category. So data augmentation plays an important role as the data set didn't have much data samples of the same category.

### 3.6 Visualization of the response

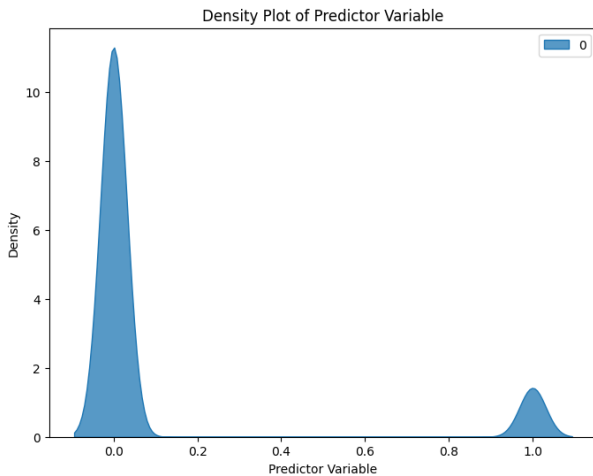


Figure 1: Before Data Augmentation

We can observe from the figure-1 and figure-2 that the predictor variable i.e., data["readmitted"], we can infer that, a peak at  $x = 0$  and another peak at  $x = 1$  suggests that our target variable is a bimodal distribution, meaning it has two distinct modes or peaks. The peak at  $x = 0$  with a higher height indicates that a significant portion of our data points clusters around this value. Similarly, the peak at  $x = 1$ , though smaller, indicates another concentration of data points.

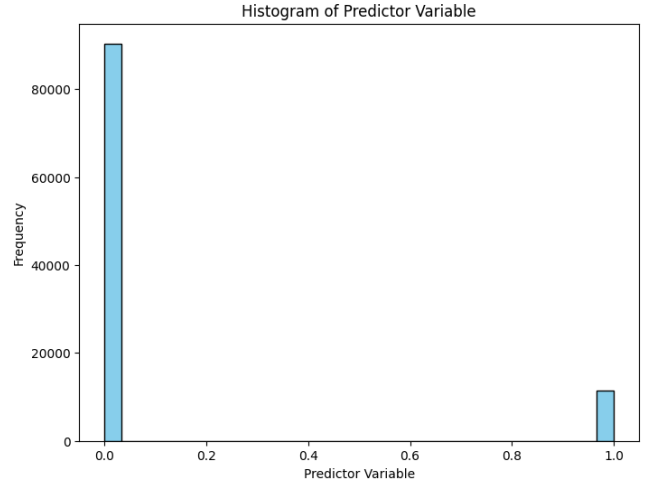


Figure 2: Before Data Augmentation

So before training the model, data augmentation is performed to balance the data set. The data set distribution after performing data augmentation 2x times can be seen in below figures: figure-3 and figure-4.

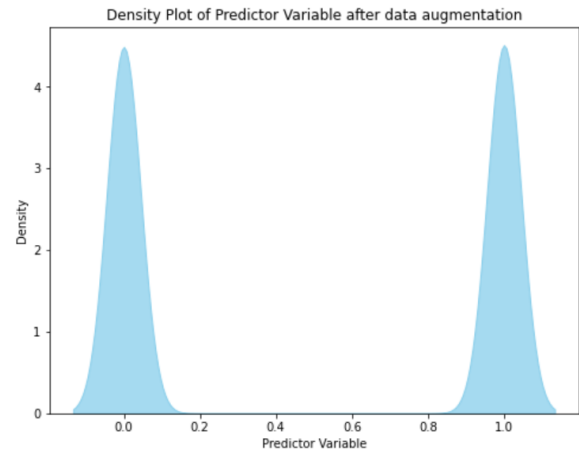


Figure 3: After Data Augmentation

### 3.7 Visualization of key predictors against the response (e.g., scatterplot, boxplot, etc.). Pick one or two predictors that you think are going to be most important in explaining the response. Your selection of predictors can either be guided by your domain knowledge or be the result of your EDA on all predictors

