Analysis of Hospitalized Diabetes Patients

between 1999-2008

Cory Chitwood, Eddy Doering, Kai Gui, Keith-Jordan Wilkinson

Section 1: Introduction

As of 2020, diabetes is a disease that afflicts approximately 10% of the US population and is the 8th leading cause of death.[1, 2] It is not uncommon for patients to be hospitalized due to diabetes or reasons relating to diabetes. Ideally upon discharge from the hospital, a patient would have their major symptoms addressed and have their disease under control. However, a few patients will have to be readmitted due to continuing complications, which is associated with unfavorable patient outcomes and high financial costs.[3] In our project, we utilize a dataset of approximately 100,000 anonymized electronic health records (EHR) from hospitalizations of diabetes patients from 1999-2008 compiled by Virginia Commonwealth University researchers from the CERNER Health Facts Database.[4] We first explore and visualize the demographics of diabetes patients represented within the dataset, comparing our findings to 2008 national demographics obtained from census data.[5] Next, we apply machine learning models to predict if hospitalized diabetes patients will be readmitted based on features of their original hospital stay. Predictions like these could improve patient outcomes by empowering primary healthcare workers to better identify and monitor at-risk patients.

Exploratory Questions:

- What is the most common age range for diabetes patients?
- How does the number of patients readmitted to the hospital change with age?
- What are common diagnoses with diabetes patients?
- How does time spent hospitalized change with patients age?
- Are there any racial disparities in hospital visits/ diabetes patients vs. national demographics?
- What features can predict readmittance rates?

Section 2: Data Exploration

First to learn more about our dataset, we needed to learn more about each individual patient and identify what conclusions we could draw from examining any preliminary patterns.

