Diabetes Risk Predictive Model Based on Age, BMI, and Physical Activity: Predictive Analysis Project

Tace Harris
Quantitative Methods in The Social Sciences, The University of Michigan

Methods

To create the model, the dataset Diabetes Health Indicators from the Center for Disease Control was used. In order to create a model to predict the chances a set of people have diabetes or not, the programming language Python was utilized. First, an exploratory analysis was conducted to determine what variables are included in the dataset and how they are distributed.

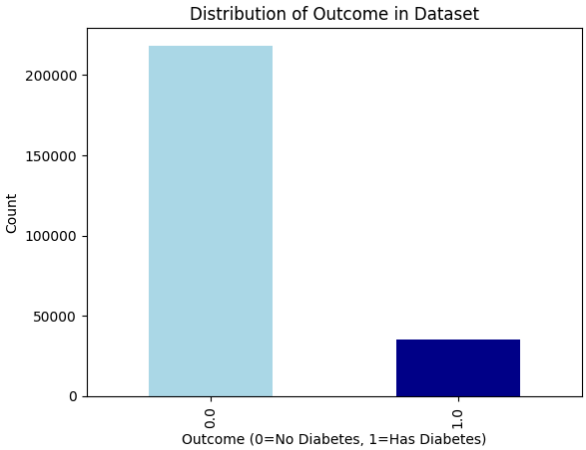
Then, three variables from the dataset were chosen to be included in the model; Age, BMI, and physical activity. Age is coded as 13 categories; 1 = 18-24, 2 = 25-29, 3 = 30-34, and so on. BMI or Body Mass Index is coded as a numerical variable with ranges from 0-65. BMI is categorized as underweight (<18.5), healthy weight (18.5-24.9), overweight (25-29.9), and obese (30 or higher). Finally, physical activity is coded as a binary variable, a 0 meaning the person does not participate in physical activity and a 1 meaning they do.

These three predictor variables were selected to represent different categories relevant to diabetes risk. The first category includes modifiable variables, such as physical activity and BMI. Physical activity is something an individual can choose to participate in or not, making it modifiable. BMI, while partially determined by genetics due to factors like body composition and metabolism, can still be modified through diet and lifestyle choices. The second category is a non-modifiable variable, Age. This variable represents how old a person is, which cannot be altered by behavior or genetics.

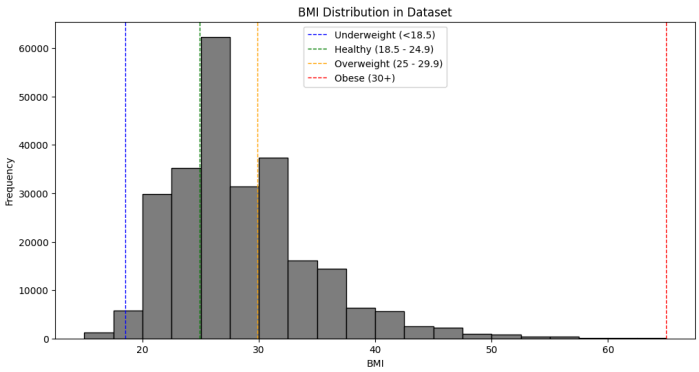
After the predictor variables were selected, a multivariate logistic regression model was built using the *statsmodel* library. Then the same model was built using the *sklearn* library to predict the likelihood of having diabetes based on the three variables. The accuracy of the model was evaluated by generating an accuracy report and confusion matrices, which were also conducted on a model where class (diabetes or no diabetes) was weighted to potentially help eliminate accuracy issues due to an imbalance of the dataset.

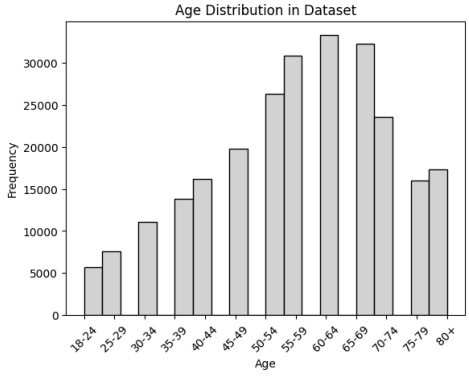
Results/Analysis

The exploratory analysis revealed that the dataset is extremely unbalanced with only 16% of the cases being diabetes (*figure 1*).



To explore the chosen variables, *matplot library* was used to create two charts, each demonstrating the distribution of the two numeric variables. The BMI distribution chart (*figure 2*) showed that most of the dataset subjects are in the healthy and overweight BMI ranges. The age distribution chart (*figure 3*) shows that most of the dataset's subjects are in their 50s-70s.



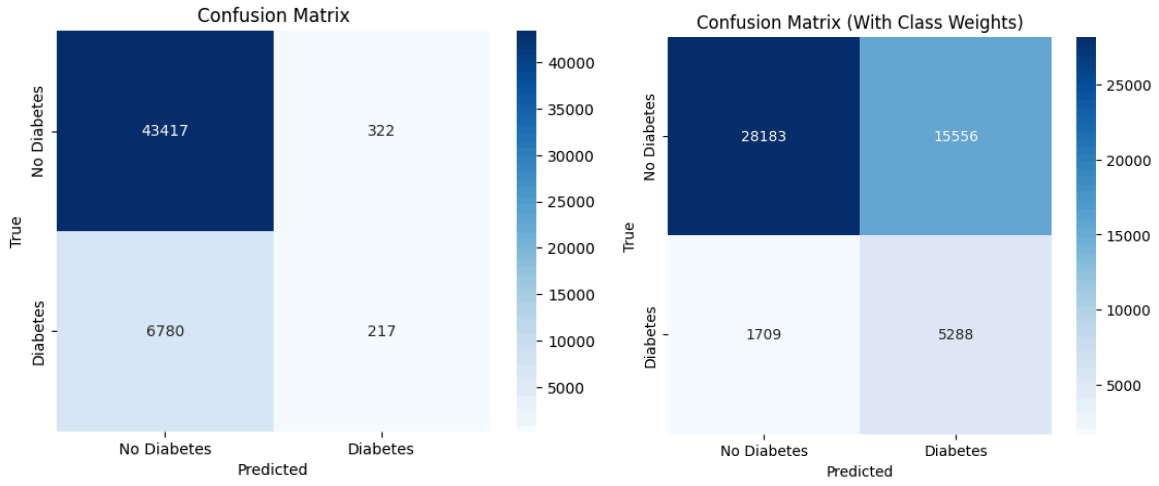


The first model created using *statsmodel* revealed that each variable is statistically significant. Specifically, age and BMI are positively correlated with the risk of developing diabetes, while physical activity is negatively correlated with the risk. The model showed that for each one-unit increase in age, the log odds of having diabetes increased by 22%. Similarly, for each one-unit increase in BMI, the log odds of developing diabetes increased by 8%. In contrast, participating in physical activity reduced the log odds of having diabetes by 42%. These results indicate that being older, overweight, and not participating in physical activity are all associated with a higher risk of developing diabetes.

Discussion

The two models created to predict outcomes had differing results. The model without class weights has an accuracy of 86%, which is relatively high. However, the performance is skewed towards correctly predicting non-diabetic individuals, with a precision of 0.86 and a recall of 0.99 for this class. This suggests that the model is very effective at identifying non-diabetic individuals but struggles significantly with predicting diabetic individuals, where the precision is low at 0.40 and the recall is very poor at 0.03. This results in a very low F1-score of 0.06 for diabetes, indicating that the model fails to correctly identify these cases.

When class weights are applied, the accuracy drops to 0.66, indicating a decrease in the overall performance of the model. However, there is a shift in how the model handles both classes. The precision for non-diabetic individuals improves to 0.94, but recall drops to 0.64, meaning the model is now more conservative about predicting non-diabetic cases and misses some of them. The precision for diabetic individuals is low at 0.25, but the recall significantly improves to 0.76, meaning the model is now much better at identifying diabetic cases (although with many false positives). The F1-score for class 1.0 is 0.38, which is still low but shows an improvement compared to the model without weights. The confusion matrices for the model without weights (*figure 3*) and the model with weights (*figure 4*) illustrate the difference in prediction accuracy.



Conclusion

In conclusion, through Python and various libraries, a model was created to predict the likelihood of a person having diabetes based on the variables age, BMI, and physical activity participation. Logistical regression analysis showed that each variable is statistically significant and impacts the likelihood of having diabetes. Of the two models created, it was determined that the unweighted model was more accurate. It offers better overall accuracy and provides strong performance in predicting non-diabetic individuals, which may be the more important class in certain contexts where it is wanted to avoid false positives for diabetic predictions. However, if correctly identifying diabetic individuals is more critical (e.g., in medical diagnosis), then a model with class weights might be more appropriate despite the drop in overall accuracy.