Predicting Readmission of Hospitalized Diabetic Patients

Nathan Stangler (stang451), Jinming Chen (chen6386), Mark Frenz (frenz079), Vivian Tsang (tsang065), Thomas Knickerbocker (knick073)

Introduction: Literature review

Hospital readmissions are a metric used to evaluate the quality and effectiveness of care provided by a hospital. There are several time periods used when talking about readmission rate, commonly 30 and 90 days, as well as 1 year. They are defined as the hospital admission subsequent to the first hospital stay, or the *index admission.* If a hospital has a high rate of readmission, it can be an indicator of low quality care, or failure to educate the patient on how to manage their transition from the hospital back to their home. Reduction of the readmission rate is important to both hospitals and patients, as hospitals are able to use resources more effectively and focus on new patients while patients receive better quality care, increasing their confidence in the healthcare system and reducing the emotional and physical impact of the hospital stay (Dhaliwal & Dang, 2024), (Spanakis et al., 2019). In particular, patients admitted with diabetes have been observed to be readmitted with a higher rate than patients admitted for other reasons. Additionally, studies have shown that diabetes specifically increases the risk of readmission within 30 days by at least 17% (Rubin & Shah, 2021). There are a wide range of reasons patients with diabetes are readmitted. One study found that key factors for readmission were poor glycemic control, comorbidities, and limited diabetes specific inpatient consults and follow-up care. For cases of secondary diabetes cases, or cases where the main reason of admission wasn't diabetes, infections were the main cause of readmission (Ostling et al., 2017). Another study found that hospital admission due to severe dysglycemia, or abnormal blood sugar levels, was a strong predictor for readmission with another dysglycemia episode. However, this can be reduced with outpatient care specific for diabetes, even if the primary reason for admission was something else (Zimmerman, 2017). This further highlights the need for understanding hospital readmission for patients with diabetes so as to know the causes and subsequently the preventions for this happening.

We use the Diabetes 130-US Hospitals for Years 1999-2008, details of which will be covered in the project introduction. This dataset has been used in several other studies, including the *Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes* (Liu et al., 2024). This study aimed to compare different deep learning and machine learning methods for predicting diabetes readmission. They applied traditional ML models supplemented with Grey Wolf Optimizer (GWO) for feature selection, and they found that out of all the methods they tested, the random forest model consistently outperformed all the others, with XGBoost coming in second place. To address the class imbalance problem of this dataset, they used the Synthetic minority over-sampling technique, or SMOTE (Chawla et al., 2002). This technique helps combat a class imbalance by reducing sampling from the majority class, and generating synthetic examples of the minority class through interpolation over existing data points. This data is then added to the training to make the class sizes more evenly distributed. These findings also aligned with those of another study, which used the same dataset and had a similar setup, using GWO as their feature selector and SMOTE as for the class imbalance problem; they also found that the random forest method produced best results. Overall, readmission of patients to hospitals is a significant challenge for the healthcare industry, particularly in patients with diabetes. High readmission rates are responsible for increased healthcare costs, healthcare delays, and greater risk for the patient's health. Hospital readmissions among patients with diabetes are often associated with greater health risks and complications compared to readmissions in the general population, thereby necessitating robust methods for accurately predicting readmissions. ML models have shown promising results in predicting readmission for diabetic patients, and we will be exploring this topic more in this project.

Introduction: Project

This report analyzes the **Diabetes 130-US Hospitals for Years 1999-2008** dataset (Clore et al., 2014), which contains hospital records of 101,766 patients who were admitted with diabetes over a ten year period. Each record represents a patient that was treated by the hospital with a stay of up to 14 days. In each instance, medication was administered and lab tests were performed. There are over 50 features including features describing personal information about the patients, medications that were administered, number of lab tests performed, and more. This problem is important because it allows medical professionals to study the effectiveness of their treatments and the specific factors that could lead to a better outcome. Hospital visits are costly and it is in the best interest of both the patient and hospital to minimize the chances of

readmission. The main goal of this report is to predict whether a patient with diabetes will be readmitted within 30 days of discharge.

To achieve the main goal, the dataset will be analyzed using three different machine learning models. The three machine learning models are logistic regression, random forest, and a neural network. All of the models will attempt to predict whether or not a patient was readmitted to the hospital using the remaining features as predictors. By analyzing the model's performance, hospitals and medical providers can see what treatments work better for which types of patients and vice versa, and they can apply this knowledge to improve the quality of care for future patients.