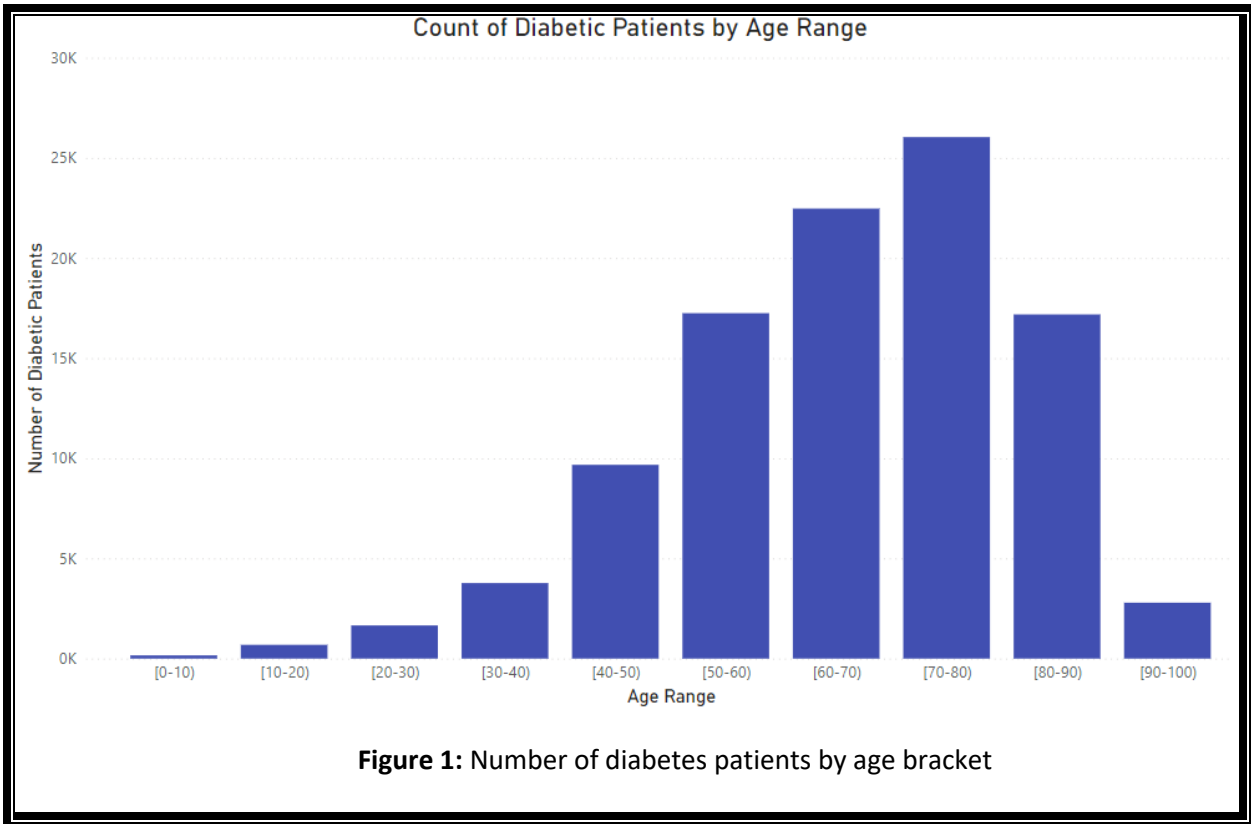

Figure 1: Number of diabetes patients by age bracket

Above is the first exploration into the age breakout groups showing what age ranges are represented most often in hospitalized diabetes patients. As shown, there is a steady increase in the number of patients beginning at age 40. This observation is backed by the article, "Age of Onset for Type 2 Diabetes" by Kristeen Cherley, where it is reported that the age range which records most new diabetes cases is between 45-65.[6] The number of patients peaks at the 70-to-80-year range, which is to be expected given the relationship of old age and medical issue. Additionally, given the average life span of a diabetic American is approximately 74.5 years-old, the drop of hospitalized patients in the 80-90 and 90-100 range is expected.[7]

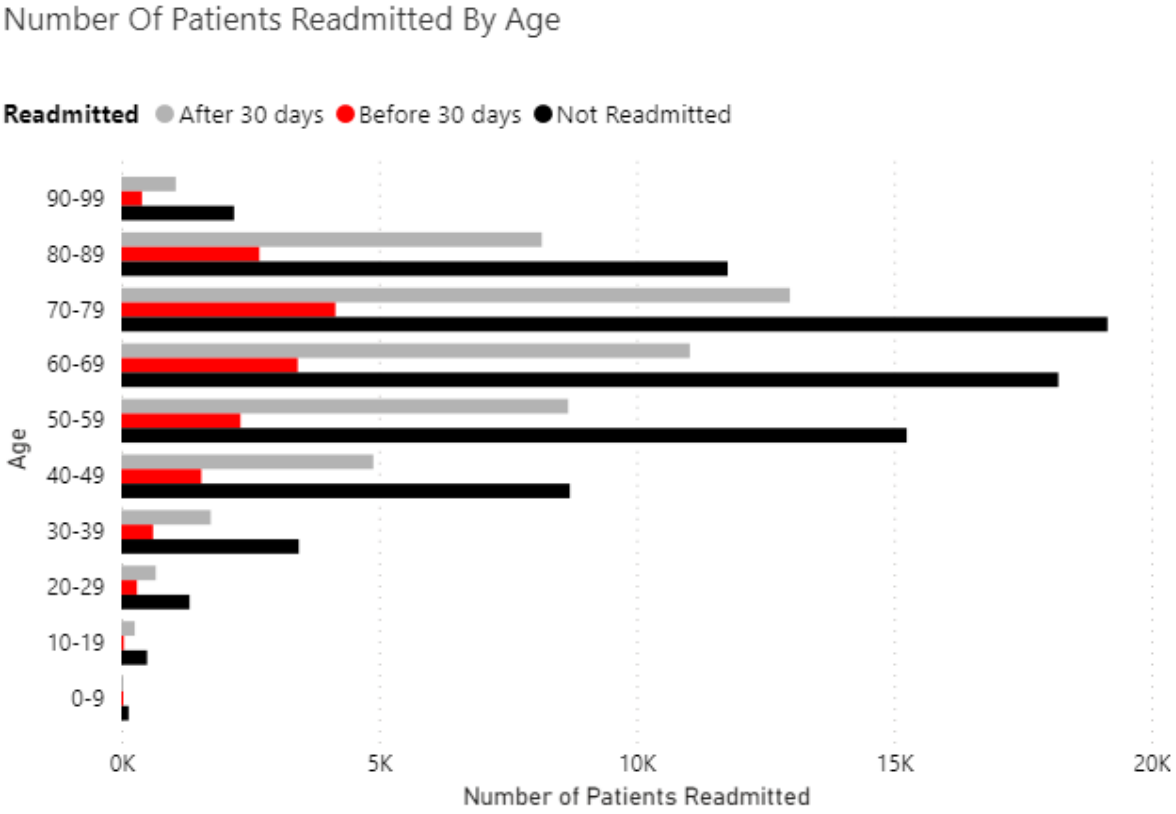

Figure 2: Readmittance status of diabetes patients by age bracket

Figure 2 shows the number of patents readmitted to hospitals within 30 days, after 30 days or not readmitted at all, and broken out by patient age range, with regards to readmittance status, for each age group in the data set, it is much more likely for patients not to be readmitted after being in the hospital. Also, with each age group, if patients were readmitted, more patients are readmitted after 30 days than before 30 days.

