Machine Learning I
Group 2 Final Project
Date: 03/11/2024

Diabetes Prediction Model

Iris Huang, Yumeng Li, Hira Stanley, Paul Pham-Ly, Michelle Zhu

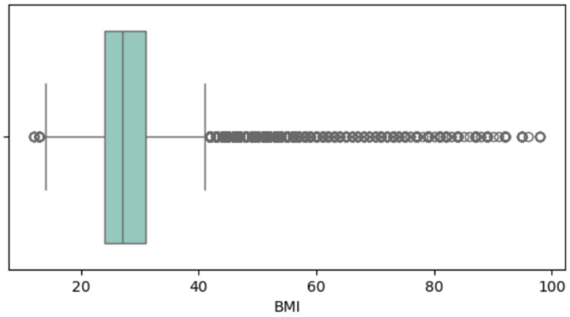
Business Problem

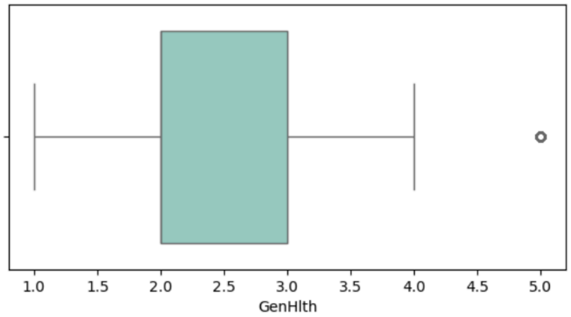
Diabetes is a chronic condition that affects millions of people worldwide with significant impacts on health and quality of life. It is becoming increasingly more prevalent due to several key reasons including urbanization, lifestyle changes, global aging population, and genetics. Understanding the patterns and relationships in the factors influencing the development of diabetes is critical for timely intervention and management. This project aims to develop a machine learning model that will accurately predict the risk of diabetes in individuals by analyzing which characteristics contribute significantly to a diagnosis.

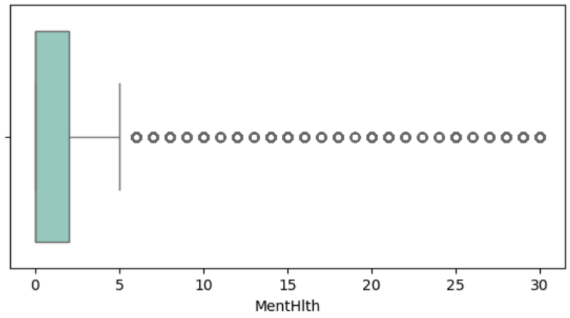
Data Description

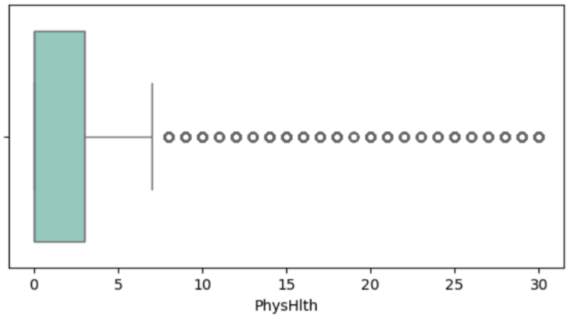
The original data is from the Behavioral Risk Factor Surveillance System (BRFSS) telephone survey conducted by the CDC in 2015 with over 400,000 responses. The data is a subset of responses that include the variables suspected to be related to diabetes. This subset contains 253,680 observations with 21 feature variables and 1 binary target variable (0 = no diabetes, 1 diabetes). The features include demographic, lifestyle, and medical history characteristics that are a mix of integers, categorical, and binary variables.

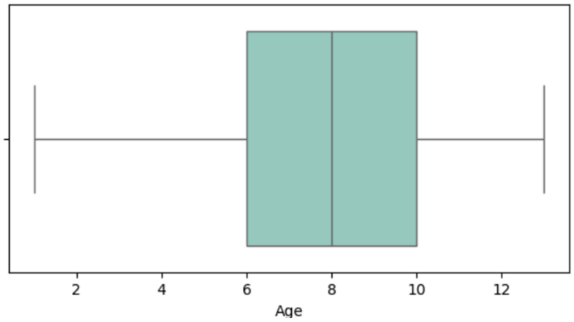
First, we convert data type from float to integer, then use box plots to check for outliers in non-binary columns (see box plots below). BMI, Mental health and Physical Health appear to have significant outliers that might skew the data.