Data Preprocessing:

The original dataset contained a significant amount of useful information to achieve the goal of gaining knowledge on what leads to readmission. However, the original dataset contained a number of areas that required preprocessing for it to be usable. One of the major issues with the dataset is a significant amount of missing values. The first step in processing the dataset was to handle missing values and remove multiple columns. The original dataset used question marks to represent missing values, which caused certain functions in python to treat them like actual values instead of missing data. In order to treat the missing values properly, they were all replaced with the numpy not a number value. The columns patient_nbr and encounter_id were dropped due to them being unique identifiers for the specific data entries, which caused them to not be useful for predictions. Alongside dropping the unique identifiers, the weight column was also dropped. The weight column was dropped due to only 3,197 out of the 101,766 patients having a recorded weight. Since the proportion of patients missing the weight value was very high, it made it not reliable to fill in missing values. Additionally, the patients who died during their initial hospitalization were dropped from the dataset since it is not practical to predict if they were readmitted.

The original dataset contained two different feature types, categorical and integer. The integer feature types were already in a form that makes it easy for the models to read, but the categorical data needed to be converted into a form that the models could use. The age column was not represented as the patient's exact age, but instead it was a categorical value which was in an interval form that would group 10 years of age together from 0 through 100. The age interval causes the value to be less precise, which could slightly increase the error in prediction when using this variable. Initially, the age categories were replaced with an index value from

1-10, but to capture the separation between the groups more for some of the models, the age categories were replaced with the center value of the age interval.

The original dataset contained the features diag_1, diag_2, and diag_3, which were treated as categorical variables. The feature diag_1 is the primary diagnosis coded as the first three digits of ICD9, and diag_2/diag_3 are the secondary diagnosis also with the ICD9 coding. These features are mostly integer values with some containing an alphabetic letter that represents a more specific diagnosis. The diagnoses without a letter are able to be represented by the models, but the ones with a letter cause issues. The ICD9 format uses integers 0-999 with some having an E or V before the number. The ICD9 values without a letter were not changed, but the values with a letter had the letter removed and a constant added to the number. If the ICD9 value included an E, it had 1000 added to the value, and values with a V had 2000 added to the value. By adding a constant value to the numeric part of the ICD9 values that include a letter, it allowed the models to work without dropping any entries.

The payer_code and medical_specialty columns were both represented categorically, each with many different possible values. Both of the features had a substantial amount of missing values, which increased the difficulty of considering how to handle the data. In order to handle the missing values, they were initially filled with a string to represent it being unknown. The different categories were then converted into a one-hot encoded column for each category. The one-hot encoded categories allows the model to consider each category separately, instead of different methods like assigning an integer code, which could cause the models to group similar integers together. The gender, race, admission_type_id, discharge_disposition_id, and admission_source_id features were also handled the same way.

The original dataset included 23 different prescription features that the patients could have been prescripted at the hospital visit. The prescription features have a value of one of four categorical values that states whether the dosage of the prescription was increased, decreased, remained the same, or was not prescribed at the visit. Each of the prescription features were looped through with the categorical dosage being covered to 0 if the prescription was not prescribed, 1 if the dosage was increased, 2 if the dosage was decreased, and 3 if the dosage remained the same. Additionally, two of the prescriptions did not have any rows after other preprocessing, so examide and citoglipton were dropped.

The readmitted feature is used as the label for predicting whether a treatment was good or not, but the original dataset had three possible categorical values. The three categories

represented whether the patient was not readmitted, readmitted in less than 30 days, and readmitted in over 30 days. In order to simplify the models, the feature was converted to 0 if the patient was not readmitted within 30 days and 1 if the patient was readmitted in less than 30 days.

There were a number of other features that required similar conversion from categorical representation to integer representation. The other columns that were converted are gender, race, change, diabetesMed, max_glu_serum, and A1Cresult. Each of these features were converted in a similar way with unknown or unmeasured values being given an integer value that represented unknown values. Following the dropping of features and conversion of categorical features, the data was converted into only numeric form. All rows that still contained missing values or invalid data were dropped, and the indices of the dataset were reset. After all of these steps were completed, the data was properly represented for all of the models to use.

