PROMINENT RISK FACTORS IN DIABETES

Oleg Fleitman

College of Engineering and Applied Science, University of Colorado Boulder, USA

ABSTRACT

A December 2023 Fortune [1] article revealed that nearly 50% of the U.S. population has Diabetes or Prediabetes, many unaware of it. This inspired a data mining project using the CDC's 2015 BRFSS dataset [2], with 253,000 entries and 17 features, to identify key Diabetes risk factors. The data was pre- processed using SMOTE to address class imbalance before applying four models: Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. While Logistic Regression had the lowest F1 score (0.66), the others achieved an F1 score of 0.83. Age, BMI, and General Health were the top three risk factors identified. It is recommended to target diabetes awareness campaigns at individuals over 45, with a BMI above 25, or those who self-rate their health poorly. Future work should involve a broader set of features and consultation with medical experts.

KEYWORDS

Diabetes, risk factors, prevention, reversing Diabetes, Type 1, Type 2, pre-Diabetes

1. INTRODUCTION

Before we dive in, I would like to ground you on what Diabetes is and the four main types of Diabetes. Diabetes is a metabolic disease where an individual has persistent elevated levels of blood glucose that over time, leads to significant damage to an individuals, eyes, kidneys, heart, nerves, and blood vessels [3].

In Type 1 Diabetes, the individuals body makes little or no insulin. Their immune system attacks and destroys the cells in the pancreas that is responsible for creating insulin, which is a hormone that regulates blood sugar. This group requires taking insulin every day to stay alive [4].

In Type 2 Diabetes the issue is not related to the immune system. Instead, the issue is twofold. First, the body is not utilizing insulin correctly to reduce blood glucose levels. Second, the pancreas does not produce enough insulin [5].

In Gestational Diabetes this occurs in females during pregnancy. Researchers are not yet aware of what causes this phenomenon; however, it’s important to recognize the risk factors as Gestational Diabetes can impact the baby’s health if left uncontrolled [6].

Lastly, there are individuals that are not diagnosed Diabetics but are nearing the threshold. These individuals are referred to as Pre-Diabetics. These individuals have a higher-than-normal blood glucose level, but not high enough to be diagnosed with Diabetes.

It is important to note that about 95% of the cases are of the form of Type 2 Diabetes,

while only 5% of cases account for Type 1 Diabetes [7]. While Type 1 Diabetes is a lifelong illness, there is a large volume of research that indicates Type 2 Diabetes is reversible if properly managed [8]. It is critical to begin management of Type 2 Diabetes earlier than later. The sooner an individual begins managing their glucose levels, the better the outcomes will be. Early management is not limited to those who have been diagnosed with Type 2 Diabetes, however. Early management can also be extended to those who are in the pre-Diabetes group to ensure they don't get to the diabetic stage. While Type 1 Diabetes is not yet reversible, the same management techniques can be utilized by Type 1 Diabetics to reduce the amount of medicine (i.e., insulin) one needs to rely on. Our analysis will not differentiate between pre-diabetics, Type 1 or Type 2 Diabetics. Instead, we will focus on identifying the top risk factors that are most relevant to identify individuals who are at risk for Diabetes in general [9]. The purpose of this work is to increase awareness through educating individuals on the importance of getting screened for Diabetes if they present themselves with such risk factors.

2. RELATED WORKS

Although there is a lot of research and literature surrounding the risk factors that are prevalent in Diabetes. Most of their applications are geared toward health professionals or government entities. Specifically, raising awareness to find new therapies to treat Diabetes, innovative ways to better diagnose individuals, or simply offering predictive models to identify those who may have Diabetes. There are limited applications in leveraging the results to run an educational campaign and nudge millions of Americans to screen themselves for Diabetes.

A paper titled Association of risk factors with type 2 Diabetes: A systematic review conducted an analysis to identify the majority of the risk factors for the incidence and prevalence of type 2 Diabetes. The paper identified "sleep quantity/quality, smoking, dyslipidemia, hypertension, ethnicity, family history of Diabetes, obesity, and physical inactivity" as top risk factors of Diabetes [10]. However, the purpose of the analysis was geared toward health professionals and government institutions to help with better diagnostic methods and prognosis of the disease. The aim of my paper is to bring awareness to the millions of Americans who are living with this disease without even realizing it with the expectation that early intervention will allow them to lead normal lives without Diabetes.