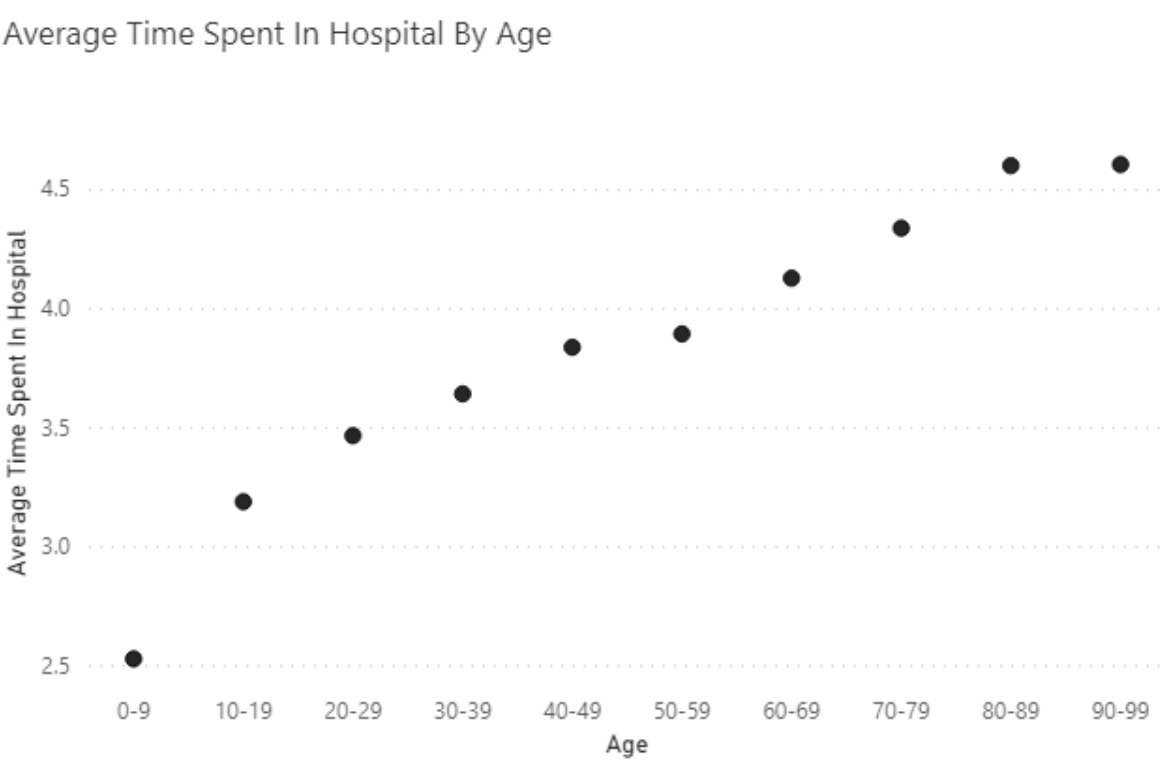

Figure 3: Average hospitalization time by age bracket

As shown in figure 3, we see clearly that the average hospitalization duration increases gradually with age. From ages 0-50 we can see that the difference in the average times is very gradual and almost plateaus as it approaches 50. Between the ages of 50 onwards a similar type of increase is seen which appears to plateau as the age range approaches 90-99. The gradual increase in time spent in hospital is not surprising. Younger individuals can recover from surgical procedures and diseases faster than older individuals. As a result, a younger individual would potentially have faster recovery times and would require less time in the hospital. Also, with age, one becomes more susceptible to complications with surgical procedures and long-term health issues. Hence, there would be more cause for an older individual to spend time in the hospital, explaining the results seen.