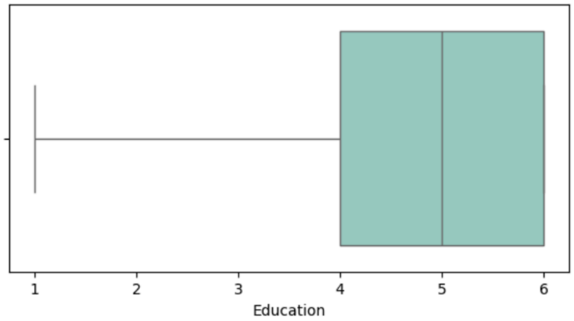


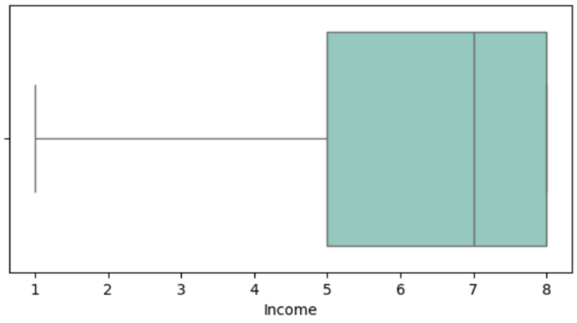






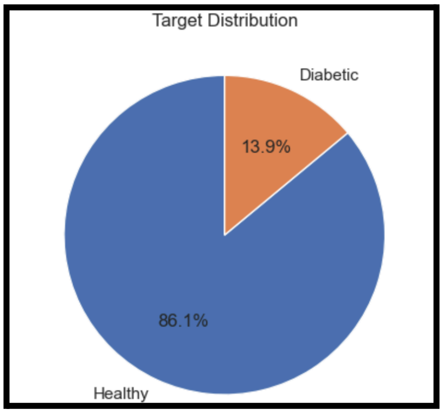


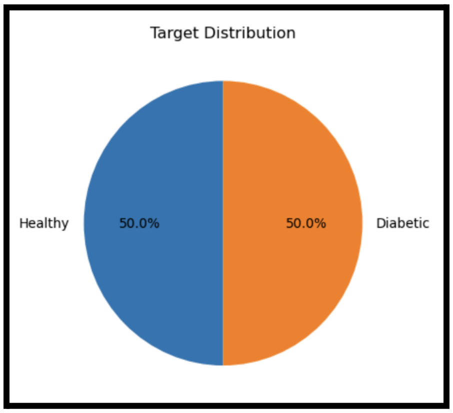




The original data set was imbalanced where 14% (35,346) of the observations are labeled 1 (diabetes) and about 86 percent are healthy. We applied SMOTE technique to oversample and Near Miss to undersample resulting in a 50/50 balanced dataset.

The resulting data set, and the one we base this project off, contains 70,691 entries, with the final target distribution being 50% diabetic, and 50% non-diabetic.





Data Table Schema

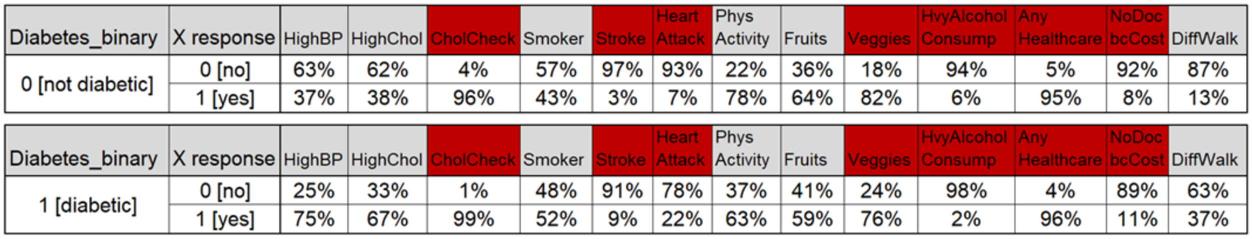
*For these binary variables, 0 = no and 1 = yes.

