Kelley Denny
Student ID 800563382
ITSC 3162 - Intro to Data Mining
Project 2 - Classification Modeling

Link to Dataset: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Introduce the Problem:

Diabetes is a serious chronic health condition that is characterized by heightened levels of sugar in the bloodstream, otherwise known as blood glucose levels and can directly result in a reduction of quality of life and life expectancy. Additionally, there are three different types of diabetes that have been identified, and these are Type 1 diabetes, Type 2 Diabetes, and Gestational Diabetes (Roglic, 4). Type 1 Diabetes is irreversible, is most typically seen in younger people, and studies have not been able to conclusively determine risk factors that contribute to the development of Type 1 Diabetes. Type 1 Diabetes is a result of the pancreas under producing insulin to clear sugar from the bloodstream, so Type 1 Diabetes patients will always have to get insulin from external sources, usually via injection (Bell, 2024). However, Type 2 Diabetes is a result of insulin resistance and is more generally understood and able to be predicted based on certain health factors and habits outside of the genetic disposition (Roglic, 4). Type 2 Diabetes is also generally viewed as being preventable or reversible with lifestyle and dietary changes (Bell, 2024). This project aims to train a classification model on a Diabetes health indicators dataset. Specifically the target variable for the project is to classify a person as Non-Diabetic, Pre-Diabetic, or Diabetic based on the most highly correlated health indicators (outside of genetic disposition) in the dataset, which were identified as BMI, High Blood Pressure (HighBP), High Cholesterol levels (HighChol), and General Health rating on a scale of 1-5 (GenHlth).

Introduce the Dataset:

This Diabetes health indicators dataset contains 253,680 rows and has 22 columns, and was obtained through Kaggle. The target variable column is ‘Diabetes012’, which is the classification of Diabetes, with 0 being Non-Diabetic (or only having Gestational Diabetes), 1 being Pre-Diabetic, and 2 being Diabetic. The features of the dataset contain different Health indicators such as Blood Pressure, Cholesterol, BMI, Heart disease, whether or not a person has had a stroke, whether or not a person smokes, and dietary and exercise habits among others.

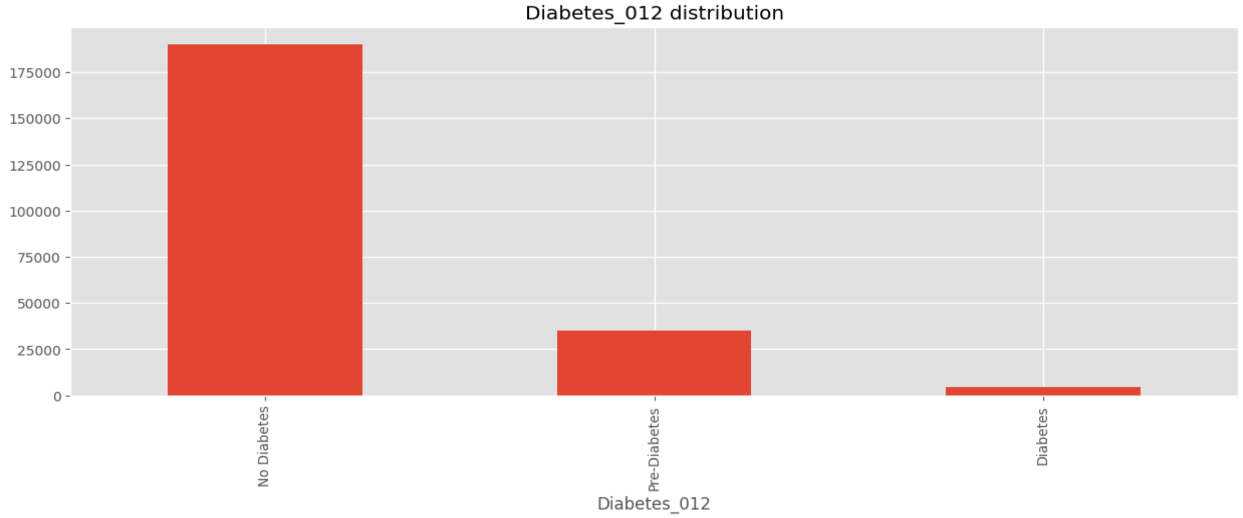
Pre-processing:

In the exploratory data analysis portion of the project I found that there were no missing values found in the pre-processing stage so I did not need to handle any missing data. I did find quite a few duplicate values in the dataset, specifically there were 23,899 duplicate rows in the data set so I dropped the duplicates, keeping the original in place. I also found that the dataset was unbalanced, as there were 190,055 Non-Diabetic entries, 35,097 Pre-Diabetic entries, and 4,629 Diabetic entries. I chose to use SMOTE prior to modeling the dataset to oversample the diabetic entries so that there would be more representation among that class so that the model could more accurately predict the Diabetic patients. I also scaled some of the features of interest—specifically BMI, General Health, Age, Mental Health, Physical Health, and Income—as these were on a larger scale compared to the 0–1 scale on which most other features were measured.

Data Understanding/Visualization:

I conducted an exploratory data analysis on the data set to get a feel for the distribution of the dataset as well as to see how the predictor variables influenced each other and the target variable. I found that the dataset was heavily skewed toward the non-diabetic variable as 83% of the observations were classified as Non-Diabetic. About 15% of the observations were classified as Diabetic, and the remaining 2% were classified as Pre-Diabetic. I created two visualizations for this distribution, one being a count plot (Figure 1) and the other a pie chart (Figure 2).

Figure 1



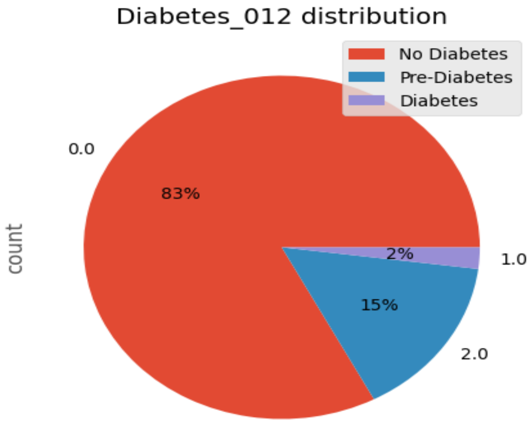


Figure 2

I then wanted to explore the correlation that the different variables in the dataset had with the Diabetes classification target variable. I created a bar plot (Figure 3) to show both positive and negative correlation that the predictor variables had with the Diabetes classification. For additional analysis I also created a heat map of all of the variables (Figure 4).

Figure 3

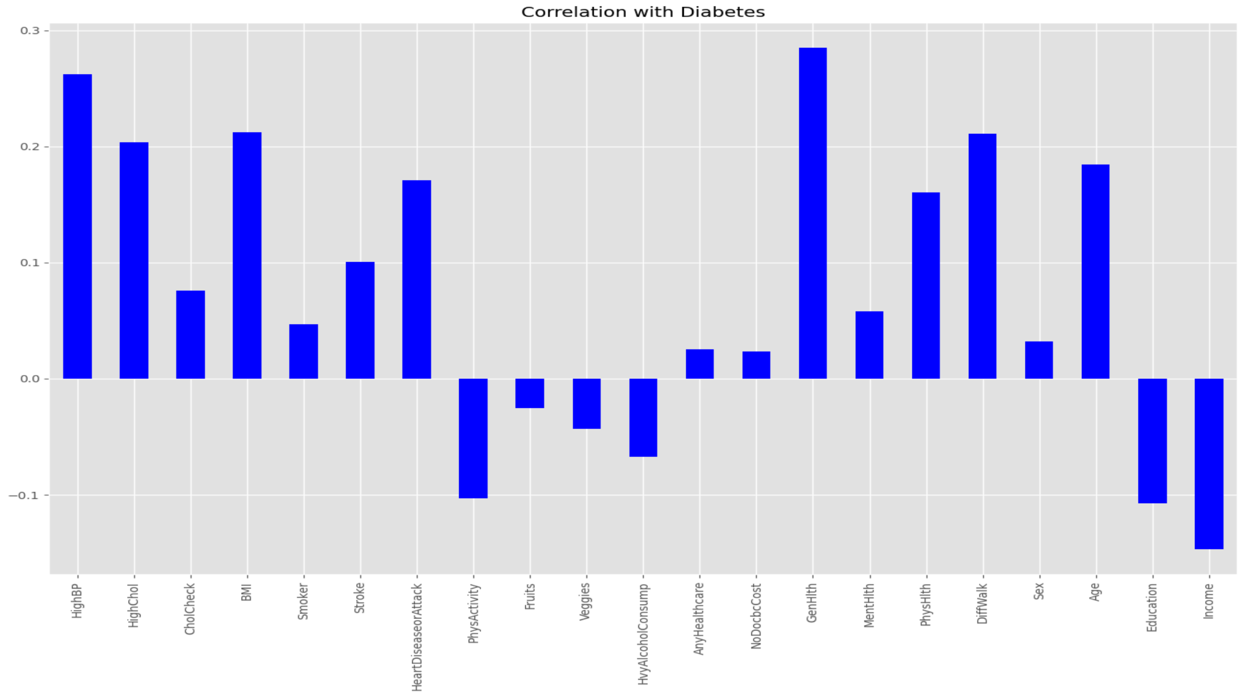
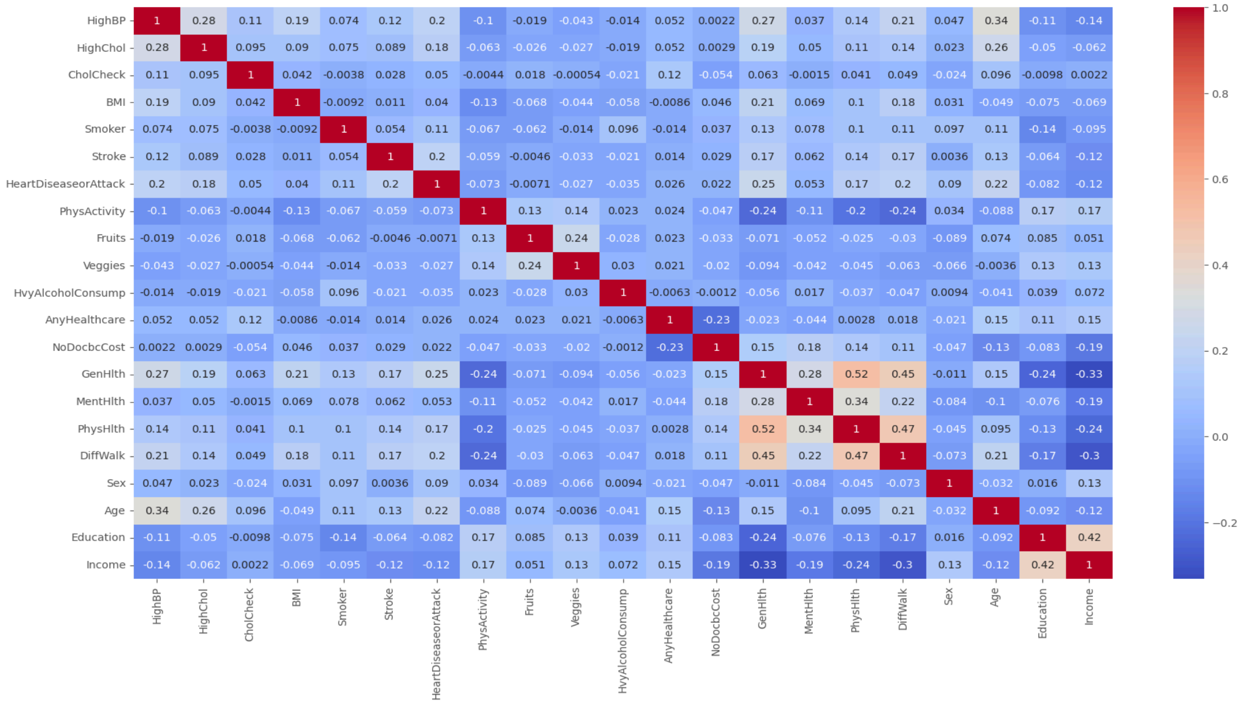
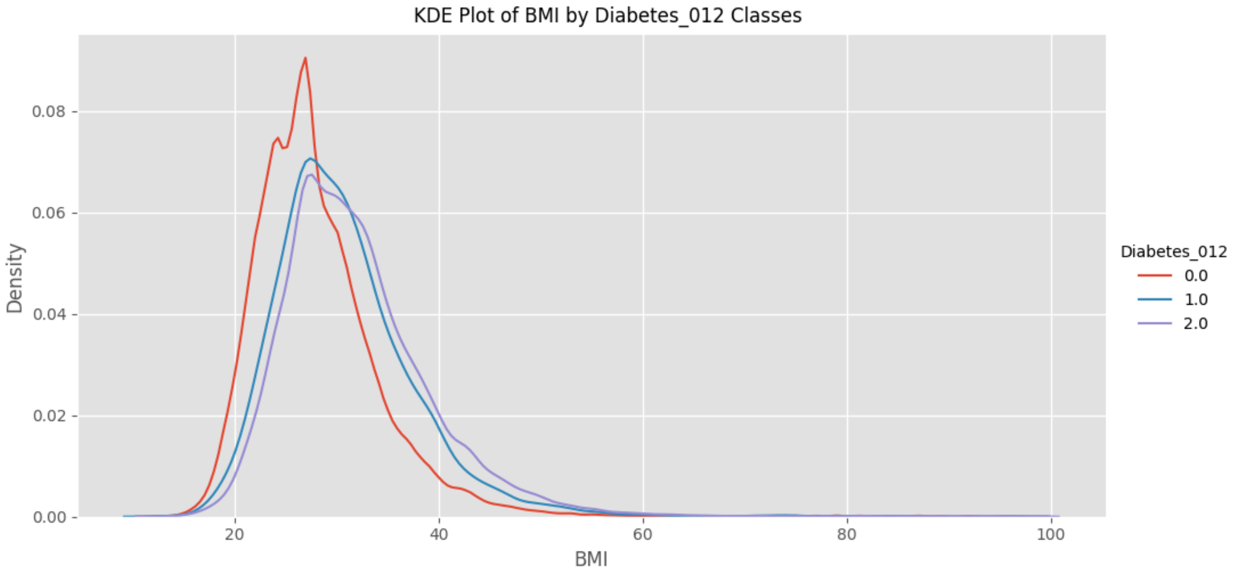


Figure 4



Through these visualizations I was able to understand that the predictors that were most positively correlated with Diabetes are High Blood Pressure, High Cholesterol, BMI, Heart Disease or Heart Attack, General Health, Difficulty with walking, and Age. Education, Income, and higher physical activity were the most negatively correlated features with the target variable. As commonly stated, correlation does not equal causation, so some of these positively correlated features could either be at play for a person developing diabetes, or they could be a result from the person developing Diabetes. Finally I wanted to see the distribution of BMI levels across the three classifications so I created a Kernel Density plot to show the distribution of BMI (Figure 5).

Figure 5



This plot shows that Pre-Diabetic and Diabetic patients tend to have a higher BMI than Non-Diabetic patients.

Some key insights from this data analysis show that some other negative health indicators such as High Blood Pressure, High Cholesterol, and higher BMI are correlated with Diabetes, and perhaps combating these indicators could help to reduce the number of people found to have either Diabetes or Pre-Diabetes. Diet and exercise are often prescribed as lifestyle modifications to help reduce BMI, blood pressure, and reverse Diabetes in Type 2 or Gestational Diabetes patients (Bell, 2024).

Modeling:

For the model selection I chose to first run a Decision Tree Classifier as it is easily able to be visualized to see exactly how this Classifier is reaching the final decision for the target variable classification. I created the model with a max depth of 6 so that the visualization would still be able to be understood, but the Classifier model would not be limited in the amount of splits it could make from the root node down the branches to the final leaves of the tree. I then evaluated the performance metrics of the model by finding the accuracy score, creating a confusion matrix, and finding the F1 score for the classifier. The accuracy score was .7319, meaning that the model was roughly 73% accurate in its classifications. This seems quite high but can be misleading on its own as I identified a class imbalance already in the dataset. The confusion matrix and F1 score will better be able to determine how the classifier model works. The confusion matrix is shown below.

```
Confusion Matrix: [[29327 0 8509] [ 522 0 386] [ 2903 0 4310]]
```

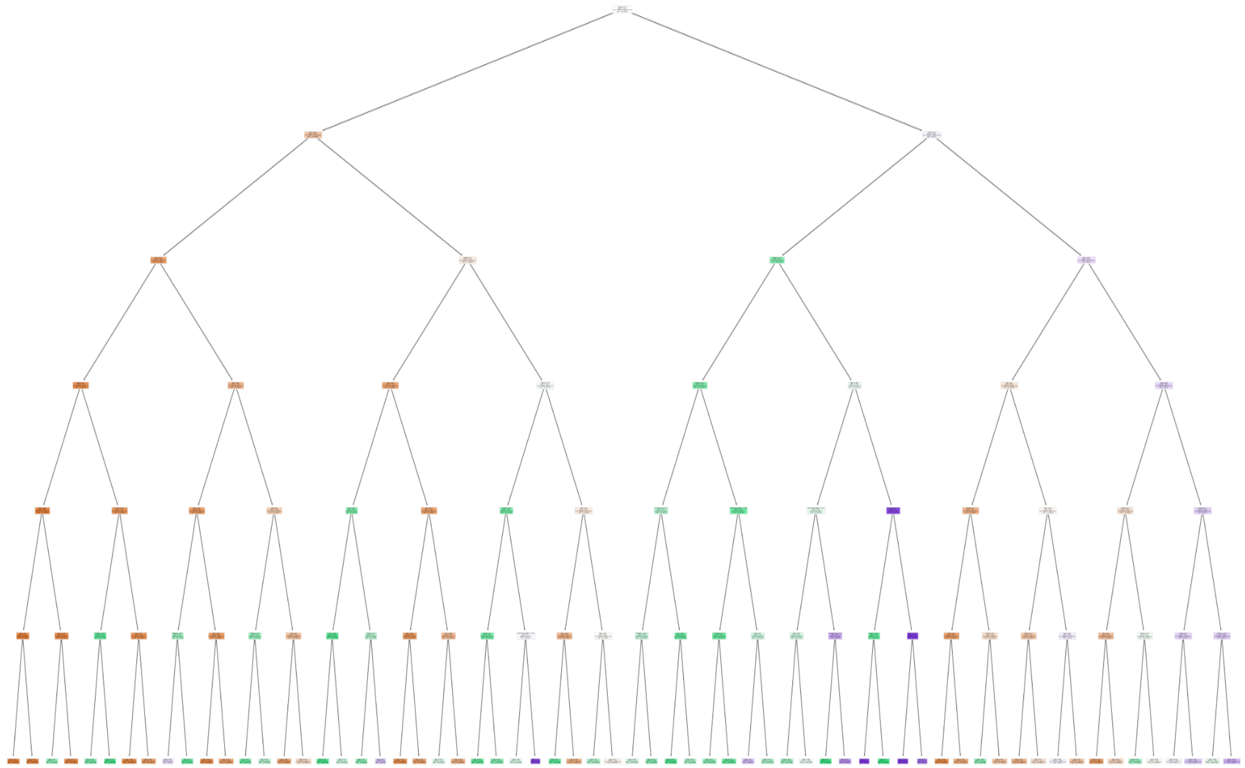
The results of this show that for Class 0 (Non-Diabetic patients) there were 29,327 true positives (Non-Diabetic patients correctly classified as Non-Diabetic) and 8,509 false negatives (Non-Diabetic patients classified as being Diabetic) with no predictions for the Pre-Diabetic class. For Class 1 (Pre-Diabetic patients) there were no true positives predicted, and there were 522 false negative predictions for Non-Diabetic classification and 386 false negative predictions for Diabetic classification. Finally for the Class 2 (Diabetic patients) there were 2903 false negative predictions for the Non-Diabetic Class, no predictions for the Pre-Diabetic class, and 4310 true positives for the Diabetic class. This shows that the decision tree classifier struggles greatly with the Pre-Diabetic classification and chooses to not make any predictions for this class. The accuracy mixed with this confusion matrix shows that there is a class imbalance and that the accuracy is truly misleading as the classification can just predict the majority class and be right a majority of the time. Finally I looked at the F1 score for the classifier and found it to be .4177. This score is the harmonic mean between precision and recall, and a score of .4177 shows that precision and recall are not well balanced. This means that either the model struggles to make correct true positive predictions, or makes a lot of misclassifications.

Next I chose to create a Random Forest classifier as it builds on the strengths of the decision tree classifier and is an ensemble method that takes in many different Decision Trees and finds the most likely outcome based on the decisions of all the trees combined. This classifier did perform better than the single Decision Tree with an accuracy score of .8055 meaning that it correctly predicts about 81% of the observations in the training data. The confusion matrix for the Random Forest classifier is shown below:

```
[[34432 130 3274] [ 707 7 194] [ 4594 40 2579]]
```

These numbers were a bit different and the true positives for each class did seem to be somewhat higher with the exception of true positives for Class 2 (Diabetic patients). The Random Forest model did predict some observations as being Pre-Diabetic as well which shows some slight improvement over the single Decision Tree. Finally the F1 score for the Random Forest classifier was .4299 which is also a slight improvement over the Decision Tree F1 score but is still quite low overall.

The visualization for the Decision Tree Classifier is pictured below:



The visualization was useful for seeing how the Decision Tree classifier works, but I was not able to get a good visual that showed the labels for each level of the tree.

Storytelling:

This project shows that medical science can still face challenges with properly classifying or diagnosing a person with disease even with the markers of health indicators. Certain machine learning models don’t perform as well when it comes to handling class imbalance and thus machine learning can still be limited in applications of medicine. In example, the Decision Tree classifier failed to label any of the observations as Pre-Diabetic and receiving this diagnosis could help a person to reverse their inevitability of developing Type 2 Diabetes. Both models had a misleading high accuracy score but with taking other metrics in like F1 score and the confusion matrix showed that the accuracy alone is not enough to measure the robustness of the machine learning model. These models could serve as a good baseline for improving upon in future studies of Diabetes classification however. Improving upon methods for handling class imbalance, changing the parameters, and adding more data to the study would help to build upon the work that these initial machine learning models have done. With improvement in this area, medical science could see better predictive powers with classifying diseases or the onset of diseases and could help people to improve their health outcomes and reverse the proclivity towards developing these diseases or disorders.

References

Bell, Ashley. “Can Diabetes Be Reversed?: Research Spotlight.” *UCLA Medical School*, 31 July 2024, medschool.ucla.edu/news-article/can-diabetes-be-reversed

Roglic, Gojka. WHO Global report on diabetes: A summary. International Journal of Noncommunicable Diseases 1(1):p 3-8, Apr–Jun 2016. DOI: 10.4103/2468-8827.184853