Another paper titled Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention conducted an analysis on various risk factors of Diabetes. The paper includes factors such as "age, sex, height, waist circumference, BMI, ethnicity, history of hypertension, and prevalent/latent Diabetes, medication use, physical activity, and consumption of berries alcohol, coffee, whole grains, fruits, vegetables, and red meat" [11]. Although their aim is focused on raising awareness for developing new therapies, it is not focused on raising awareness for Americans to get screened for the disease.

Lastly, another paper titled Identifying risk factors associated with type 2 Diabetes based on data analysis establishes that Triglyceride and hemoglobin have been identified in this study as the most influencing factors for developing type 2 Diabetes. This is unsurprising as hemoglobin levels are a diagnostic tool to determine whether someone has Diabetes. The paper offers no real-world applications other than a predictive model based on specific data points [12].

3. PROPOSED WORK

3.1.Data Understanding

The data set contains 253,680 entries, including 1 target variable with 3 classes – those with no Diabetes, those with type 2 Diabetes and those with Prediabetes. In addition, the data contains 17 features including (1) whether an individual has high blood pressure, (2) whether an individual has high cholesterol, (3) whether the individuals has checked their cholesterol levels in the last 5 years, (4) the individuals body mass index, (5) whether the individual is a smoker who smoked at least 100 cigarettes in their lifetime, (6) whether the individual has been told they previously had a stroke, (7) whether the individual has heart disease or had a heart attack in the past, (8) how many days of physical activity did the individual participate in the last 30 days, (9) whether the individual consumed 1 or more fruit servings a day, (10) whether the individual consumed 1 or more veggies a day, (11) the individual would be considered one who has heavy alcohol consumption, (12) individuals general health on a scale from 1 to 5, inclusive, (13) whether the individual had a specific number of mental health days that were not good in the past 30 days, (14) whether the individual had a specific number of days where their physical health was not good in the past 30 days, (15) whether they have difficulty walking or climbing stairs, (16) the individuals gender, and (17) the individuals age.

It is important to note that there were two additional features included as part of this dataset. Specifically, education and income. Based on my domain knowledge, I have decided to remove both variables from the dataset. Although one could argue that more educated people could potentially have less chance at developing Type 2 Diabetes as they would practice blood glucose management techniques through lifestyle and diet, it would be difficult to make that conclusion as the feature just stipulates highest education level. It does not stipulate whether the individual has education in the topic of Diabetes. Similarly, it can argue that higher income individuals may have greater access to better medical care, especially in the US. However, I want to steer away from such generalizations as an argument could be made that higher income earners have less time to see a doctor. The purpose is to understand what risk factors are prevalent in having Diabetes so that we can educate the American public in getting screened through campaigns. Thus, in this paper we will be leveraging classification models to help answer our question.

Exploratory Data Analysis (EDA) was conducted on the dataset to ensure strategic preprocessing is actioned prior to model development. The dataset was examined to ensure no missing values are present. In addition, a correlation analysis was produced to get a sense of how each variable interacts with one another.

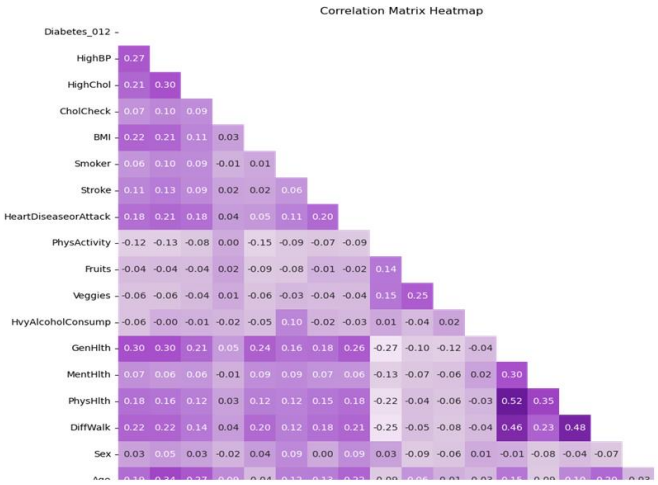

Figure 1. Correlation matrix showing how each variable correlates with each other.

distribution analysis was also produced in the form of mini histograms to gain a better understanding of whether the target and features are balanced.