Variable	Variable Definition	Data Type	Notes
Diabetes_binary*	Diagnosed with diabetes	binary	Near 50/50 split
HighBP*	High blood pressure	binary	Near 50/50 split
HighChol*	High cholesterol	binary	Near 50/50 split
CholCheck*	Cholesterol check in 5 years	binary	97% is 1
BMI	body mass index	integer	40% in 25-30 (overweight), 21% in 18-24 (healthy), 21% in 31-35 (obese I), 17% (obese II-III), histogram is right skewed
Smoker*	Smoked at least 100 cigarettes in entire life	binary	Near 50/50 split
Stroke*	Ever told you had a stroke	binary	94% is 0
HeartDiseaseorAttack*	Coronary heart disease (CHD) or myocardial infarction (MI)	binary	85% is 0
PhysActivity*	Physical activity in past 30 days not including job	binary	70% is 1
Fruits*	Consume fruit 1 or more times per day	binary	61% is 1
Veggies*	Consume vegetables 1 or more times per day	binary	79% is 1
HvyAlcoholConsump*	Consume >=14 drinks per week (adult men) or >=7 drinks per week (adult women)	binary	96% is 0
AnyHealthcare*	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc.	binary	95% is 1
NoDocbcCost*	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?	binary	91% is 0
GenHlth	General health condition rating 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor	categorical	33% is 3 28% is 2 19% is 4 12% is 1 8% is 5
MentHlth	# of days of poor mental health in last 30 days	integer	68% is 0, majority of rest spread across 1-5, 10, 15, 20, or 30 days

PhysHlth	# of days with physical illness or injury in past 30 days	integer	56% is 0, 21% is 1-7, majority of rest spread across 10, 15, 20, or 30 days
DiffWalk*	serious difficulty walking or climbing stairs	binary	74% is 0
Sex	0 = female 1 = male	binary	Near 50/50 split
Age	14-level age category 1 = 18-24yrs 2 = 25-29yrs 3 = 30-34yrs etc	categorical	53% in 8-11 (55-74yrs) with largest category of 15% in 10 (65-69yrs), histogram is left-skewed
Education	Highest grade or year of school completed 1 = never or only kindergarten 2 = Grades 1-8 3 = Grades 9-11 etc	categorical	37% is 6 (college graduate), 28% is 5 (some college or technical school), 28% is 4 (high school graduate)
Income	8-level scale of annual household income 1 = <$10,000 2 = <$15,000 3 = <$20,000 etc	categorical	29% is 8 ($75k+), 16% is 7 ($50-75k), 15% is 6 ($35-$50k)

Data Exploration

To better understand the roles of each variable, we analyzed the spread of the data across all the explanatory variables with respect to the target variable. This was done by cross tabulation analysis. We found that the following variables highlighted in red are likely to be uncorrelated with diabetes due to having a high percentage (>75%) of one particular response whether they had diabetes or not. For example, most respondents in the data set answered no to having had a stroke in the past for both the diabetic and non-diabetic group.



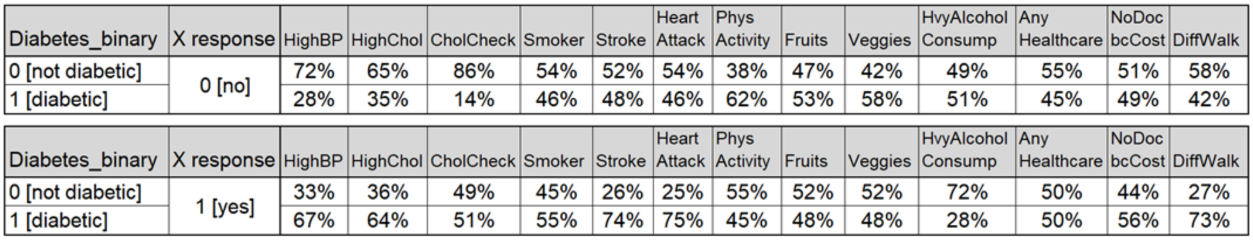
Diabetes_binary	X response	HighBP	HighChol	CholCheck	Smoker	Stroke	Heart Attack	Phys Activity	Fruits	Veggies	HvyAlcohol Consump	Any Healthcare	NoDoc bcCost	DiffWalk
0 [not diabetic]	0 [no]	63%	62%	4%	57%	97%	93%	22%	36%	18%	94%	5%	92%	87%
	1 [yes]	37%	38%	96%	43%	3%	7%	78%	64%	82%	6%	95%	8%	13%

Diabetes_binary	X response	HighBP	HighChol	CholCheck	Smoker	Stroke	Heart Attack	Phys Activity	Fruits	Veggies	HvyAlcohol Consump	Any Healthcare	NoDoc bcCost	DiffWalk
1 [diabetic]	0 [no]	25%	33%	1%	48%	91%	78%	37%	41%	24%	98%	4%	89%	63%
	1 [yes]	75%	67%	99%	52%	9%	22%	63%	59%	76%	2%	96%	11%	37%

Interestingly, blood pressure and cholesterol levels hint at correlation with diabetes as indicated by their distribution going in opposite directions depending on their diabetic status. For both variables, >60% of non-diabetic respondents answered no to having high levels, while >60% of diabetic respondents answered yes to having high levels. The remaining variables did not exhibit interesting patterns that suggest correlation with the target variable at this point.