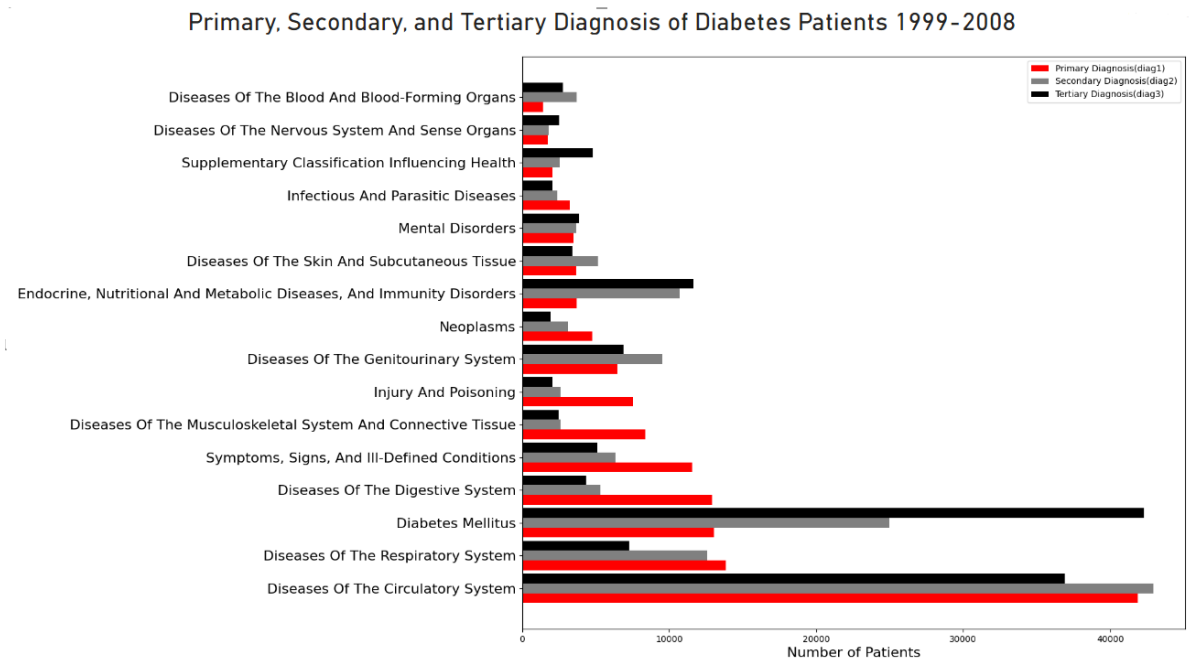

Figure 4: Diagnoses of Diabetes Patients

Figure 4 shows the most common diagnoses of hospitalized diabetes patients. In our dataset three diagnoses were tracked, a primary diagnosis, a secondary diagnosis, and a third or tertiary diagnosis. We explored the relation between these diagnoses and the number of patients of each diagnosis. The largest diagnoses within the dataset are diseases of the respiratory and circulatory system and diabetes mellitus. It is interesting to note that there diabetes mellitus is most common as a tertiary diagnosis which could be explained by patients being hospitalized for complications from their diabetes, such as diseases of the respiratory system or circulatory system.

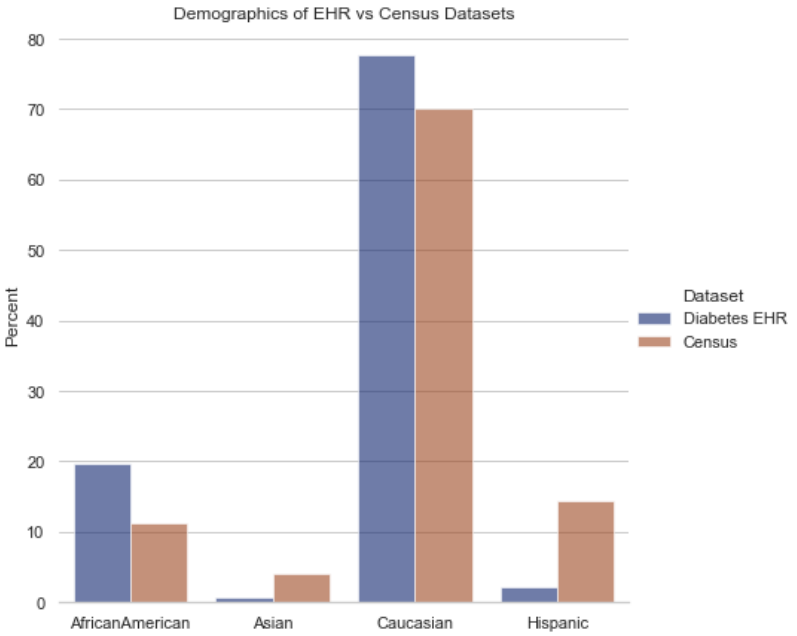

Figure 5: Racial Demographics of Diabetes Patients vs Census Demographics by Percent