As most of our data is categorical, we performed Chi Square test to identify the key predictors. And the threshold of the

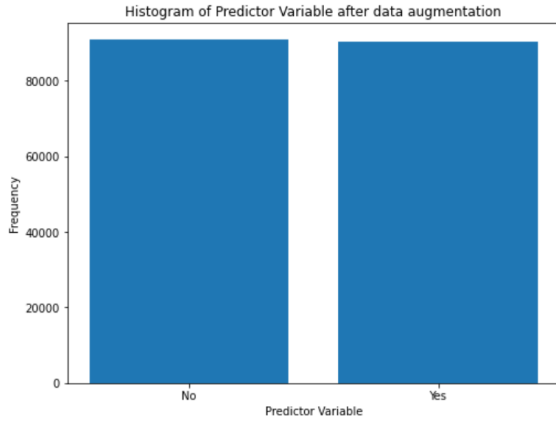


Figure 4: After Data Augmentation

p-value is set as 0.001. The list of attributes along with their p-value after the chi-square test is performed are:

patient_nbr	1.6231957707516826e-07
race	3.693156699943775e-58
gender	1.899934547488499e-08
age	5.1046708632083964e-48
admission_type_id	4.3734314457319205e-79
discharge_disposition_id	0.0
admission_source_id	1.4887499868070666e-220
time_in_hospital	5.4395995682428616e-86
medical_specialty	1.1076254534233351e-192
num_lab_procedures	2.8288254393930463e-30
num_procedures	1.805647819549774e-46
num_medications	2.6572469734526495e-159
number_outpatient	2.301684230538321e-247
number_emergency	0.0
number_inpatient	0.0
diag_1	0.0
diag_2	2.9492520066796185e-270
diag_3	1.5758156561089971e-232
number_diagnoses	6.003769374844997e-291
max_glu_serum	1.2134248248296836e-10
A1Cresult	1.4313260301707168e-11
metformin	4.19890158536121e-18
repaglinide	3.2200658568991624e-11
glipizide	7.395345675080531e-10
rosiglitazone	1.84283775119927e-06
acarbose	6.0690444639684226e-05
insulin	4.264056271676487e-103
change	9.964594376477834e-49
diabetesMed	1.1694104072313729e-85
readmitted	0.0

After examining the above table results, here are our top performers: 'number\_inpatient', 'number\_diagnoses', 'encounter\_id', 'number\_emergency', 'patient\_nbr', 'diag\_1\_428', 'number\_outpatient', 'time\_in\_hospital',

'diabetesMed\_Yes', 'num\_lab\_procedures', 'diag\_2\_250'.

To get clear understanding of how the important features are related to the target variable, graphs are plotted. First plot depicts the relation between "number\_inpatients" vs "Readmission".

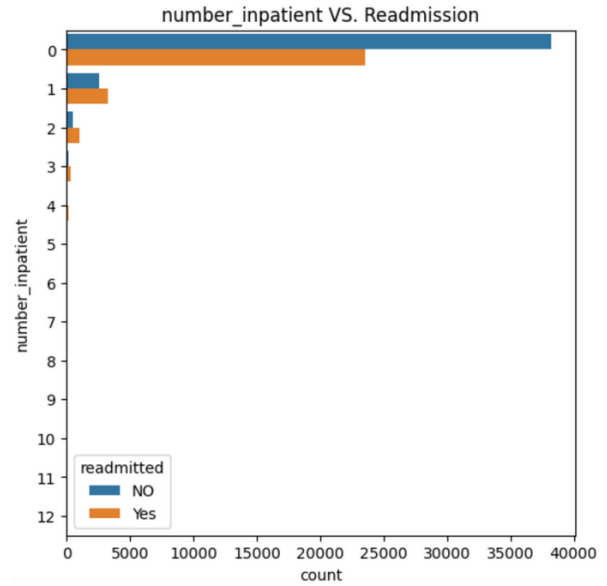


Figure 5: Depicting the relation between number of inpatients and target variable

It can be interpreted that with at least one inpatient has high chances of readmitting again.