Next, we viewed cross tabulation by the response variable in the table below. Again, blood pressure and cholesterol levels suggest a correlation with diabetes. For respondents with low levels, <40% were diabetic, and for respondents with high levels, >60% were diabetic. Stroke and HeartDiseaseorAttack have interesting results, where for those who have not had such an event in the past are split about 50/50 non-diabetic and diabetic, and those who have are split about 25/75.

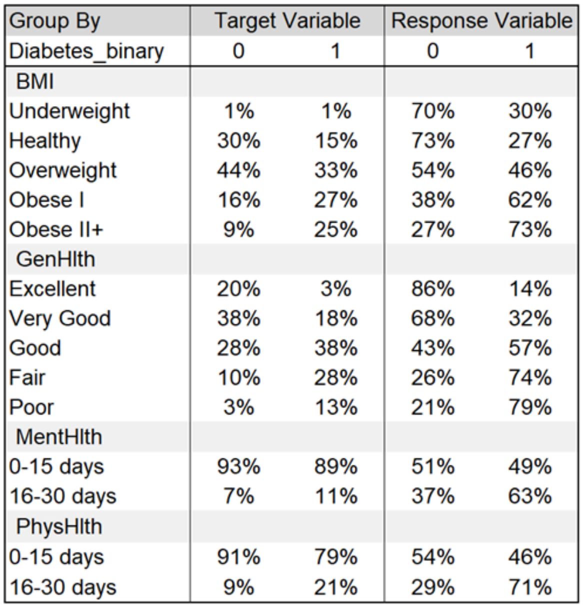
CholCheck, AnyHealthcare, and NoDocbcCost continue to exhibit signs of no correlation to diabetes, as suggested by their near 50/50 split whether the respondent answered yes or no. CholCheck showed 86% of respondents answering no to be not diabetic. However, this could be due to the lack of need to check cholesterol levels if there are no signs of diabetes.



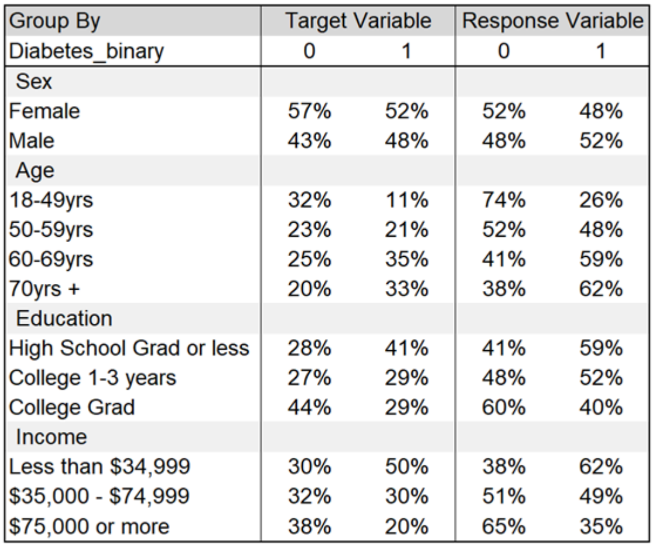
Diabetes_binary	X response	HighBP	HighChol	CholCheck	Smoker	Stroke	Heart Attack	Phys Activity	Fruits	Veggies	HvyAlcohol Consump	Any Healthcare	NoDoc bcCost	DiffWalk
0 [not diabetic]	0 [no]	72%	65%	86%	54%	52%	54%	38%	47%	42%	49%	55%	51%	58%
1 [diabetic]		28%	35%	14%	46%	48%	46%	62%	53%	58%	51%	45%	49%	42%

Diabetes_binary	X response	HighBP	HighChol	CholCheck	Smoker	Stroke	Heart Attack	Phys Activity	Fruits	Veggies	HvyAlcohol Consump	Any Healthcare	NoDoc bcCost	DiffWalk
0 [not diabetic]	1 [yes]	33%	36%	49%	45%	26%	25%	55%	52%	52%	72%	50%	44%	27%
1 [diabetic]		67%	64%	51%	55%	74%	75%	45%	48%	48%	28%	50%	56%	73%

For the multinomial and ordinal categorical variables, the tables below show cross tabulation analysis grouped by target variable (ex. sum of BMI spread for target = 0 is 100%) and by response variables (ex. sum of target variable spread is 100% for BMI = Underweight). Looking at the cross analysis side by side provides interesting insights about which variables potentially play a role in classifying diabetic and non-diabetic individuals.