In Figure 5, The racial breakdown of hospitalized diabetes patients is plotted against the approximate racial breakdown of the US in 2008. It is observed that African American and Caucasian groups have higher representation in the hospitalized patients than one might expect based on the population size of those groups. Likewise, Asian and Hispanic groups have lower representation in the hospitalization dataset than what the census would suggest. This finding demonstrates that hospitalization rates may be influenced by external factors in addition to population size.

To explore possible reasons for why different racial groups are hospitalized for diabetes at different rates, a census dataset examining the uninsured rate of different racial groups was utilized. We hypothesize that hospitalization rates for a racial group may have an inverse relationship with a group's uninsured rate. In general, a racial group that has a high uninsured rate may avoid hospitalization due to high financial burden. To test this hypothesis, the normalized ratio of hospitalized diabetes patients against US population was

plotted for each racial group. A ratio over "1" indicates when a racial group is overly represented in hospitalized diabetes patients based on population size, while a ratio below "1" indicates the inverse. Next the uninsured rate was normalized against the uninsured rate for the entire US and plotted. In the resulting visualization, figure 6, we see that our hypothesis appears to be debunked. Hispanic and Caucasian groups fit our hypothesis of inverse relationship between uninsured rate and hospital admittance, but African American and Asian groups do not.

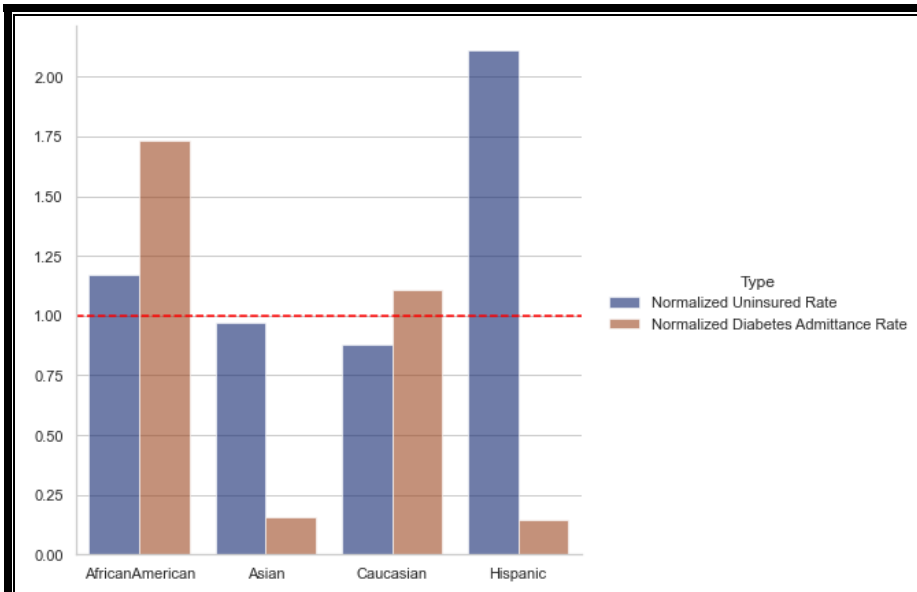

Figure 6: Normalized Uninsured Rate vs Normalized Diabetes Admittance Rate

For additional exploration, it may be productive to explore other explanations for the demographic disparity in hospitalized diabetes patients. The CDC 2020 National Diabetes Statistics Report finds that people of Hispanic origin have the second highest rate of diagnoses for diabetes at 12.5%, while non-Hispanic white individuals have the lowest rate of diagnoses for diabetes at 7.5%, showing that the discrepancy in hospitalizations is not because Hispanic or Asian groups are less susceptible to diabetes than Caucasian groups. An alternative explanation may be that Hispanic and Asian groups may have a larger proportion of recent immigrants that may not be comfortable navigating the US Hospital system due to language or cultural barriers.

If this were the case, the low rate of hospitalization for Asians and Hispanics may indicate a failure of healthcare access that may be remedied with services such as interpreter access or printing hospital correspondence in additional languages.

Section 3: Data Analysis and Machine Learning

The primary analysis of the diabetes EHR dataset was to train a machine learning model to predict whether a diabetes patient would be readmitted to the hospital based on the features of their initial hospital stay. The dataset provided multiple values to indicate "readmitted" status, including "readmitted less than 30 days", "readmitted after 30 days", and "not readmitted at all". With this task in mind, we initially focused our attention on common classification algorithms including Random Forests Classifier, K-Nearest Neighbors Classifier, Ridge Classifier, and SVM Classifier. However, based on the size of our dataset, SVM optimization proved to require too much processing power, so SGD optimization was used as an alternative model.

Model	Base Accuracy	Optimized Accuracy	Improvement	Optimized AUC	Approx. Runtime (min.)
Random Forest Classifier	0.58421	0.65523	12.156%	0.72	5
Ridge Classifier	0.57886	0.65068	12.406%	0.71	2
SGD Classifier	0.54353	0.63233	16.339%	0.69	15
KNN Classifier	0.49331	0.63648	29.023%	0.62	25
Ensemble Voting Classifier	N/A	0.65175	N/A	N/A	60

Table 1: Summary of Optimized Machine Learning Models

In initial attempts of model fitting, it became apparent that predicting short term hospital readmittance within 30 days is challenging due to the small number of data points. Figure 2 shows that very few hospital visits result in readmittance within 30 days, resulting in models that very rarely make predictions for short term hospital readmittance. Although the model would consistently miss predictions for short-term readmittance, the overall accuracy score remained high due to the small number of short-term readmittance in our dataset. Despite being accurate, a model that always predicts "No" is not particularly useful. To

address these issues, we decided to shift our goal to predict whether a hospital visit would result in readmittance, both before and after 30 days. Refactoring our dataset in this manner led to a more equal distribution of readmitted statuses that was easier to optimize our models around. Initial attempts at model fitting appeared promising, and we turned our models for better performance (Table 1).

Tuning Parameters:

- Ridge Classifier was tuned with GridSearchCV, scanning alphas from 0.1-100. Ultimately, an alpha of 70 provided the highest performance.
- Random Forest Classifier was tuned with GridSearchCV, setting n_estimators to 500 and scanning parameters criterion ("gini" or "entropy"), and "max_features" ("sqrt", 0.2-0.5). Ultimately, criterion = "gini" and max_features = "sqrt" provided the best combination of performance and low runtime.
- SGD Classifier was tuned with GridSearchCV. The alpha values tested were: 0.0001, 0.001, 0.01, .1, 1, 10 and 100. The number of iterations tested (n_iter_n_change) were: 1, 5 and 10. The penalty values tested were l1 and l2. Lastly, the optimal 'loss' value was tested using 'hinge' and 'log'. The combination that provided the best score was alpha= 0.0001, n_iter_n_change = 5, penalty = l1, loss = log.
- KNN classifier was tuned by varying the leaf size, number of nearest neighbors, and the p-value, or the "Power parameter for the Minkowski metric Tuning." The leaf size was varied from 0 to 30, the number of nearest neighbors was varied from 1 to 50, and the p value was either 1 or 2. This was a daunting task because the KNN classifier took 5 mins to run and with 3000 different variations of hyperparameters being tested the task would have taken 250 days to complete. Instead, we change the number of cores and used a test set of the data to make the model creation process faster and observed an increase in the mean accuracy score with a leaf size= 1, p= 1, n_neighbors=11.

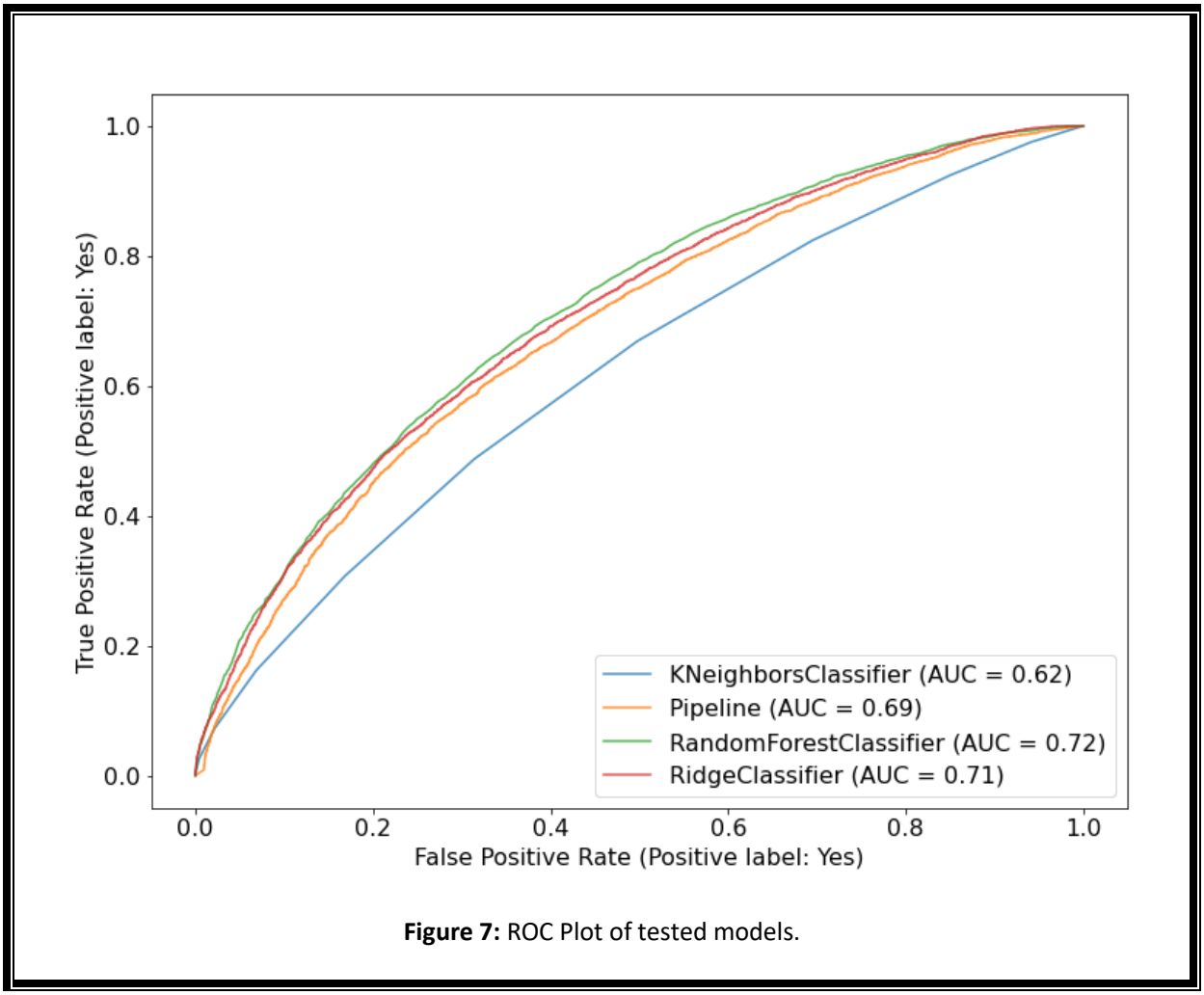

Figure 7: ROC Plot of tested models.

With optimized models in hand, comparisons of model accuracy could be made (Table 1). Random Forest Classifier and Ridge Classifier both yielded strong results with relatively fast runtimes. Random Forrest Classifier also shows the strongest performance in the ROC plot, with an Area Under the Curve (AUC) of 0.72 (Figure 7). We hoped that we could improve performance by combining the predictions of multiple independent models using the Ensemble Voting Classifier, but unfortunately, the accuracy of the combined model fell short of Random Forrest alone. The correlation matrix for the strongest model, Random Forests is also presented in Figure 8, showing that the there is room for improvement in reducing the number of false positives and false negatives.