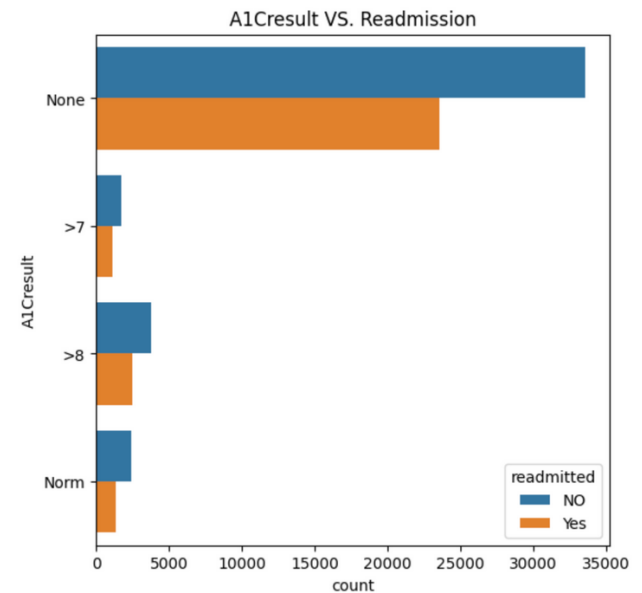


Figure 6: Depicting the relation between A1C\_result vs target label

The "A1C\_result" feature is an important measure of glucose control, which is widely applied to measure performance of diabetes care. The measurement of HbA1c at the time of hospital admission offers a unique opportunity to assess the efficacy of current therapy and to make changes in that therapy if indicated (e.g., HbA1c > 8.0% on current regimen). So this is an important feature which determines if the patient gets readmitted within 30 days of discharge. And from the above plot of A1cresult vs readmitted features, it is interesting to note that readmitted people don't have their blood sugar level measured.

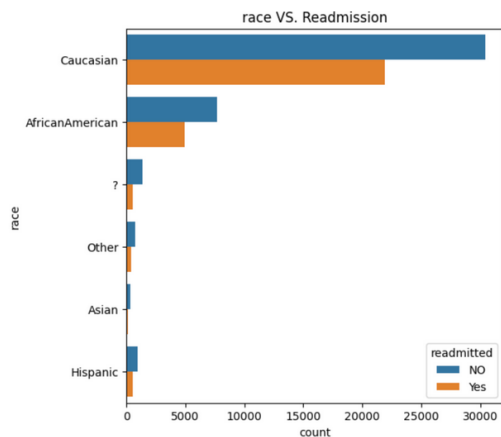


Figure 7: Depicting the relation between race vs target label

In a general sense we think people with different races have different food habits and different cultures which might change their immune system. So "race" feature is considered to determine how it effects the readmission of the people.

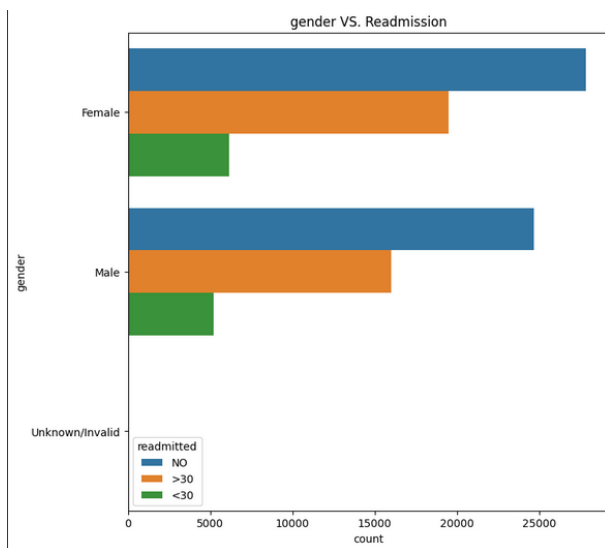


Figure 8: Depicting the relation between gender vs target label

It can be interpreted that females are readmitted more often than males. The factors which could attribute to this are hormonal influences such as menstrual cycle, socioeconomic factors, or Psychological factors.