Group By	Target Variable		Response Variable	
Diabetes_binary	0	1	0	1
BMI				
Underweight	1%	1%	70%	30%
Healthy	30%	15%	73%	27%
Overweight	44%	33%	54%	46%
Obese I	16%	27%	38%	62%
Obese II+	9%	25%	27%	73%
GenHlth				
Excellent	20%	3%	86%	14%
Very Good	38%	18%	68%	32%
Good	28%	38%	43%	57%
Fair	10%	28%	26%	74%
Poor	3%	13%	21%	79%
MentHlth				
0-15 days	93%	89%	51%	49%
16-30 days	7%	11%	37%	63%
PhysHlth				
0-15 days	91%	79%	54%	46%
16-30 days	9%	21%	29%	71%



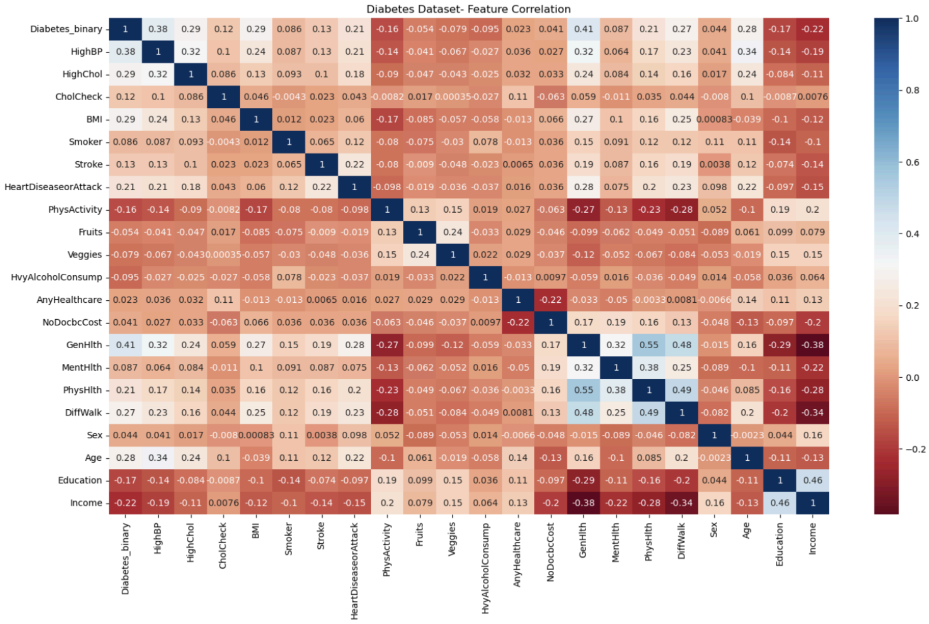
Group By	Target Variable		Response Variable	
Diabetes_binary	0	1	0	1
Sex				
Female	57%	52%	52%	48%
Male	43%	48%	48%	52%
Age				
18-49yrs	32%	11%	74%	26%
50-59yrs	23%	21%	52%	48%
60-69yrs	25%	35%	41%	59%
70yrs +	20%	33%	38%	62%
Education				
High School Grad or less	28%	41%	41%	59%
College 1-3 years	27%	29%	48%	52%
College Grad	44%	29%	60%	40%
Income				
Less than $34,999	30%	50%	38%	62%
$35,000 - $74,999	32%	30%	51%	49%
$75,000 or more	38%	20%	65%	35%

It can be noted that BMI has similar spread for respondents with and without diabetes, but for people with diabetes, there is a higher percentage of obese individuals. Variables related to health conditions showed that the majority of diabetic respondents rated themselves with generally favorable health physically, mentally, and generally. However, looking at the data grouped by response variable, it is notable that in the unhealthier subsets of respondents, there is a higher percentage of diabetic individuals. For example, 86% of GenHlth = Excellent respondents are non-diabetic, whereas 79% of GenHlth = Poor respondents are diabetic. This pattern is seen for MentHlth and PhysHlth as well.

