Diabetes in America: Insights and Analytics

Rohan Bansal
Stanford University
robansal@stanford.edu

Rohan Cherivirala
Stanford University
rohanc12@stanford.edu

Ryan Nurwono
University of California, Berkeley
ryan_nurwono@berkeley.edu

Nikhil Suresh
Stanford University
ncsuresh@stanford.edu

1 Non-Technical Executive Summary

1.1 Introduction

Diabetes is a group of chronic metabolic disorders that are generally characterized by high blood sugar levels. Reduced levels of insulin production or incompatibility of insulin with the host are the root causes of diabetes [1]. Lifestyle changes over the past half century have been a significant contributor to increasing rates of diabetes and obesity in both developed and developing countries. Diabetes has been ranked highly on the international health agenda as it becomes ever more prevalent. Moreover, in developing countries where treatment is not aimed at chronic conditions, expensive new drugs are unaffordable for millions with limited resources [4]. Unilaterally increasing costs for care estimates put the total cost of diagnosed diabetes increasing by at least 25 percent every five years for the past few decades. With over 37 million Americans that currently have diabetes and expensive recurring medical costs, reducing costs from patient visits and prescriptions would greatly improve general healthcare related to diabetes [3].

1.2 Problem Statement

Our team decided to address these issues by providing better foresight into the process, specifically by focusing on being able to predict outcomes of diabetes patients given the available demographic data and treatment information. In this paper, we discuss the technical aspect of transforming and performing quality control on the data set in order to determine the most significant features and our definition of a successful outcome. We also investigate the implications and suggestive capabilities of the predictions in determining which features are most revealing, using data on diabetes treatment over a 10 year span (1999-2008) from 130 US hospitals. We wanted to investigate this issue to help determine the most important factors in successful diabetes treatment. We hope these results will help improve treatment of diabetes and decisions being made by personnel in hospitals.

1.3 Key Findings and Significance

Our model focuses on personal information and clinical data, such as race, gender, age, number of lab procedures, to predict the relative success of a diabetes patient in being discharged (to home or elsewhere) and being readmitted (never, within 30 days, after 30 days). The model is 2.5x more accurate than making the baseline prediction of randomly guessing, which is indicative that there are distinct factors that pose significance in the prediction process. We also consistently determined with statistical significance that the most important factors that came with this decision were time spent in hospital, number of inpatient visits, age, and result of first diagnosis. Repeated training of the model on different training sets resulted in the same, relevant factors being expressed, which highlights that our model was able to pick up on general features and filter out some level of noise in a dataset which had a large amount of data and some relatively sparse features. This type of model could be used in the future to perform generic searches on larger datasets to determine best approach to care for various demographics. For example, this could help doctors determine how many tests and which medications should be prescribed to an African-American woman who was suffering from diabetes in order to maximize her future health outcomes. We can also identify these types of discrepancies between demographics and ideally help shape general decision making and inform medical experts on new methods of evaluation for patient care.

2 Technical Exposition

2.1 Data Cleaning/Transformation

The dataset we analyzed consisted of data on diabetes treatment over 10 years (1999-2008) from clinical care at 130 US hospitals. The raw dataset contained 101766 unique patient interactions and represents 71518 unique patients. There were a trivial amount of unknown or missing values for race, a field which was presumed to figure heavily into patient outcomes, so we chose to replace missing race values with the mode (as the data was heavily white) while we removed the <5 rows which were missing gender. The resulting gender and race breakdowns are shown in Figures 1 and 2, and we found roughly 80% of the dataset was white and nearly 55% was female.

We then analyzed missing values to determine which fields were prevalent enough to merit remaining in our analysis. The weight and payer_code fields were missing in over 95% of the rows, so we completely eliminated them. Similarly, medical specialty appeared in less than 50% of the rows, so we chose to eliminate that column due to the combination of a large amount of categories with such sparse representation. Among the medications provided in the dataset, and chose to remove medications which were present in less than 0.2% of the total records, resulting in the removal of 12 out of 23 total medications. This could be a potentially problematic decision, as we recognize that certain medications could be prescribed in very specific and serious manifestations of diabetes. However the lack of data for these medications was likely to lead to skewed results, and the elimination of these less represented fields allowed us reduce potential issues later in our analysis, as well as be more confident in the results our model arrived at.

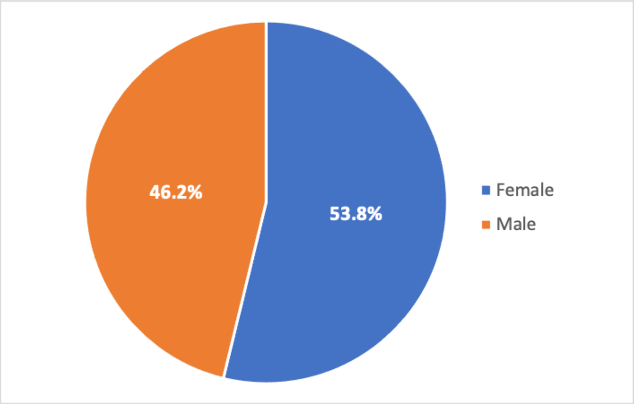

Figure 1: Gender breakdown in the cleaned dataset

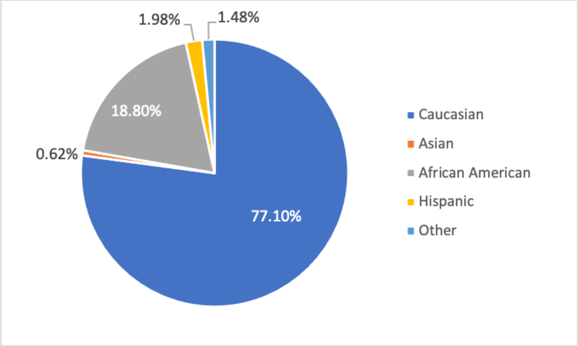

Figure 2: Race breakdown in the cleaned dataset

In order to determine success in the outcome of a diabetes patient, we primarily focused on two of the fields in the dataset–the readmittance and the discharge description. The readmittance entries were categorized into three groups, those who were not readmitted, those who were readmitted after a month or longer, and those who were readmitted within a month. On the contrary, there were 30 different discharge descriptions, however the data was heavily skewed as nearly 60% of the entries were discharged to home, which we considered a positive outcome. By manually searching through the remaining descriptions, we categorized these outcomes as negative or positive. For example, we realized that many discharged categories involved transferring the patient to another facility. Through our research, we found this usually came from the lack of resources or availability from the current facility, and thus evaluated the numbers as a negative outcome[2]. We also found one outcome that could potentially be considered neutral, but it appeared in only 3 entries, so we considered it trivial. Thus, we chose to binarize the discharge descriptions as "Discharged to home" which was a positive result and "Other" which was considered negative. These two fields allowed us to generate 6 output buckets for prediction which were progressively worse. For example, a patient who was discharged to home and never readmitted would be purported to have a better outcome than one who was discharged to home but readmitted within a month.

This output category was significant for a few different reasons. Firstly, there was very little direct correlation (<.03) between discharge description and readmittance, which seemed contradictory. This lack of correlation does however indicate that our output variable is significant and captures some type of latent outcome and it is not simply redundant, because the two consisting variables capture different information. Additionally, we can identify some clear trends in this output variable with features that we know should be related such as age. fig. 3 highlights these trends as the output is almost exclusively highly positive and this trend gradually inverts as the age increases leading to primarily negative outcomes in the upper age demographics.

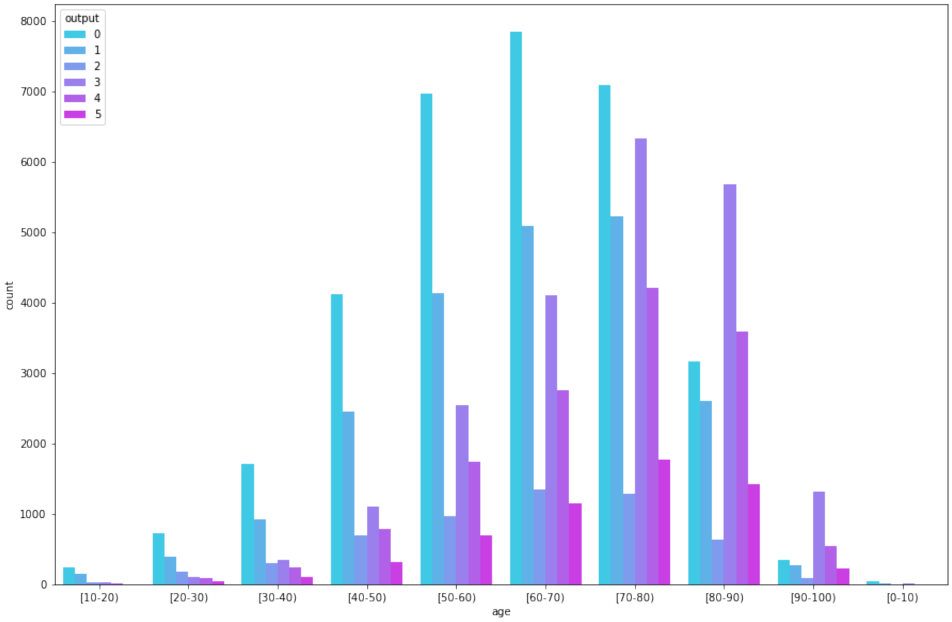

Figure 3: Distribution of output outcomes amongst various age groups

The breakdowns for readmittance and our generated output field can be seen in figs. 4 and 5 and they highlight the skewed nature of the initial dataset. Over 50% of the raw data was assigned to the "discharged to home" description leading to relatively sparse representations of the majority of the remaining descriptors, but creating discretized buckets by combining with readmittance allowed us to create a significantly less skewed distributions for our target variable. Similarly, over 50% of the entries were never readmitted and only 11% were readmitted within 30 days which represents a skewing towards positive outcomes, so our categorization helps alleviate some of this bias and create a better sliding scale of general patient outcome

To handle outliers, we first determined outlier counts for each relevant numerical feature and then attempted to logically rationalize and decide whether the outliers should be replaced with a more representative value depending on qualitative analysis and percentage of values which were outliers. We primarily chose to keep outlier values because they did not seem to come from erroneous data collection and would provide some important clues about interesting behavior in these features. If provided more data, and more time, it would be illustrative to perform the modeling with both the outliers kept, replaced with central characteristics, and removed to determine if this decision had a statistically significant impact on the performance.

Discussing the significance of our success metric in output, the correlation found between a numerized version of readmittance and discharge_disposition_id was near 0, (-0.03474). This demonstrates that the influence of incorporating one feature with the other is not redundant, and our assumptions included this version of output being a more informative depiction of success.

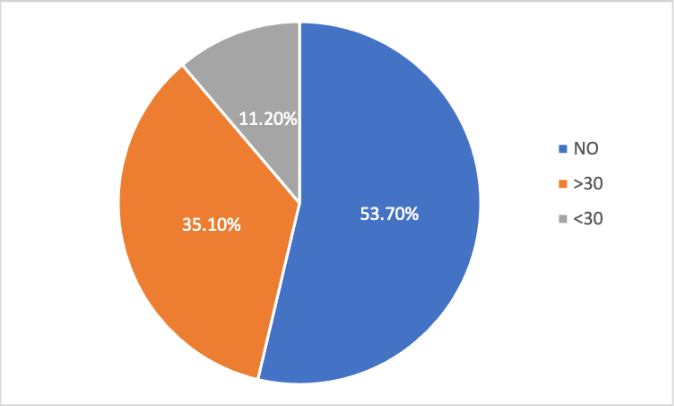

Figure 4: Re-admittance breakdown in the cleaned dataset

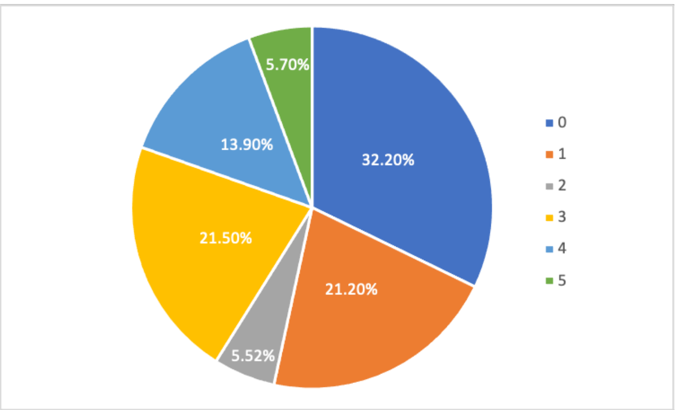

Figure 5: Output buckets breakdown in the cleaned dataset. A higher bucket number represents a poorer outcome.

2.2 Feature Exploration

After our cleaning process, we had 32 total input features, of which 21 were categorical and 11 were numerical. This included a combination of independent, predetermined variables such as race, previous medical history and gender, as well as treatment related variables including testing, medications, and diagnosis. For modeling, we had to both reduce the number of features which we were utilizing to improve accuracy/predictions and increase speed while also creating aggregations for some of the categorical data to reduce sparsity. We did basic statistical testing to identify any direct correlations and found limited correlations between many of our indicator classes and the output we were aiming to predict. However, certain features such as age seemed to have a medium correlation, which in the medical field can often be a fairly significant relationship. fig. 6 demonstrates some of these correlations.

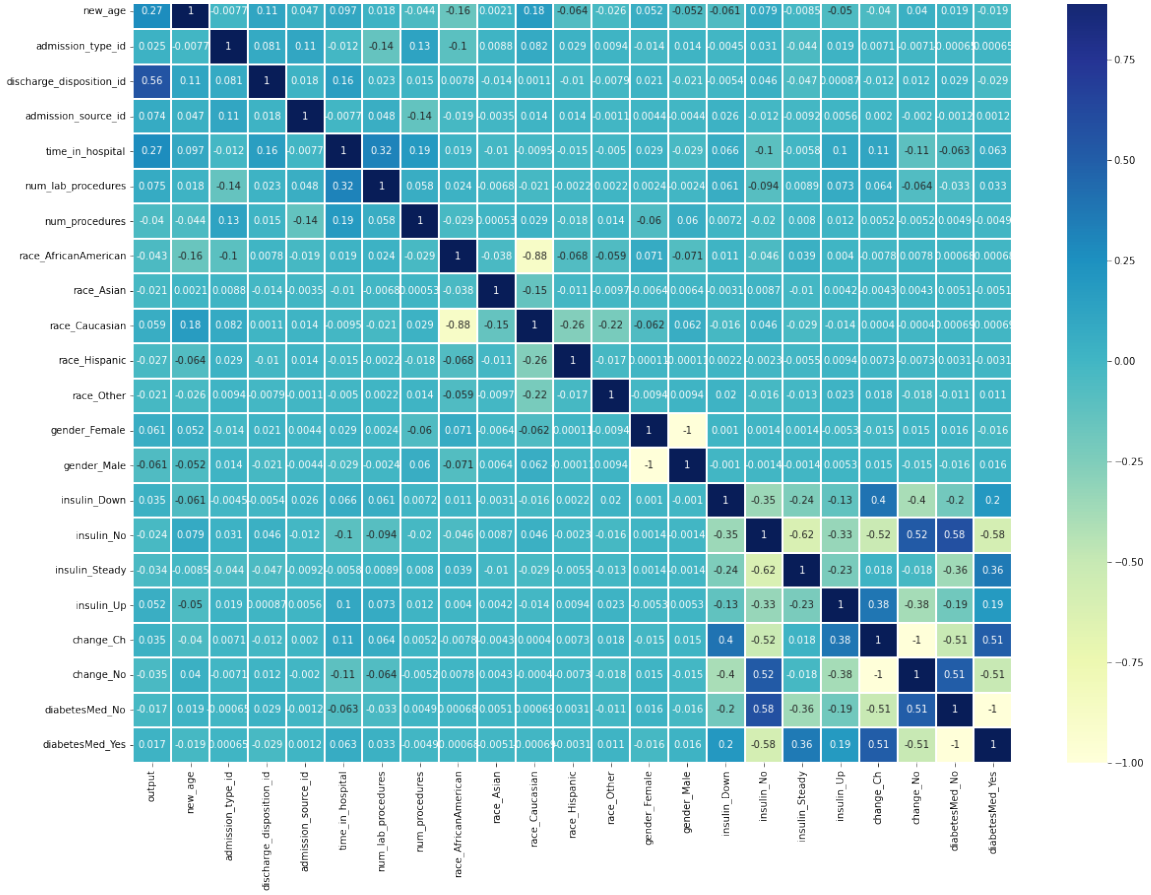

Figure 6: This heatmap displays the correlations between various selected features with each other

We made a variety of processing and extraction-related decisions here to try to optimize the balance between speed and accuracy. Most notably, we created bins for the diagnosis codes (which were organized by their ICD sequence) and followed the ICD's own binning process for creating 20 unique diagnosis classes. We relied on one hot encoding for the remaining categorical features due to the relative separation in the features, however, this approach did leave us more susceptible to issues with a lack of data. For example, with many of the medications, the dataset contained <5 entries in which the drug was prescribed or increased. We chose to eliminate these types of classes from the data because of their extremely low frequency (as they appeared in less than 0.01% of the entries), and some examples of the eliminated fields are provided in table 1.

Class	Count
acarbose_Up	10
acarbose_Down	3
glyburide-metformin_Up	8
glyburide-metformin_Down	6

Table 1: Some of the classes which were eliminated from data analysis due to sparsity

This process of cleaning and categorization left us with **18** categorical and **14** numerical features, resulting in **80** unique classes.

2.3 Modeling

For modeling, we initially chose to implement a mirkowski KNN due to the large number of classes and lack of access to the necessary hardware required for deep learning. In order to properly train and test this model, we set aside 20% of the provided set to act as testing data and used the other 80% as training data. Upon training the test, we found that it took nearly 2 minutes to properly train the KNN model and calculate our predictions for the training and testing set. We also opted to use recall as our metric for quantifying the accuracy of the model because it provided a clear way of identifying the model's chances of selecting the correct class and allowed us to compare directly to the random baseline. The KNN model with a k-neighbors value of 100 had a 35% accuracy on the training set and a 33% accuracy on the testing set. This is, however, statistically significant, even if we compare to the baseline as the random chance of predicting a correct class out of 6 options is simply 16%. Thus, the model performed about two times as well as the random baseline. Despite this seemingly positive result, the nearly identical performance of the KNN between the train and test sets for a variety of hyperparameter classes pointed to the model routinely underfitting the data and being unable to capture some of the hidden and more complicated relationships between the various features present in our dataset. Even when increasing the neighbors to increase variance, the model was unable to improve its performance to over 35%, even on the test set, emphasizing that it may simply be unable to properly be fitted to the dataset provided.

Because KNN's are frequently hampered by datasets with too many classes if the representation is too sparse (curse of dimensionality) and our performance remained below 50%, we then elected to implement a decision tree to compare and generate predictions on. The decision tree had a variety of benefits, including the ability to tune its max-depth and its ability to find hidden, non-linear relationships between certain variables. Decision trees can also be train and run much more quickly due to their lazy loading of data, which allowed us to perform more runs and gridsearch across a wider variety of hyperparameters. We decided to vary the maximum depth of the decision tree and randomly shuffle the data 15 times for each value of the depth. We then determined the direct recall of this model to determine the accuracy, identical to the above process for KNN, and were able to get an average train recall of 44% and test recall of 39%. The depth played a significant role in these decisions and it appeared that the ideal hyperparameter configuration had a depth of 10 and the training performance was slightly better than the testing performance, as can be seen in table 2. This performance is statistically significant as it represents a performance of over 2.5x the random baseline and over a 10% increase from the KNN performance.

Additionally, we were able to identify the most heavily expressed features in the decision tree to attempt to provide some additional contextual framing, and the top 5 features are presented in table 3. These percentages were determined by running the decision tree 25 times with a max depth of 12 and averaging the weights. The top features, such as time spent in hospital and number of previous inpatient visits, are obvious candidates in predicting future health and thus the fact that they are highly expressed in the classifier indicate that it is able to determine comprehensible nodes.

Depth (N)	Training Accuracy	Testing Accuracy
5	39.36%	38.98%
10	43.75%	39.24%
15	56.80%	37.44%
20	77.32%	33.23%
25	92.45%	30.81%

Table 2: Performance of our Decision Tree model given various depths

Feature	Percentage of Prediction
Time spent in hospital	21.79%
Number of inpatient visits	16.47%
Age (80-90 years old)	8.92%
1st Diagnosis	5.88%
Age (70-80 years old)	5.85%

Table 3: Top 5 features with the highest impact on the Decision Tree prediction

3 Conclusion

In this paper, we have shown that, using a machine learning model built on the provided dataset, we are able to predict the outcome of a given person's stay at a hospital given that they have diabetes based on different features and characteristics of their stay. This result is significant in that it can be used to predict the outcome of a given person's stay at a highly significant rate, thus providing valuable insight on the outcome of someone's treatment.

Acknowledgments and Disclosure of Funding

Thanks to Citadel and Correlation One for sponsoring the event.

References

[1] 2020. URL: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444.

[2] Isla Hains. *Transfer troubles*. URL: https://psnet.ahrq.gov/web-mm/transfer-troubles.

[3] *What is diabetes?* 2021. URL: https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=Diabetes.

[4] Paul Z Zimmet, Dianna J Magliano, William H Herman, and Jonathan E Shaw. "Diabetes: a 21st century challenge". *The Lancet Diabetes Endocrinology* (2014).