There are other feature plots which are visualized for better understanding of the dataset.

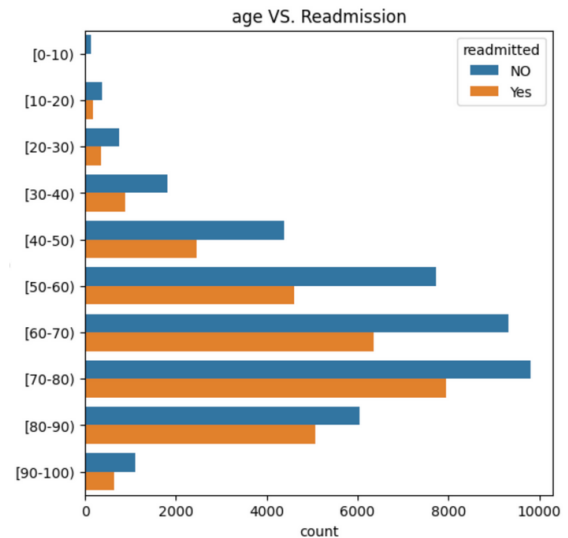


Figure 9: Age vs Readmission

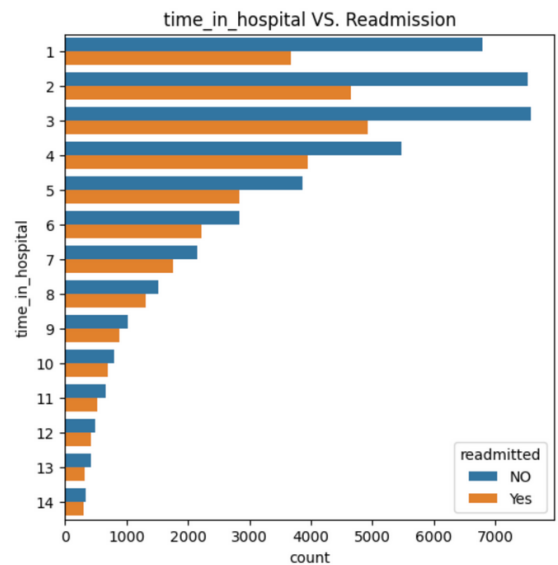


Figure 10: Time in hospital VS Readmission

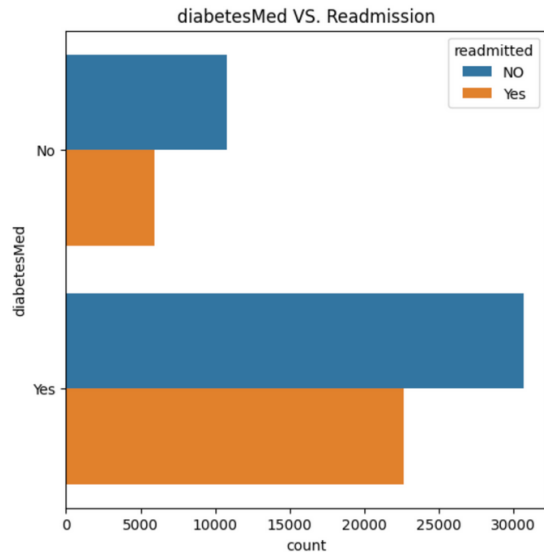


Figure 11: Diabetes\_medication VS Readmission

## 4 SUMMARY OF MACHINE LEARNING MODELS

### 4.1 Justify your model choices based on how your response is measured and any observations you have made in your EDA

The model which we used initially is Support Vector Machine which is abbreviated as SVM. The use of the model SVM is best fit for predicting our categorical target because we got very good true positives (2622) and true negatives (723) with minimal false negatives (786). The goal of our model is to predict if a patient will be readmitted within 30 days of discharge given the medical and demographic information of the particular patient. In light of the above goal, out of 6816, we only have 786 samples which are classified as false negatives and have 2622 samples which are correctly classified as positive labels. It will be a risk for life, if our model had high false negatives however, our model is able to predict the positive label correctly. Considering these metrics, our initial assumption is that SVM is a good fit for our model. On the whole, SVM gave an F1 score of 0.56 with 0.6 recall and 0.49 precision.