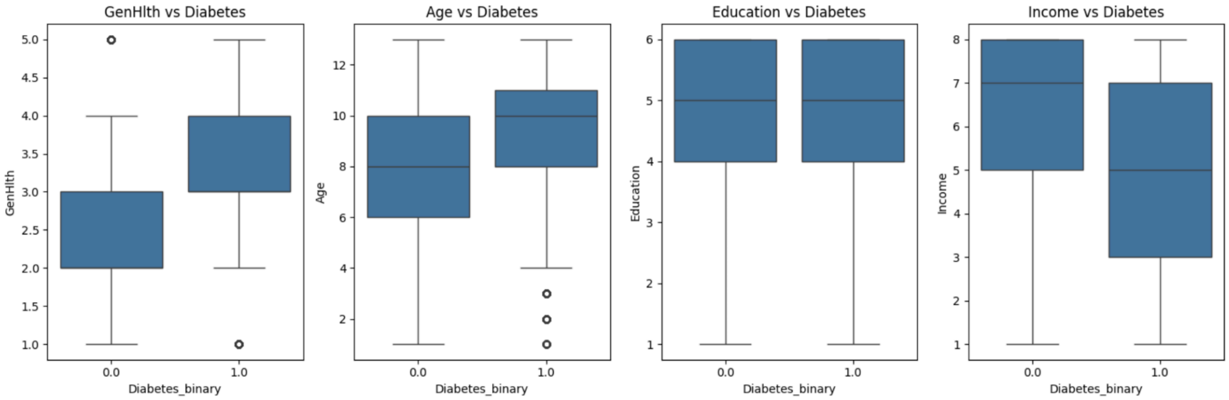
Age is spread rather evenly amongst non-diabetic respondents; however, it skews towards elderly individuals in the diabetic subset. There is also a higher percentage of non-diabetic individuals for younger respondents and the opposite for older respondents. A similar pattern is observed in the Education and Income variables, where non-diabetic respondents have a roughly even spread across categories but exhibit higher percentages of lower education and lower income for diabetic respondents. Grouped by response variables, a higher percentage of lower education and lower income respondents reported diabetic. Sex appears to be an unremarkable variable as indicated by near 50/50 split whether grouped by target or response variable. It must be noted that there may be bias introduced by the added layer of grouping of the response variable categories.

EDA

Our initial correlation matrix, prior to any transformations, revealed some important preliminary relationships. Certain health factors, such as high blood pressure, BMI, and self-reported general health on a scale of 1-5 had the strongest predictive power on our target variable. It is also important to note that several variables had an inverse relationship with our target variable. For example, income, education, and physical activity all move in the opposite direction of diabetes. Lastly, although not high enough to be considered multicollinear, certain variables have strong correlations that we would consider dropping, such as general health, physical injuries, and difficulty walking. Intuitively, they all represent various indications of physical ability.



We also took deeper consideration of selected variables against our target variable (Diabetes_binary) and plotted their distributions; some interesting relationships were revealed. For instance, individuals with diabetes tended to be older on average and had lower average incomes compared to those without diabetes. These trends highlight potential sociological factors and disparities associated with diabetes.



We conducted a Variance Inflation Factor test to ensure there was no multicollinearity present. All variables revealed a score of less than 2.0, meaning there was a moderate but acceptable level of multicollinearity present.

Model Selection & Training

The dataset was split into 80% training and 20% testing to ensure model generalization. The split was stratified to maintain class balance. The training set was used for model fitting and hyperparameter tuning, while the test set provided an independent evaluation of model performance.

Model Selection

Several machine learning models were trained and evaluated using accuracy, F1-score, precision as performance metrics:

- The baseline model Logistic Regression: This baseline model provided an F1-score of 0.76. While simple, it struggled with complex interactions between features.
- Support Vector Machine (SVM): SVM achieved an F1-score of 0.78 but took longer to train due to its computational complexity.
- K-Nearest Neighbors (KNN): The KNN model obtained an F1-score of 0.72, which is lower compared with other models.
- Naive Bayes: The Naive Bayes model obtained an F1-score of 0.73.
- Decision Tree Classifier: The Decision Tree Classifier achieved an F1-score of 0.65, the lowest among all models, due to its tendency to overfit training data.
- Random Forest Classifier: Random Forest improved generalization, achieving an F1-score of 0.75.

- Gradient Boosting Classifier (GBM): GBM, the best performing model, showed strong performance with an F1-score of 0.77 but required hyperparameter tuning for optimal results.
- XGBoost Classifier: One of the best-performing models, achieving an F1-score of 0.77. Hyperparameter tuning might improve its performance later.

Overall, XGBoost and Gradient Boosting models are two models that used ensemble methods to improve their performance and finally gave the highest F1-scores among all the models that tested. Therefore, we are going to hypertune these two models to see if they could give results that are more accurate.