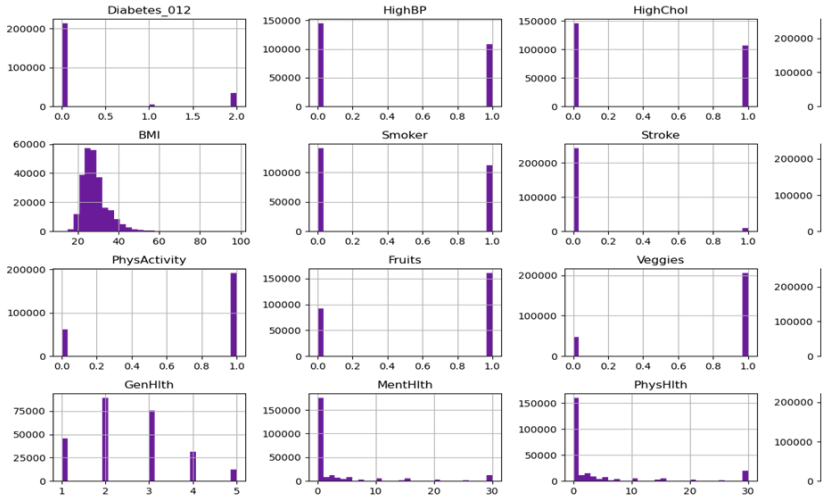

Figure 2: The first histogram shows the target followed by feature variables.

3.2. Data Preprocessing

The dataset was observed to be heavily unbalanced. Several techniques are available to deal with unbalanced data. One of the more common ways is resampling the training set. Specifically, under-sampling or over-sampling. Under-sampling balances the dataset by reducing the size of the copious feature, while over-sampling is leveraged when the quantity of data is not sufficient and the goal is to increase the size of the smaller scarce sample [13]. One of the more popular methods of over-sampling is leveraging the Synthetic Minority Over- Sampling Technique (SMOTE). "*SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line*" [14]. This is the method that will be used in preprocessing the data to account for the heavily imbalanced nature of the data.

3.3. Data Modeling

Given the dataset is in the form of a classification problem, I have chosen to start by running the following four models – Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine (SVM). While the former three models ran quite efficiently, the SVM model proved to run inefficiently. Further research showed that XGBoost "… can offer better performance on binary classification problems with a severe class imbalance [15]. Therefore, I replaced SVM with XGBoost to complete my analysis.

3.4. No Diabetes vs. Type 2 Diabetes

First, I explored the feature importance between those with no Diabetes vs. those with Type 2 Diabetes. Although the feature importance varied slightly by model, Random Forest and XGBoost agreed on BMI, General Health, and Age as the most important. Logistic Regression showed slightly different results and replaced Age with High Blood Pressure. In addition, Gradient Boosting replaced BMI with Mental Health though the two were not far off from each other.