Moreover, the EDA revealed that the data is bimodal, which means that there are two classes for the predictor variable. And as SVMs are commonly used for classification tasks by aiming to find a hyperplane that best separates different classes, they are well-suited to our classification of bimodal data. By default SVMs provide a natural way to identify feature importance by examining the support vectors. So in our feature set, there are features which play a crucial role in separating classes such as "A1Cresult", "diabetesMed" which our SVM model can highlight and make use of them. And also

as we have finalized fifteen features to train our model, obviously the high dimensional non-linear complex relationship between the features and the predictor variable which can be done easily using SVM. Because SVMs can be effective in finding a hyperplane that separates different categories.

Though we tried to fit the SVM model with different kernels, the model didn't improve on accuracy. The reason might be due to more categorical features, in which case we should opt for models which best perform with most of the categorical features in the dataset. So with the above observation made, the models which we tried this time are "Random Forest", "XG Boosting", and "Neural Networks". The choice of these models depends on the feature set diversity and type of features.

As we have mostly categorical features, "Random Forest" is used to fit the data set. But random forest gave a test error rate as 51% and train error rate as 22%. Though the random forest is giving high training accuracy, the number of false negatives are high compared to false positives. And in our problem, high number of false negatives is fatal so this model can't be used in practice. So we shifted from "Random forest" to XGBoosting algorithm, which automatically encodes the features in the data set.

Though we used the balanced data set to train the "XG-Boosting" model, we couldn't observe much improvement in the test error rate. The "XGBoosting" algorithm gave the test error as 51% and train error rate as 47%. As the test error rate on the data set remained same and train error rate increased, we can make a note that the model is not learning much. So a model which can learn the complex patterns of the data set is needed. And the model which is best in learning complex features of the data set is Neural Networks.

So to avoid underfitting and overfitting, the model which we choose is a "fully connected Neural Networks". As the most famously used categorical classification models were enabled to capture the complexity of the feature set in our data set, we opted for "Neural networks" which are mostly used to capture the complex feature patterns. After training the neural networks by using the balanced data set, we finally got the test error rate as 50% and train error rate as 42%. The test error rate is decreasing with significant test accuracy to 50%.

### 4.2 Report the results from at least two different models - For each model, report the model's test error. Justify your choice and For each model, discuss how well the model fits the data.

**Random Forest Classifier:** The test error rate given by the Random Forest Classifier(RFC) is 51% with train error rate as 22%. Though the random forest is giving high training accuracy, the number of false negatives are high compared

to false positives. And in our problem, high number of false negatives(as shown in fig:12) is fatal so this model can't be used in practice.

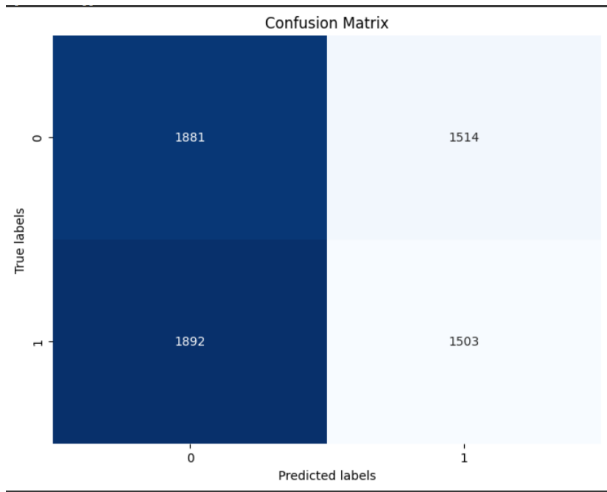


Figure 12: Random Forest Classifier Confusion Matrix

**XGBoosting:** The test error rate given by the XGBoosting algorithm is 0.51 with train error rate as 0.47. On the balanced data set, XGBoosting gave good train accuracy and but test accuracy is not improved. Even though the number of false negatives are less when compared to the number of false positives(as shown in fig: 13), neural networks model should be tried to learn the complex features of the data set.