Following the dataset being processed into the proper form, the preprocessed dataset needed to be formatted to be used by the models. The models predict whether or not the patient was readmitted, so the label is either 0 or 1, with 0 being that the patient was not readmitted and 1 if the patient was readmitted. The labels are the column vector Y, which is just the readmitted column. The predictors that the models train on are all of the remaining feature columns, which is the matrix X. The feature columns that train the models are in different scales, which can cause the model to not accurately predict the label. In order to prevent the issue that the different scales create, the values in all columns are separately scaled to a standard normal distribution. The scaling causes each column to have a mean of 0 and a variance of 1, which allows the models to predict the label based on the predictors that are in the same scale.

The data was then split into a training set and test set each with a predictors matrix and a label vector. The training set was 80% of the data, and the testing set was 20% of the data. However, the data was significantly imbalanced. The imbalanced data was not an issue for the testing set, but it did cause issues for the training set. The training set contained only around 11% of entries that had patients that were readmitted, which caused the models to favor predicting that patients were not readmitted rather than learning the features. In order to prevent the models from favoring the most common label, a few different methods of balancing the training data were tried. One of the methods tried was balancing using SMOTE (Chawla et al., 2002), which balances the data by generating samples of the class with fewer entries. After using SMOTE to balance the data, the training set contained the same amount of entries where

the patient was readmitted as where the patient was not readmitted. The second method was using balanced class weights parameters in the models, which handles the imbalance in the training process. The balanced training set using SMOTE improved the models slightly, but the models still were favoring the class with more real entries. Using balanced class weight parameters causes the models to focus on predicting based on the features instead of the imbalance, so this method was used for all of the models.

The logistic regression and random forest models are able to use the numpy training and test sets directly, but the neural network requires tensors. The predictors matrix and labels vector for both the training and test set were converted into PyTorch tensors and input into a tensor dataset, which were then used in data loaders. The data loaders used batch sizes of 64 for training and test data. The training data loader shuffled the data, which prevented batches being the same each epoch.

Machine Learning Models Implementation:

The logistic regression model is implemented using sklearn's LogisticRegression. The model fits the training data using balanced class weights due to the data imbalance. The model then predicts the readmitted label for all of the test data, which is then used to analyze the model. The logistic regression model is mainly used as a baseline model to help understand how the other models perform.

The random forest model is implemented similarly to the logistic regression model. The model is fit to the training data using balanced class weights then is used to predict the readmitted label for all of the test data. We set up the tree with 50, 100, or 200 estimators to capture more features, a limited depth to a maximum of none, 5, or 10 to prevent overfitting, and with 1, 2, or 5 minimum sample leaves to predict minorities. The random forest model uses 5-fold cross validation to find a good number of decision trees to use, maximum depth, and minimum sample leaves. The cross validation method used was grid search and the best value for the parameters was selected from the previous possible amounts based on the parameters that give the best balanced accuracy.

The neural network is implemented using a PyTorch network to train and evaluate the model. The network has a linear layer with 128 output features, which is then passed through a one dimension batch normalization layer. The output of the batch normalization is then passed

through to a ReLu, and a dropout layer with a conservative probability of 0.3. The output of this stage is then passed through a similar process with a linear layer with 64 output features, a batch normalization, ReLu, and a dropout layer. The output layer has a linear layer that has 2 out features, which relate to the two classes of readmission. We use batch normalization to soften the gradient and stabilize the results. We use Dropout to enhance the model's robustness and prevent overfitting (Hinton et al., 2012). The model is trained using the Adam optimizer with a learning rate of 0.0001. The model uses a cross entropy loss function class weights that are balanced based on the amount of patients in each readmitted class. The neural network is trained for 25 epochs with its performance calculated for each epoch.

The performance of the models is analyzed using the same four metrics, test accuracy, AUC, recall, and a confusion matrix. The main goal of this report is to predict the readmission of patients with diabetes, so the recall was the main metric that was focused on. The recall was the main metric because it is the proportion of true positives that were correctly predicted, which is the proportion of readmission correctly predicted. The recall and accuracy have a trade off where the recall could reach its maximum of 100% if it were always predicted that a patient was readmitted, but it would cause a low accuracy due to the majority not readmitted class being entirely mislabeled. The balance between accuracy and recall would be best chosen based on the cost of mislabeling each class, but for this report it was assumed that the accuracy should be around 65%.

Analysis of Machine Learning Models:

The three different machine learning models provided information about the prediction of readmission for diabetic patients. The metrics for test accuracy, AUC, and recall gave comparable results to understand each model's performance. The results of the linear regression model and the random forest model can be seen in Table A.