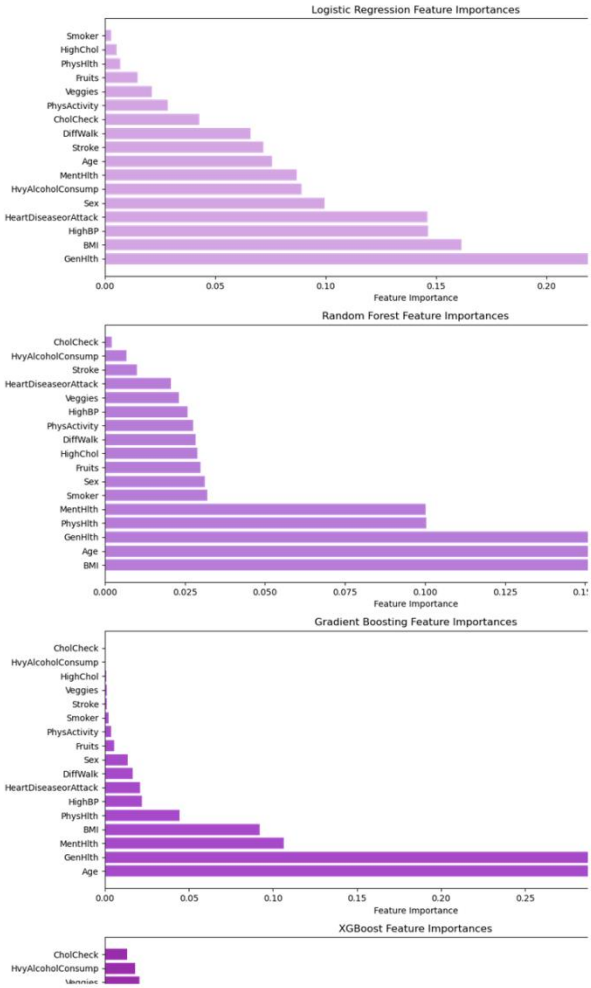

Figure 3: No Diabetes vs. Type 2 Diabetes feature importance based on Logistic Regression, Random Forest, Gradient Boosting, and XGBoost models.

3.5. No Diabetes Vs. Prediabetes

Next, I explored the feature importance between those with no Diabetes vs. Prediabetes. Unsurprisingly, the models produced similar results. However, with this group it is observed that Random Forest, Gradient Boosting, and XGBoost models agreed that the top three important features in predicting Prediabetes were BMI, General Health, and Age. Similar to the no Diabetes and Type 2 Diabetes group, Logistic Regression concluded that High Blood Pressure is more important in predicting Prediabetes than Age. I found it interesting that being a smoker was not a good predictor of a person to have either Type 2 Diabetes or Diabetes. However, the results make sense that eating vegetables, which contain fiber, reduce the risk factors of Diabetes.

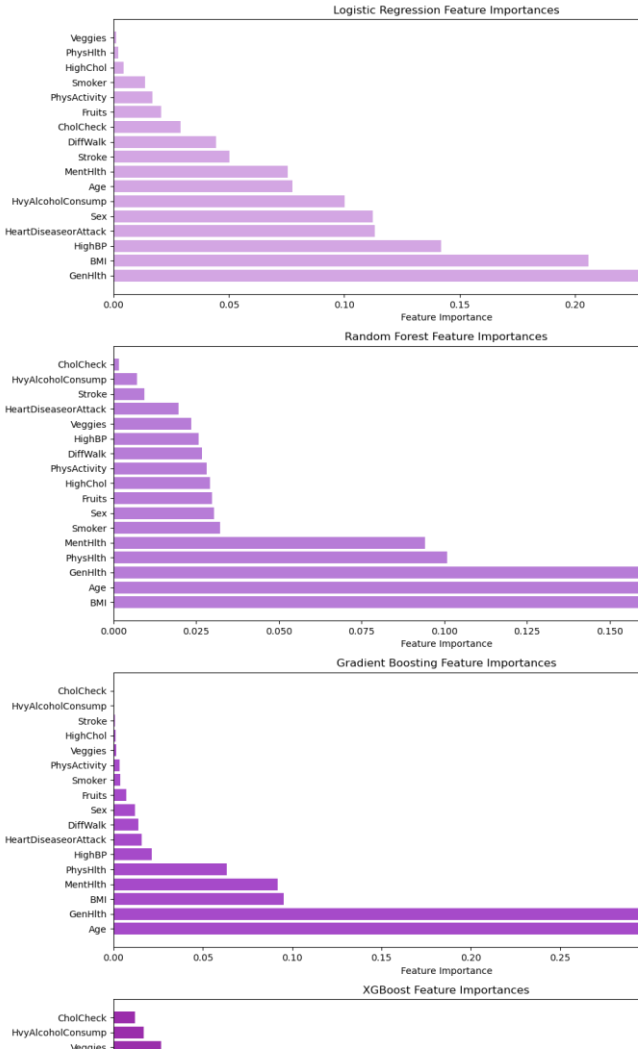

Figure 4: No Diabetes vs. Prediabetes feature importance based on Logistic Regression, Random Forest, Gradient Boosting, and XGBoost models.

4. EVALUATION

4.1. Model Performance

Each of the four models were evaluated using consistent metrics including Accuracy, Persistence, Recall and F1. Although Accuracy may not be the best measure for highly unbalanced datasets, we did leverage SMOTE technique to mitigate the unbalanced nature of the data. However, our focus will primarily be F1 measure which is a balance metric between Persistence and Recall.