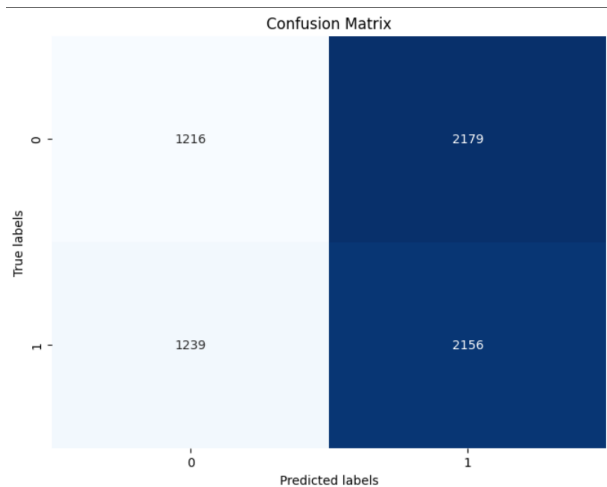


Figure 13: XGBoosting Confusion Matrix

The intuition behind choosing neural networks is in case of random forest and XGBoosting the test accuracy is not improved much. This indicates that the data set might have complex features which are not learnt by basic machine learning models. Whereas Neural Networks is proven to be best in learning complex features in the data set, so we are

proceeding with neural networks.

**Neural Network:** Considering the finalized model which is neural network model with 100 hidden units in the single neural network layer, reported the test accuracy as 0.50 and test error rate as 0.50, indicating that approximately 50% of instances are misclassified. Here are the observations from the Confusion matrix. False Positives (FP): 2065, True Positives (TP): 2082, True Negatives (TN): 1330, and False Negatives (FN): 1313. the model exhibits a reasonable level of misclassification.

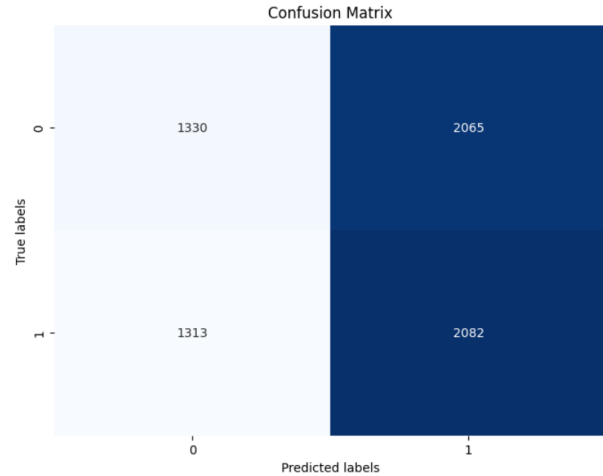


Figure 14: Confusion matrix of the final model - Neural Network

The misclassification and the test error rate of 50% can be attributed to the number of false positives and false negatives predicted by our model. False negatives in our context determine that a person will not get readmitted even though the true chance of getting readmitted within 30 days of discharge is high. So in our case, false negatives are critical and also fatal when compared to false positives. And as false negatives are less compared to the false positives, we can consider that model has performed good enough.

### 4.3 Briefly discuss which model fits the data better

The Neural networks gave the train accuracy as 58% and test accuracy as 0.50%. The train accuracy of 58% indicates that, during training, the neural network was able to correctly predict the target variable for approximately 58% of the training data. The model has learned to some extent from the training data.

Moreover if neural network is evaluated with different machine learning models discussed above for predicting diabetes, neural network model has the lowest count of false negatives as 1313. This suggests that the neural network is performing well in terms of sensitivity or recall. The neural networks



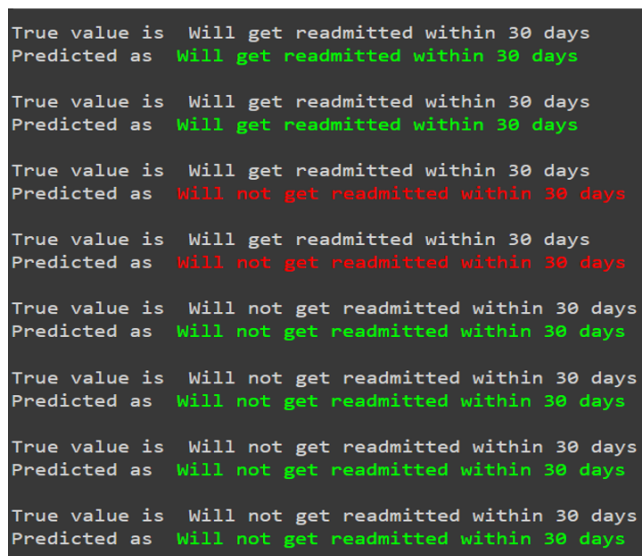
gave the sensitivity as " 0.6133". A sensitivity value of 0.6133 for a neural network model on diabetic data indicates that the model correctly identifies approximately 61.33% of the actual positive cases.

The focus on false negatives is crucial in healthcare scenarios, especially when dealing with diseases like diabetes. False negatives represent instances where the model incorrectly predicts that a patient does not have the condition when, in reality, they do. In the context of healthcare, this can have serious consequences, as it may result in a failure to identify and treat individuals who actually have the disease, leading to delayed or missed medical interventions.

Hence, the neural network is reasonably effective in identifying individuals with diabetes, showing a better-than-random performance. It may be particularly adept at recognizing certain patterns associated with diabetes in the data set.

#### 4.4 For the model that fits the data best, make predictions for at least three cases of interest. One option is to show changes in predicted outcomes for changes in one of the predictors, holding all other predictors constant. Another option is to calculate predicted outcomes for particular cases of interest from the data set, or for hypothetical cases that are of interest

Our best performer is the Neural Network model. The predictions were made for two classes i.e., if the patient is readmitted within 30 days of discharge or the patient doesn't get readmitted within 30 days of discharge. And the prediction results are as shown in figure 15.



True value is	Predicted as
Will get readmitted within 30 days	Will get readmitted within 30 days
Will get readmitted within 30 days	Will get readmitted within 30 days
Will get readmitted within 30 days	Will not get readmitted within 30 days
Will get readmitted within 30 days	Will not get readmitted within 30 days
Will not get readmitted within 30 days	Will not get readmitted within 30 days
Will not get readmitted within 30 days	Will not get readmitted within 30 days
Will not get readmitted within 30 days	Will not get readmitted within 30 days
Will not get readmitted within 30 days	Will not get readmitted within 30 days

Figure 15: Predictions

## 5 Summary and Conclusion

### 5.1 Going back to the question that has motivated your project, how would you answer that question given the results of your analysis?

The motivation behind this project is to answer the question: "Can we construct a reliable predictive model to ascertain the likelihood of hospital readmission within a 30-day period for diabetic patients?". Given the models sensitivity of 61.33%, it can be said that goal is achieved. This metric aligns with the project's goal of minimizing false negatives, which is critical in healthcare scenarios. While other accuracy metrics like test accuracy and train accuracy provide additional insights, the emphasis on sensitivity underscores the model's success in capturing positive instances in the dataset. Further refinements and optimizations may be explored to improve overall performance, but the neural network has shown promise in addressing the specific objective of identifying cases of diabetes.

### 5.2 Think about domain experts in the field you have analyzed. What can they learn from your project? How could the results of your analysis inform their work?