Metric	Logistic Regression		Random Forest	
Accuracy	0.674426		0.688422	
AUC	0.615903		0.617485	
Recall	0.539885		0.525342	
Confusion Matrix	12074	5376	12383	5067
	1044	1225	1077	1192

Table A - Metrics of ML Models

The metrics were analyzed for if their ability to predict whether a patient was readmitted was greater than randomly picking one class or always picking the majority class. The accuracy of the models would need to be greater than 50% in order to be better than randomly guessing at risk patients if the data was balanced. However, the data was significantly imbalanced with a much smaller percentage of the data being readmitted, which could lead to a high accuracy by just picking the most common class. The AUC and recall metrics are analyzed to help determine if the model is picking the most common class.

The logistic regression model was used as a baseline to compare the other models. The model gave an initial understanding of the dataset and the area of expected performance. The model was able to predict readmission with an accuracy and AUC of over 0.6, which suggests that the features are correlated to a patient with diabetes being readmitted to the hospital. The recall of the model was 0.539, which means that over half of the patients readmitted were correctly predicted. However, the model incorrectly predicts 27% of patients as being readmitted.

The logistic regression model was also used as a baseline because it is able to analyze which features had the most impact on the prediction. The ten features that have the most predictive ability and their coefficients are shown in Table B. The feature number_inpatient, which is the number of inpatient visits of the patient in the year preceding the encounter, had the largest impact on predicting readmission. The number of visits a patient had in the previous year suggests they are likely frequently admitted, so it is likely they will be readmitted again. The features for discharge_disposition id 22, 3,5, and 2 were also important, which represent discharges to another hospital or facility with skilled nursing. Patients discharged to specialized care facilities likely have recurring severe health issues and will be readmitted. The feature payer_code_Unknown also had a large impact on the prediction, which could be missing

Feature	Coefficient
number_inpatient	0.358489
discharge_disposition_id_22	0.190960
discharge_disposition_id_3	0.156484
payer_code_Unknown	0.114953
medical_specialty_Gynecology	-0.113043
discharge_disposition_id_5	0.111317
medical_specialty_Pediatrics-Endocrinology	-0.100928
diabetesMed	0.096208
discharge_disposition_id_2	0.092068
tolbutamide	-0.084718

Table B - Top 10 Logistic Regression Coefficients

values, either from a hospital not listing the insurance or an uncategorized insurance, or uninsured patients. The large positive coefficient suggests that patients with less common insurances or without any insurance are more likely to be readmitted. Patients admitted by medical specialists that treat conditions unrelated to diabetes symptoms are less likely to be readmitted likely due to their initial hospital visit being unrelated. Patients prescribed diabetes medication increases the likelihood of readmission likely due to their condition being worse, but specific medications like tolbutamide can decrease the likelihood of readmission.

The random forest model was found to give the best performance with 200 decision trees a maximum depth of 10, and minimum sample leaves of 5, which was found using the cross validation. Figure A plots the different parameters that were selected through cross validation against the mean balanced accuracy of the forests fit with that parameter. The three plots for the three parameters all show that as the value of the parameter increases, the mean balanced accuracy increases. The plots show why the best parameters were selected to be the largest option due to the mean balanced accuracy being the largest when the parameter is the largest. The parameter value options were limited to the chosen options to fit the assumed accuracy of 0.65 so that the recall could be maximized.