4.2. No Diabetes vs. Type 2 Diabetes

Observing each of the four model's performance for the group between no Diabetes and Type 2 Diabetes, it is observed that the lowest performance is attributed to the Logistic Regression model. Surprisingly, Random Forest, Gradient Boosting, and XGBoost models performed fairly similar. The F1 score across each of the three models held up with a respectable score of 0.83 compared to 0.66 with Logistic Regression.

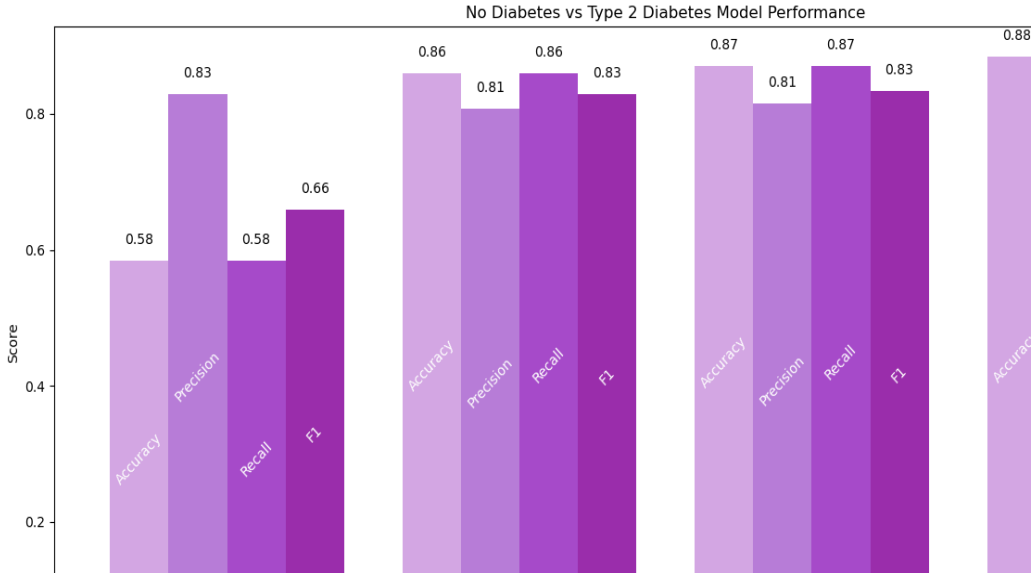

Figure 5: No Diabetes vs. Type 2 Diabetes model performance.

4.3. No Diabetes vs. Prediabetes

Observing each of the four model's performance for the group between no Diabetes and Prediabetes, it is observed that the lowest performance is also attributed to the Logistic Regression model. Random Forest, Gradient Boosting, and XGBoost models performed also fairly similar in this group. Unsurprisingly, the F1 score across each of the three models held up with a respectable score of 0.83 compared to 0.66 with Logistic Regression.

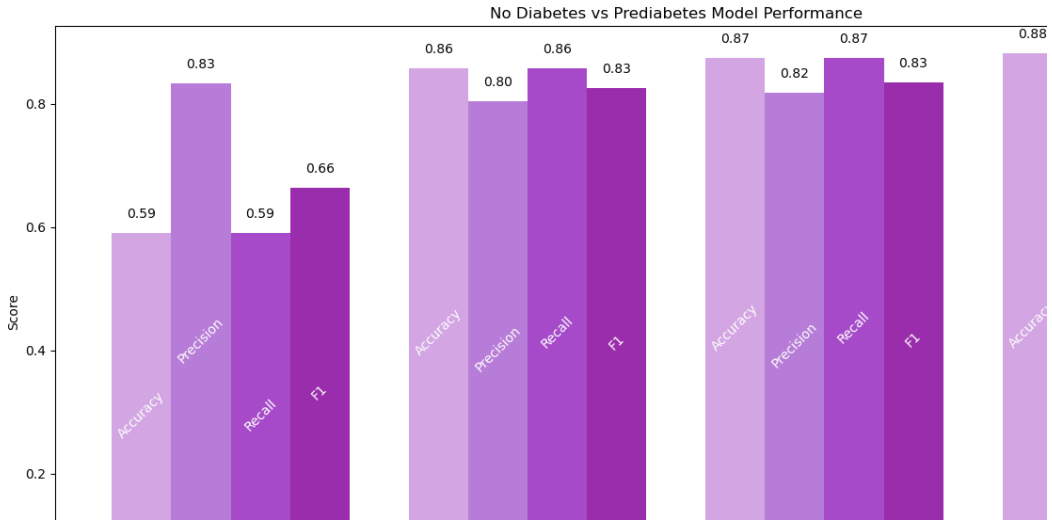

Figure 6: No Diabetes vs. Pre-Diabetes model performance.