Model Training

To optimize XGBoost and Gradient Boosting models, RandomizedSearchCV, GridSearchCV and recursive feature elimination were applied.

The random search process involved **5-fold cross-validation** with **50 different hyperparameter candidates**, totaling **250 fits**. The key parameters tuned in randomized search included:

- Learning rate (η): Tested values in the range [0.01, 0.3]. The best-selected value was 0.038.
- Maximum depth: Ranged from 3 to 15. The optimal depth was found to be 6.
- Number of estimators: Evaluated between 50 and 500, with the best result at 253 trees.
- Subsampling rate: Values between 0.5 and 1.0 were tested, selecting 0.90 as the optimal fraction of samples used for training.
- Regularization parameters (alpha & lambda): Used to control model complexity. Alpha (L1 regularization) = 0.17, and Lambda (L2 regularization) = 7.63 provided the best performance.
- Gamma: A hyperparameter controlling split penalties, with the best value found at 0.46.

The best cross-validation F1 Score obtained was **0.7733**.

After identifying a promising region of hyperparameters from RandomizedSearchCV, a GridSearchCV was conducted to fine-tune the most influential parameters. The refined grid search tested:

- Learning rate: [0.03, 0.04]
- Max depth: [5, 6]
- Number of estimators: [200, 250]
- Subsample: [0.85, 0.90]
- Regularization parameters (alpha, lambda): gamma [0.4, 0.5]

However, the GridSearchCV result didn’t beat the randomized search result above, achieving an F1-score of **0.7731** on the test set.

We also tried to apply recursive feature elimination to the best model gained: XGBoost model with random searched parameters. The process chose 9 features, ['HighBP', 'HighChol', 'CholCheck', 'HeartDiseaseorAttack', 'HvyAlcoholConsump', 'GenHlth', 'DiffWalk', 'Age', 'BMI_Category']. However, the F1-score is lower than the original version, so we decided to use all the features available.

Following our hyperparameter tuning and evaluation of the XGBoost model, we extended our analysis to Gradient Boosting (GB) to determine whether it could provide improved predictive performance for identifying individuals with diabetes. Using a similar optimization approach, we employed RandomizedSearchCV to find an initial set of promising hyperparameters and further refined them with GridSearchCV.

Hyperparameter Tuning for Gradient Boosting

We began by conducting a RandomizedSearchCV over a broad hyperparameter space, similar to our approach with XGBoost. The best-performing hyperparameters were then refined with GridSearchCV, testing a narrower range centered around the optimal values.

The key hyperparameters tuned in Randomized Search for Gradient Boosting included:

- Learning rate (η): Values tested in the range [0.01, 0.3], with the best-selected value being 0.0744.
- Maximum depth: Tested values from 3 to 10, with the optimal depth found to be 3.
- Number of estimators (trees): Evaluated between 100 and 500, with the best result at 350 trees.
- Subsampling rate: Values between 0.5 and 1.0 were tested, with 0.70 selected as the optimal fraction of samples used for training.
- Regularization parameters: Minimum samples required for splitting a node and minimum leaf samples were optimized.

The best hyperparameters from Randomized Search were:

- Learning rate: 0.0744
- Max depth: 3
- Number of estimators: 350
- Subsample: 0.70
- Min samples split: 2
- Min samples leaf: 3

Using these values, GridSearchCV was employed for fine-tuning, reducing the search space to:

- Learning rate: [0.2, 0.3]
- Max depth: [3, 4]
- Number of estimators: [200, 250]
- Subsample: [0.85, 0.90]
- Min samples split: [3, 4]

- Min samples leaf: [1, 2]

After evaluating 192 candidate models, the best hyperparameters identified were:

- Learning rate: 0.2
- Max depth: 3
- Min samples leaf: 2
- Min samples split: 3
- Number of estimators: 200
- Subsample: 0.90

The best cross-validation accuracy obtained for Gradient Boosting was **0.7734** from grid search.

Based on the outputs from the Gradient Boosting and XGBoost models, we can compare their performance using precision, recall, and F1-score to determine which model is better suited for identifying people with diabetes.