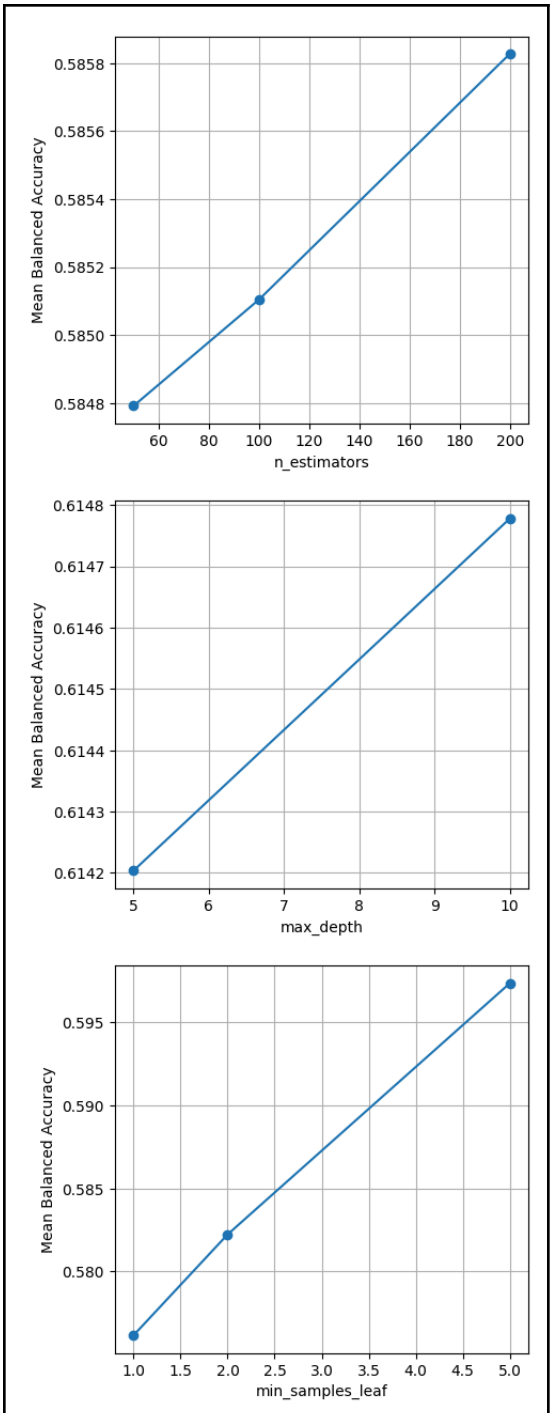

Figure A - Cross Validation Visualization

The random forest model had a higher testing accuracy and AUC than the logistic regression model, but with a recall. The recall of the model was 0.525, which means 52.5% of

the patients readmitted were correctly predicted. The confusion matrix shows that the model incorrectly predicts fewer non-readmitted patients as readmitted than logistic regression, which causes a lower recall but higher accuracy.

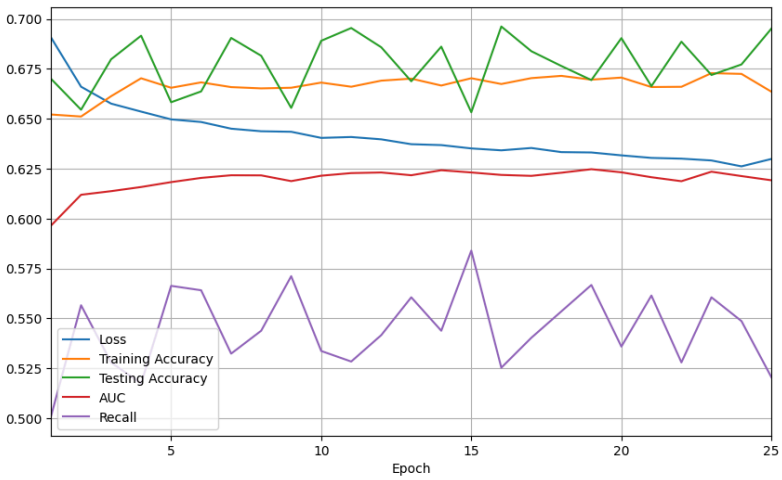

Figure B - NN Metrics Visualization

The neural network model has 25 epochs with each epoch having an accuracy, AUC, recall, and confusion matrix. However, the neural network also has a training accuracy and a training loss value. The training loss, training accuracy, testing accuracy, AUC, and recall are graphed in Figure B. The training accuracy, testing accuracy, and recall significantly increases and decreases between epochs. However, the overall accuracy increases as the epochs increase, and recall decreases as the epochs increase. The decreasing recall as the accuracy increases suggests that the network over trains towards the majority class, which causes the earlier epochs to be better trained for recall, and the increasing accuracy is likely fitting towards the percent of patients not readmitted.