5. DISCUSSION

Diving deeper into each of the three features that were ranked as highly important in the modelling phase, I will explore which age groups, Body Mass Indexes, and General Health groups are associated with higher Diabetes risks.

5.1. Age

Unsurprisingly, the older an individual is the higher at risk they are at developing Prediabetes or a form of Diabetes. Individuals who are 60 and older have the highest risk followed by those who were between 45 and 59. Although the disease does not shy away from impacting young people, the goal is to identify the target we will focus on in building our awareness campaign and encourage folks to get screened.

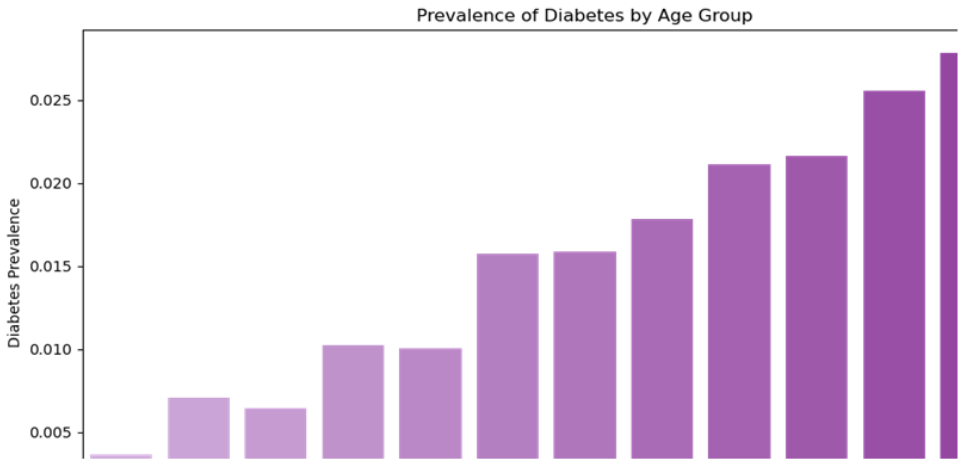

Figure 7: Prevalence of Diabetes by Age Group.

5.2. BMI

Not surprising, the more weight an individual has relative to their height, the higher the risk for having either Prediabetes or form of Type 2 Diabetes. The data clearly shows a direct relationship with a higher weight to height ratio increases the risk of Diabetes. For example, a BMI of greater or equal to 40 is associated with a much higher risk of Diabetes than a BMI of someone with less than 18.5.

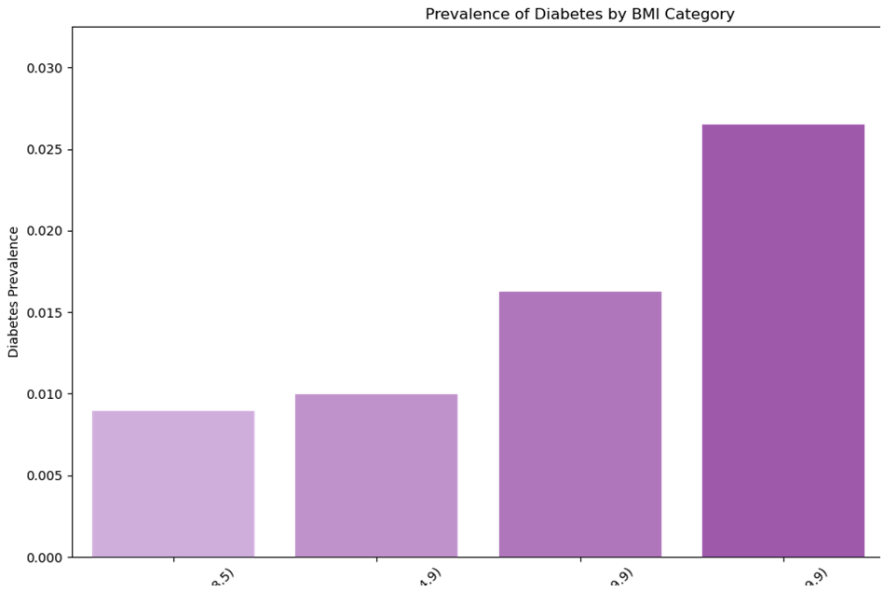

Figure 8: Prevalence of Diabetes by BMI Category.