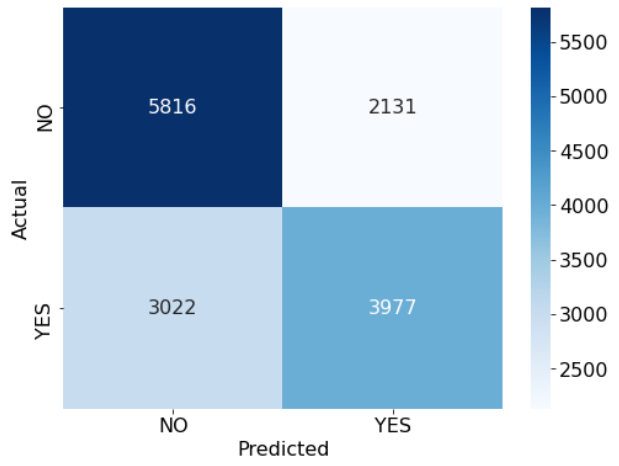

Figure [8]: Correlation Matrix of Optimized Random Forest Predictions

Feature	Importance
num_lab_procedures	0.0514
num_medications	0.0486
number_inpatient	0.0399
time_in_hospital	0.0365
number_diagnoses	0.0281
num_procedures	0.0248
number_outpatient	0.0173
number_emergency	0.0163
gender_Male	0.0137
insulin_Yes	0.0102

Table 2: Most important features in optimized Random Forest Model

Within the optimized Random Forrest model, we look the top features that influence the model's decision-making process (Table 2). Many of these features were not the subject of our initial data exploration process, such as the number of lab procedures, the number of distinct medications

taken during the visit, and the number of hospital procedures conducted. Certain features that we explored through visualizations did appear amongst the most influential features for the random forest model, such as the features of hospitalization duration and the number of diagnoses.

Section 4: Conclusion and Discussion

In conclusion, in line with our exploratory questions, we investigated the demographics of hospitalized diabetes patients, looking at age, race, common diagnoses, etc., finding that hospitalized diabetes patients are older, more likely Caucasian or African American, and often have co-diagnoses of circulatory or and respiratory diseases. We have trained and optimized four different machine learning classifier algorithms to predict hospital readmittance based on features of an initial hospital stay. To further improve model performance, we propose expanding the features of our dataset by reevaluating what data is available within the Cerner Health Facts Database. As we did not have access to the original database, our options for exploring alternative features were limited. If we did have access to the original database of electronic health records, however, we may also have found that some potentially informative features may not be routinely collected or recorded, which was the case with a patient's weight in the utilized dataset. Certain data that may have been helpful in predicting readmittance, such as weight, were unfortunately missing from our dataset. In some cases, incomplete information is a solvable problem for data professionals, but in our case, changing how hospitals collect data around a hospital stay would not be feasible.

In addition to changing how features are selected, we recommend that future work can focus on the specific medical specialty associated with the hospital visit or to focus on a patient's primary diagnosis. In our data exploration, we have seen that diabetes patients are hospitalized with a variety of different diagnoses, and when patients are hospitalized for different reasons, the features that influence future readmittance would be expected to differ. Thus, training different models for different categories may improve predictive power while reducing processing time. Potential roadblocks in such an approach may be that it may be

difficult to predict readmittance for uncommon diagnoses, but this problem also exists in the case of a more general model. If strong models can be developed for individual categories of diabetes patients, these better models could help improve patient outcomes by enabling medical professionals to monitor patients most at risk of requiring additional medical care.

References

1. Ahmad FB, Anderson RN. The Leading Causes of Death in the US for 2020. JAMA. 2021;325(18):1829–1830. DOI: 10.1001/jama.2021.5469
2. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020. Link
3. McIlvennan CK, Eapen ZJ , Allen LA. Hospital Readmissions Reduction Program. Circulation. 2015;131:1796–1803. DOI: 10.1161/CIRCULATIONAHA.114.010270
4. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. Biomed Res Int. 2014:781670. DOI: 10.1155/2014/781670
5. U.S. Census Bureau. HIC-9_ACS. Population Without Health Insurance Coverage by Race and Hispanic Origin: 2008 to 2019. Health Insurance Historical Tables - HHI Series. Link
6. Cherney, K. (2018, July 6). Age of onset for type 2 diabetes: Risk factors and more. Healthline. Retrieved November 8, 2021, from https://www.healthline.com/health/type-2-diabetes-age-of-onset.
7. Tachkov, K., Mitov, K., Koleva, Y., Mitkova, Z., Kamusheva, M., Dimitrova, M., Petkova, V., Savova, A., Doneva, M., Tcarukciev, D., Valov, V., Angelova, G., Manova, M., & Petrova, G. (2020). Life expectancy and survival analysis of patients with diabetes compared to the non diabetic population in Bulgaria. PloS one, 15(5), e0232815. DOI: 10.1371/journal.pone.0232815