Metric	Value	
Training Loss	0.635163	
Training Accuracy	0.670327	
Testing Accuracy	0.653279	
AUC	0.623125	
Recall	0.583958	
Confusion Matrix	11557	5893
	944	1325

Table C - Metrics of Neural Network

The metrics from the 15th epoch are shown in Table C, which is the epoch that had the best recall. The neural network had a worse testing accuracy than the logistic regression model, but it had a slightly larger AUC. The recall of the neural network was around 0.04 larger than the logistic regression model, which means it predicts an additional 4% of readmitted patients correctly. The

model incorrectly predicts 30% of patients as being readmitted compared to the logistic regression model incorrectly predicting 27% of patients as being readmitted. The additional 3% incorrect predicting of patients being readmitted gains an extra 4% of correct readmission predictions, although the 4% of correct readmission predictions is a much smaller number of patients. Depending on the cost of incorrectly predicting patients being readmitted against incorrectly predicting patients not being readmitted, this could be an acceptable difference. An AUC of 0.623 implies the neural network correctly ranks a readmitted patient higher than a non-readmitted one only 62.3% of the time, which is slightly better than a random model. This could be because the inner complexity of our dataset is high, and the model is having difficulty distinguishing between classes.

Conclusion:

This report studied the use of machine learning models to predict the readmission within 30 days of hospitalized diabetic patients using a dataset of 130 hospitals with over 100,000 entries. The dataset was preprocessed to handle missing values and incompatible data types for use in a logistic regression model, random forest model, and a neural network. The performance of the models for predicting the readmission of hospitalized diabetic patients shows their possible use to reduce readmission.

The logistic regression model was shown to be a solid baseline by performing better than chance in all metrics, and being able to correctly predict over half of the readmitted patients. The random forest model achieved the highest accuracy but had the lowest recall of 52.5%, suggesting that the model fit toward the majority class more than the other models. The neural network was able to correctly predict the highest percentage of readmitted patients with a recall of 58%, but with a slightly lower accuracy than the baseline logistic regression model.

The likelihood of readmission of patients was shown to be increased factors including having had many previous admissions, poor insurance, and discharge to specialty care facilities. Readmission was decreased by factors like the initial visit being with an unrelated medical specialist, and certain prescriptions like tolbutamide. The performance of the models shows that the features are correlated with the likelihood of a patient being readmitted due to the improved ability of identifying readmissions compared to randomly guessing. However, the rather low recall and accuracy suggests that there will still be a large error in prediction. The use of more data features could assist with improving the predictions and interpretations, and would be a good next step for continued work.

While the readmission of diabetic patients cannot always be correctly predicted, hospitals have the required data to begin identifying the most at risk patients. The performance of the models shows that their use has the possibility of decreasing the cost to hospitals and patients by reducing readmission, and improving the quality of care for at-risk patients. The implementation of the concepts discussed in this report are important for improving the care for diabetic patients.

References:

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). Diabetes 130-US hospitals for years 1999–2008 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5230J

Dhaliwal, J. S., & Dang, A. K. (2024). Reducing hospital readmissions. In StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK606114/

Hinton, Geoffrey E., et al. "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors." ArXiv:1207.0580 [Cs], 3 July 2012, arxiv.org/abs/1207.0580.

Liu, V., Sue, L., & Wu, Y. (2024). Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes. Journal of Medical Artificial Intelligence, 7. https://doi.org/10.21037/jmai-24-70

Ostling, S., Wyckoff, J., Ciarkowski, S. L., Donihi, A. C., & Pfeifer, L. (2017). The relationship between diabetes mellitus and 30-day readmission rates. Clinical Diabetes and Endocrinology, 3(3). https://doi.org/10.1186/s40842-016-0040-x

Rubin, D. J., & Shah, A. A. (2021). Predicting and preventing acute care re-utilization by patients with diabetes. Current Diabetes Reports, 21(9), 34. https://doi.org/10.1007/s11892-021-01402-7

Spanakis, E. K., Umpierrez, G. E., Siddiqui, T., Zhan, M., Snitker, S., Fink, J. C., & Sorkin, J. D. (2019). Association of glucose concentrations at hospital discharge with readmissions and mortality: A nationwide cohort study. The Journal of Clinical Endocrinology & Metabolism, 104(7), 2775–2785. https://doi.org/10.1210/jc.2018-02575

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, Article 781670. https://doi.org/10.1155/2014/781670

Zimmermann, E. (2018, June 21). Diabetes complications are a risk factor for repeat hospitalizations, study shows. Mayo Clinic News Network. https://newsnetwork.mayoclinic.org/discussion/diabetes-complications-are-a-risk-factor-for-repeat-hospitalizations-study-shows/

Code:

https://colab.research.google.com/drive/13NXxOZQvRAjo3uKC8G5Tng567uKa3-kP