5.3. General Health

I found it quite surprising that those that assessed their general health as Fair had a higher prevalence of Diabetes than those that assessed their health as Poor. However, the overall trend is intact such that those that are considered healthy or said to have Very Good health have a lower prevalence of Diabetes than those that are Fair or in Poor health.

It is important to note that the data is based on survey responses. These results could also be skewed by bias noise of individual responses. For example, if an individual may interpret their health as *Fair* when really it would be considered *Poor*. This could partly explain the findings that those with *Fair* general health have a higher prevalence in Diabetes than those with *Poor*.

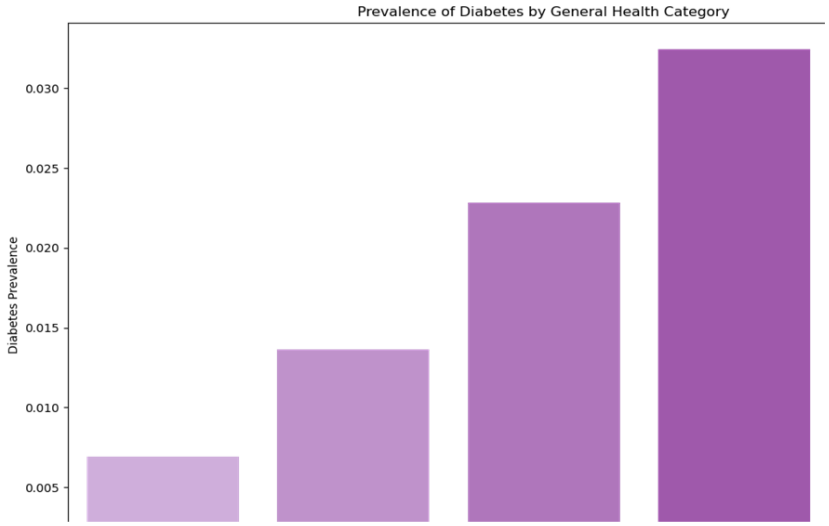

Figure 9: Prevalence of Diabetes by General Health Category.

6. CONCLUSION

In this analysis I was able to extract the top three important features that are associated with higher risk of Diabetes. These features include Age, BMI, and General Health. The models used in determining feature importance proved to perform well based on F1 scores.

The knowledge in this analysis can be leveraged to produce an awareness campaign and encourage folks that present with such risk factors to get screened for Diabetes and start taking control of their blood glucose levels. Specifically, those who are 45 and older, who have a BMI 25 or more, and / or those who do not consider their general health to be very good should get their blood glucose levels tested and screened for either Prediabetes or Diabetes in general.

For the future, I recommend the analysis be conducted against a larger set of features. In this paper only 17 features were reviewed. Moreover, I would attempt to see if a balanced dataset can be obtained without relying on resampling techniques to see if the results change. Perhaps, conducting more surveys with those that are Diabetics. In addition, I recommend engaging subject matter experts (SMEs) in the medical field, specifically, Endocrinologists, to gauge whether results presented in this paper are consistent with what SMEs see in their patients with Diabetes. Lastly, I would engage the SMEs on what they believe the risk factors to be and include these features in the analysis to see if our conclusions change after re-running the models.

REFERENCES

[1] Fortune. 2024. Half of U.S. Population Has Diabetes or Prediabetes, Experts Say. Retrieved August 19, 2024 from https://fortune.com/well/article/Diabetes-Prediabetes-obesity-half-united-states-population-insulin-wegovy-type1-type2-signs-symptoms/.

[2] Centers for Disease Control and Prevention. 2024. Behavioral Risk Factor Surveillance System (BRFSS). Retrieved August 19, 2024 from https://www.cdc.gov/brfss/index.html.