The model, despite having an overall accuracy of 58%, demonstrates a valuable capability in accurately predicting patients who are likely to be readmitted within 30 days based on their medical data. This information holds significant potential for improving patient care and hospital practices.

By leveraging the model's predictions, medical professionals can proactively focus on patients classified as category one – those predicted to be readmitted. This targeted approach allows doctors to implement personalized care plans and interventions for these individuals. For instance, doctors can consider adjusting medications, scheduling timely follow-up appointments, or initiating additional monitoring measures post-discharge.

The predictive insights generated by the model serve as an early warning system, enabling hospitals to take preventive actions and deliver timely interventions. This not only enhances patient outcomes but also contributes to the overall improvement of treatment standards within the healthcare facility.

Furthermore, the model's results empower hospitals to optimize resource allocation by directing attention and resources towards patients with a higher likelihood of readmission. This strategic utilization of healthcare resources not only improves efficiency but also aids in reducing healthcare costs.

In summary, despite the modest overall accuracy, the model's ability to identify patients at risk of readmission



within 30 days provides a valuable tool for healthcare professionals. Integrating these predictive insights into clinical workflows allows for more targeted and proactive patient care, ultimately leading to enhanced treatment outcomes and elevated standards of healthcare delivery.

### 5.3 Identify one way that your project could be improved if you had more time and resources to work on this project. For example, what additional data would you gather? What alternative data cleaning decisions would you make? What additional models would you estimate?

Our project could be improved by working on the noise present in the dataset. Due to the presence noise, the model's test accuracy is not improving much though mostly the data is cleaned and preprocessed. The data cleaning methods for removing noise includes handling rare categories for categorical features, leveraging domain-specific knowledge to identify and handle anomalies or outliers in categorical data. This may involve consulting subject matter experts or using external data sources for validation.

## 6 Dataset Link

**UCI Machine Learning Repository Link:**  
<https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>

## 7 Contributions

**Sandya Rani Prasad** - Performed exploratory data analysis, data visualization and cleaning of the data set. EDA allows us to gain a comprehensive understanding of the dataset, including its structure, distribution, and key characteristics. This understanding is essential for making informed decisions throughout the analysis. A clean and well-understood dataset is developed by Sandya which is essential for building accurate and effective models. Moreover, sandya contributed to writing this document by providing content for the sections - "Introduction", and "Summary of EDA".

**Venkata Subbu Sai Hiranmayee Machavolu** - Worked on the initial model which is Support Vector Machine. Working on the initial model, a Support Vector Machine (SVM), has provided valuable insights into the characteristics of the dataset, particularly revealing the presence of imbalanced data. After finding that the data set is imbalanced, Hiranmayee made the data set balanced, and tried to tweak the SVM model parameters to fit SVM on the balanced data set and improve accuracy of the model. Recognizing this imbalance is crucial

as it can impact the model's performance and the reliability of its predictions.

**Charan Basireddy** - Worked on XGBoosting, Random Forest classifier. Investigating the dataset using advanced machine learning models such as XGBoosting and Random Forest classifier has provided valuable insights into the complexity inherent in the data. The results indicate that the features within the dataset possess intricate relationships and patterns that go beyond the capabilities of basic machine learning models. Additionally, Charan contributed to the documentation by adding content to few sub sections of "Summary of machine learning models" section.

**Lakshmi Gayathri Rangaraju** - Worked on final model which is neural network. In the pursuit of an optimal predictive model, extensive experimentation and fine-tuning were conducted with a neural network architecture, marking the final stage of the modeling process. Notably, a multifaceted approach was adopted, encompassing data augmentation, feature engineering, and meticulous tuning of learnable parameters to enhance the overall accuracy of the neural network. Furthermore Lakshmi contributed to the sub sections of "Summary of machine learning models" and to the "Summary and Conclusion" section.

## 8 References:

- <https://www.hindawi.com/journals/bmri/2014/781670/>