Comparison of Models

Model	Precision	Recall	F1-Score
Gradient Boosting	0.7449	0.8041	0.7734
XGBoost	**0.7456**	**0.8031**	**0.7733**

Analysis

- Precision: Precision measures how many of the positive predictions were actually correct. XGBoost now has a slightly higher precision (0.7456) compared to Gradient Boosting (0.7449), meaning it produces slightly fewer false positives.
- Recall: Recall measures how many actual positive cases were correctly identified. Gradient Boosting has a slightly higher recall score with 0.8041.
- F1-Score: The F1-score balances precision and recall, with higher values indicating better overall classification performance. Gradient Boosting has a slightly higher F1-score (0.7734) than XGBoost (0.7733), suggesting a marginally better balance between precision and recall.

Deciding Model

- Since F1-score is the most critical metric for imbalanced datasets (such as identifying diabetes cases, where false negatives and false positives both have significant consequences), we prioritize the model with the highest F1-score.
- Gradient Boosting slightly outperforms XGBoost across all three metrics, including recall, and F1-score.
- While the difference is small, Gradient Boost is the preferred model because it has the best overall balance of precision and recall, making it the more effective choice for

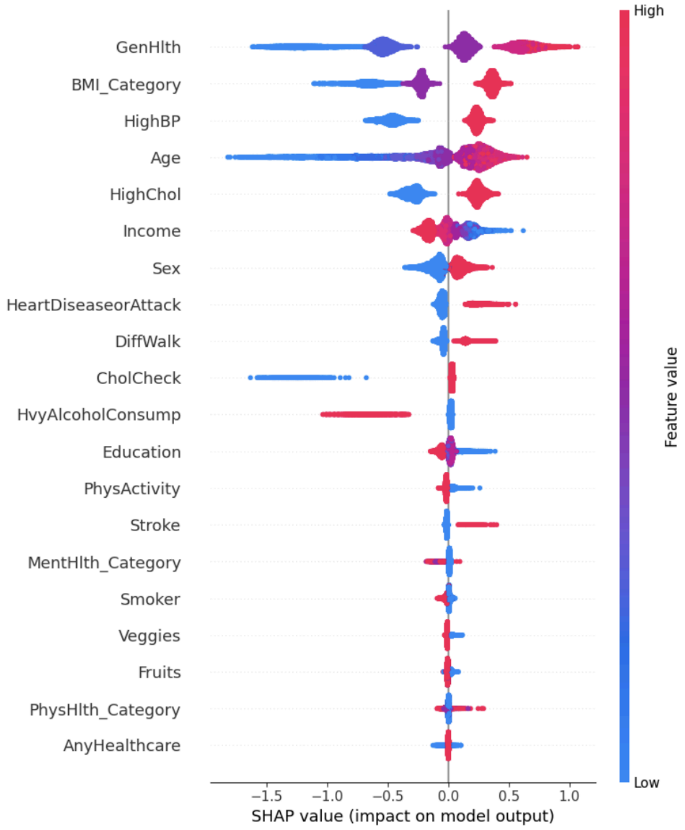
diabetes prediction. It also has a higher recall score, meaning that it has better ability to detect people who actually have diabetes.

Thus, we choose Gradient Boosting as the final model for deployment.

Model Interpretation

The real-world impact of this model in the healthcare setting should focus on attaining high recall, where the goal is to reduce missed diagnoses of diabetes. When left undiagnosed, unmanaged, or untreated, the disease poses a huge risk to an individual's life. Our final model achieves a recall of almost 80%. The precision it achieves is approximately 75%, meaning it has a moderate rate of flagging individuals as diabetic when they are not. Our final model achieved a reasonable trade-off between the two with an F-1 score of 77%, which is moderately good, but allows room for improvement.

In order to maintain model explainability, we also utilized SHAP to interpret the variables. GenHlt, BMI, blood pressure, age, and cholesterol proved to be the five features with the greatest impact on determining if an individual has diabetes. Poorer general health, higher BMI, blood pressure, age and cholesterol mean a higher risk towards diabetes.



Future Work

Further research could focus on identifying and incorporating additional features that may enhance the model's predictive power. Exploring new data sources related to medical history, lifestyle, and demographics could provide a more comprehensive understanding of the factors influencing diabetes risk. Additionally, since the current dataset is from 2015, it would be beneficial to analyze more recent data to determine if patterns and trends in diabetes diagnoses have changed over time, potentially leading to more accurate and up-to-date models.

Additionally, rather than outputting a binary outcome of diabetes or no diabetes, we can also explore regression methods that would instead output a probability of having diabetes. With increased data points, we could also explore a multi-class classification where the outputs would be no diabetes, prediabetes, or diabetes, to see how the features influence each class. Overall, this machine learning model holds the potential for broader applications, including the simulation of clinical trials and the study of various other medical conditions.

References

Data Source
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Column Descriptions
https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf