CALIFORNIA POLYTECHNIC STATE UNIVERSITY, POMONA

Dylan Gonzalez

Data Mining Final Project Report

May 9th, 2024

CIS 4321

TABLE OF CONTENTS

INTRODUCTION

In a recent article by the Centers for Disease Control and Prevention, "*During August 2021–August 2023, the prevalence of total diabetes was 15.8%, diagnosed diabetes was 11.3%, and undiagnosed diabetes was 4.5% in U.S. adults"* (Jane A. Guira et al., 2024).

This report is going to explore the chronic medical condition, Diabetes. This condition affects millions of people worldwide and poses significant challenges to public health systems. With the rising prevalence of diabetes, data-driven approaches have become increasingly important for understanding the disease, identifying risk factors, and improving diagnosis and treatment. This report analyzes a diabetes dataset to explore trends, uncover patterns, and derive insights using statistical and machine learning techniques. The goal is to enhance our understanding of the factors associated with diabetes and demonstrate how data analytics can support better healthcare decisions.

PROBLEM STATEMENT

The purpose of this report is to explore the features that influence the likelihood of a diabetes diagnosis and to assess whether we can accurately predict the presence of diabetes based on a given dataset. Accurate data handling in the healthcare domain is crucial to ensure reliable diagnostic predictions and to minimize the risk of misdiagnosis. Understanding the factors contributing to a diagnosis not only has significant implications for individual patient health, but also informs broader preventive strategies.

This analysis aims to determine how likely a person is to be diagnosed with diabetes based on specific features and to highlight how the resulting insights can further research efforts, support early detection, and guide effective treatment strategies.

UNDERSTANDING THE DATASET

In the initial exploration of the diabetes dataset, there are a total of 253,680 entries and 22 columns. The entries for each column are all mostly numerical. Binary coding was used to indicate whether the individual has the said attribute or not for most of the columns. The only two columns that were ordinal, after cleaning, were “diabetes” and “age”.

The features that I believed needed to be modified were “diabetes”, “gender”, and “age”. Changing these three variables would make further data analysis readable. The target variable for this dataset is the column, “diabetes”. According to the dataset codebook, 0 indicates No Diabetes, 1 indicates Prediabetes, and 2 indicates Diabetes. “No Diabetes” takes up 84% of the dataset, which is heavily imbalanced. I balanced the target variable to prevent a bias in the model by combining both “Prediabetes” and the “Diabetes” entries together. Later in the analysis, I resampled the dataset so that “No Diabetes” has a total of 43,000 instances and “Diabetes” has 39,997 instances.

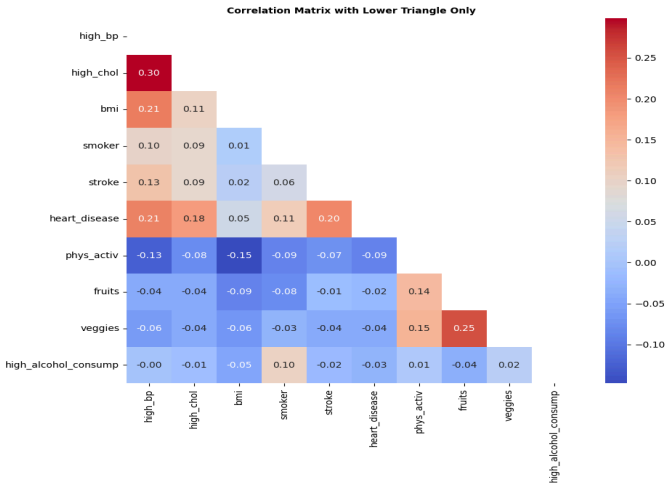
The “gender” column was also in binary. According to the dataset codebook, 0 indicates female and 1 indicates male. I mapped out the genders from binary to a string, which then changes the entire column into nominal, categorical data. The “age” column was tricky. The entries were numerical and ordinal. For example, 1 was meant to indicate the individual fell between the ages of 18-24, 2 indicated ages 25-29, and so on and so forth. The way I handled this was to map out the numerical data according to the dataset codebook and make it categorical, having the label of ages be inputted instead of a single integer.

After cleaning the data, there were a total of 13 columns and a total of 253,680 entries without resampling. With resampling, there are still the same amount of columns and 82,977 entries.

DATA ANALYTICS AND RESULTS

Performing exploratory data analysis, I discovered interesting results. Using the "groupby" function in python, I grouped "diabetes" and "gender" first. Females with diabetes had 21,015 instances and males with diabetes had 18,962 instances. Further analysis, the age group with the most cases of diabetes are 65-69. Observing this, we can assume that as the person advances in years the more likely they are to develop diabetes. The gender that is at a higher risk with this are females. Physical activity is important for health and observing the data, I found out that individuals who *are not* physically active and have diabetes are less than the number of individuals who *are* physically active with the same diagnosis. Interestingly enough, individuals who *are not* physically active and *have no* diabetes outweighs individuals with diabetes and no physical activity. This could be due to the different lifestyles a person has and the types of nourishment they feed their bodies that contribute to them being diagnosed with this disease.

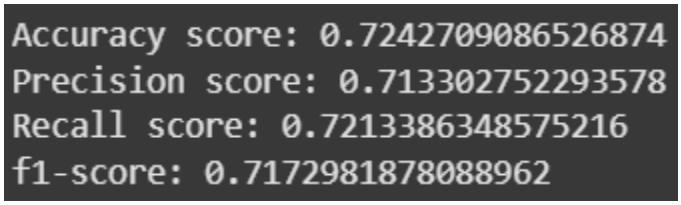
The correlation matrix on the right describes which features are highly correlated with each other by their coefficients. While the values are small, we can still analyze a few and their significance. We can observe that high cholesterol (high_chol) and high blood pressure (high_bp) have a correlation score of 0.30. This means that people who have high blood pressure tend to have high cholesterol as well and vice versa. The next variables that are correlated are fruits and veggies with a score of 0.25. This means that these foods also contribute to the model in a



positive manner. When one of these variables increases, the other increases as well and vice versa. The variables together could contribute to someone being diagnosed with diabetes. The types of nutrition they give their bodies which in return affects their blood pressure and cholesterol levels.

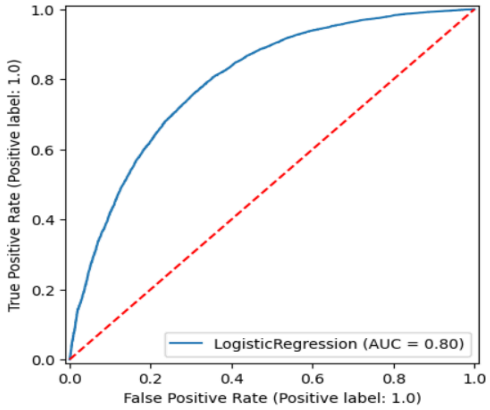
The algorithm I chose to predict if a person has diabetes or not is the Logistic Regression Model. Logistic Regression works well with both categorical data and binary data. It is easy to train the model and can handle a relatively large dataset, making it highly scalable. In this case, about 80,000 entries and 23 columns were used when the data was resampled and encoded.

Evaluating the metrics of the data, the dataset is not overfitted. The training score is 0.7199 and the testing score is 0.7247. These are acceptable scores, but can be improved. The scores indicate a good balance and the model is able to generalize well to new data without introducing a bias. The overall performance metrics of the model are acceptable, but again, can be improved. The accuracy score came out to be 0.7242, which means the model is able to classify about 72.42% correct predictions to the total number of correct predicitons. This metric does not give us the complete picture of how well the model is doing, so we observe the other metrics. Precision score measured 0.7133, which means the model is able to genuinely predict about 71.33% of individuals to have diabetes. In the context of the problem, there is a 71.33% chance the diagnosis is accurate. Recall score measured 0.7213, which means the model identifies 72.13% truly have the disease. This is a crucial metric because this implies that about 24% of patients with this disease could go undiagnosed. The f1-score calculates a balance between recall and precision scores and it measured 0.7172. The performance of the



model to correctly diagnose patients with diabetes is 71.72%. In the context of this problem, there is room for improvement since correctly diagnosing someone with this disease is crucial and not administering medication to someone who does not have the disease and could cause harm due to an error.

The Receiver Operating Characteristic (ROC) Curve, which evaluates the performance of a binary classification model. The blue curved line represents the ideal point. The closer the curve is to the top-left corner, the better the performance. The red diagonal line represents the threshold for the model and if the curve is closer to the threshold, then the model is unreliable. The Area Under the Curve (AUC) metric summarizes the classifier's performance by giving us a number. An AUC of zero indicates a bad classifier and 1 indicates a good classifier. Evaluating the ROC Curve, the model performs generally well with an AUC score of 0.80. This means the model has good predictive performance.



The overall metrics showing how the model is able to predict an individual with or without diabetes as follows. Someone being classified with no diabetes has consistent scores of 0.73 or 73%. We can assume that the reason why it was able to give better metrics for someone with no diabetes is due to the amount of samples and the complexity of the problem. Someone being classified with diabetes has a relative consistent score of 0.72 or 72%. Again, this could be due to the complexity of the problem and how patients have similar characteristics but different diagnoisis.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	71
1	0.97	0.91	0.94	43
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

CONCLUSION

After performing various metrics and visuals that describe the diabetes dataset, can we accurately predict someone being diagnosed with diabetes based on the given traits? I would say yes, but there needs to be improvement on the data collection. I believe the outcomes I got from analyzing the dataset is due to not only the complexity of the data, but how broad the data collecting was and the limited features. The “age” column gave a range of ages instead of specifying the exact age of the patient, which does not help the physician accurately diagnose someone. I also think introducing other features such as A1C levels and glucose levels can help the model better identify someone having diabetes. Most features from the original dataset were irrelevant such as income, health insurance plan, and education. Collecting specific data and being specific with the numbers or categories instead of giving a broad range can help the model’s performance and give physicians and researchers higher scores for accuracy, recall, precision, and f1 metrics.

SOURCES

Jane A. Gwira, M.D., M.P.H., Cheryl D. Fryar, M.S.P.H., and Qiuping Gu, M.D., Ph.D. (2024). Prevalence of Total, Diagnosed, and Undiagnosed Diabetes in Adults: United States, August 2021–August 2023. *National Center for Health Statistics* https://www.cdc.gov/nchs/products/databriefs/db516.htm

CDC BRFSS (2015). Diabetes Health Indicators Dataset. *Kaggle* https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download