[3] World Health Organization. (n.d.). Diabetes. Retrieved August 19, 2024, from

https://www.who.int/health-topics/Diabetes#tab=tab_1.

[4] National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). What is Diabetes? Retrieved August 19, 2024, from https://www.niddk.nih.gov/health-information/Diabetes/overview/what-is-Diabetes.

[5] Mayo Clinic. (n.d.). Type 2 Diabetes: Symptoms and causes. Retrieved August 19, 2024, from https://www.mayoclinic.org/diseases-conditions/type-2-Diabetes/symptoms-causes/syc-20351193.

[6] Mayo Clinic. (n.d.). Gestational Diabetes: Symptoms and causes. Retrieved August 19, 2024, from https://www.mayoclinic.org/diseases-conditions/gestational-Diabetes/symptoms-causes/syc-20355339#.

[7] U.S. House of Representatives Diabetes Caucus. 2024. Facts and Figures. Retrieved August 19, 2024 from https://Diabetescaucus-degette.house.gov/facts-and-figures#:~:text=Type%201%20Diabetes%20(Body%20cannot,90%2D95%20percent%20of%20cases.

[8] UCLA School of Medicine. 2024. Can Diabetes Be Reversed? Retrieved August 19, 2024 from https://medschool.ucla.edu/news-article/can-Diabetes-be-reversed#:~:text=There's%20no%20cure%20for%20Diabetes,routine%20of%20diet%20and%20exercise.

[9] Mayo Clinic. (n.d.). Prediabetes: Symptoms and causes. Retrieved August 19, 2024, from https://www.mayoclinic.org/diseases-conditions/Prediabetes/symptoms-causes/syc-20355278#:~:text=Prediabetes%20means%20you.

[10] López, M., Blomberg, J., and Rodríguez, A. 2020. Prevalence of Hypertension and Its Association with Diabetes in Primary Care. International Journal of Clinical Research 12, 3 (2020), 205-214. DOI: 10.1007/s11606-020-05759-0. Retrieved August 19, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8050730/.

[11] Lima, F. M., Costa, T. M., & Dias, C. G. 2014. Adaptive Control of a Quadrotor UAV with a Robust Disturbance Observer. International Journal of Robotics Research 33, 8 (Aug. 2014), 1234-1249. DOI: 10.1177/0278364914542031. Retrieved August 19, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4166864/.

[12] Waleed Noori Hussein, Zainab Musahim Mohammed, Amani Naama Mohammed. 2022. Identifying risk factors associated with type 2 Diabetes based on data analysis. (Jan. 2022), 100-110. DOI: 10.1016/j.jpsychores.2022.03.012. Retrieved August 19, 2024, from https://www.sciencedirect.com/science/article/pii/S2665917422001775.

[13] G. J. V. Reddy. 2017. 7 Techniques to Handle Imbalanced Data. KDnuggets. Retrieved from https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html.

[14] J. Brownlee. 2019. SMOTE Oversampling for Imbalanced Classification. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[15] Brownlee, J. (2021, October 13). XGBoost for imbalanced classification. Machine Learning Mastery. https://machinelearningmastery.com/xgboost-for-imbalanced-classificatio/.

AUTHOR

Oleg Fleitman is a distinguished leader with over a decade of experience in Finance, Risk Management, Artificial Intelligence (AI), Machine Learning (ML), and Human Resources. His rewarding career spans across Fortune 500 companies operating globally, including sectors such as Financial Services, Manufacturing, Healthcare, Government, and Start-Ups. Known for blending technology, data science, and business strategy, Fleitman has a proven track record of delivering practical solutions to complex challenges. Renowned for developing custom in-house solutions, Fleitman continuously pushes the envelope in uncharted territories, particularly in enhancing risk and control mechanisms within global financial institutions. His leadership in managing international teams within startup environments and implementing AI-powered solutions has consistently driven business results while upholding the highest ethical standards. With an academic background in Finance and Computer Science and an MBA with double majors in Management Analytics and Brand Management, Fleitman exemplifies transformative leadership at the intersection of Financial Services and Artificial Intelligence. His work not only addresses the most pressing needs of organizations but also advances innovative approaches to driving growth and operational efficiency across industries.