Final Project

STA 4724

By: Elishel Sason, Jacob Slack, Matthias Rathbun

Table of Contents

Purpose

The aim of our project is to better understand the risk factors of diabetes through creating an accurate model at predicting whether a patient has diabetes based on their given health indicators. Diabetes is a prominent health issue as about 1.5 million deaths are directly attributed to diabetes each year worldwide, and it is important for people to understand the risk factors associated with this issue in order to work towards diabetes prevention. A prediction model can hold potential applications in healthcare as a tool for early screening and identification for diabetes.

Background

The dataset we used was from The Behavioral Risk Factor Surveillance System by the CDC which is an annual health study conducted through telephone surveys. There were 253,680 observations in this study from 2015 ; each observation counts as a single patient's response. For the target variables, the dataset would classify whether each participant is either: not diabetic with 0, prediabetic with 1, or diabetic with 2. The dataset consists of 21 features which are health indicators based on demographics, lab data, and survey results. The feature columns of each participant's health include factors such as their BMI, blood pressure, general health, and others, which are put in either binary or integer values. For example, the high blood pressure feature is entered as a binary value with 0 being no high blood pressure and 1 being high blood pressure. Features such as physical health were measured as integers, as physical health would measure the amount of days during the past 30 days that the patient's health was not good and would be entered on a scale of 1-30 days.

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
253675	0.0	1.0	1.0	1.0	45.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0
253676	2.0	1.0	1.0	1.0	18.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0
253677	0.0	0.0	0.0	1.0	28.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0
253678	0.0	1.0	0.0	1.0	23.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0
253679	2.0	1.0	1.0	1.0	25.0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0

253680 rows × 22 columns

The dataset

Exploratory Data Analysis

Once we had an understanding of the basic contents of the dataset, the next step was to analyze the actual data that it provided. The first step in doing so was to take a look at the general distributions of our features and target variables. We created histogram distribution plots for each feature (see Appendix A). When analyzing these plots it was immediately apparent we had an imbalanced dataset with respect to our target variable. We then decided to take a look at feature means across the classes of our target variable, where we could see the averages for each feature when the patient's were considered diabetic (2), prediabetic (1) , or not diabetic (0). These averages can be seen below.

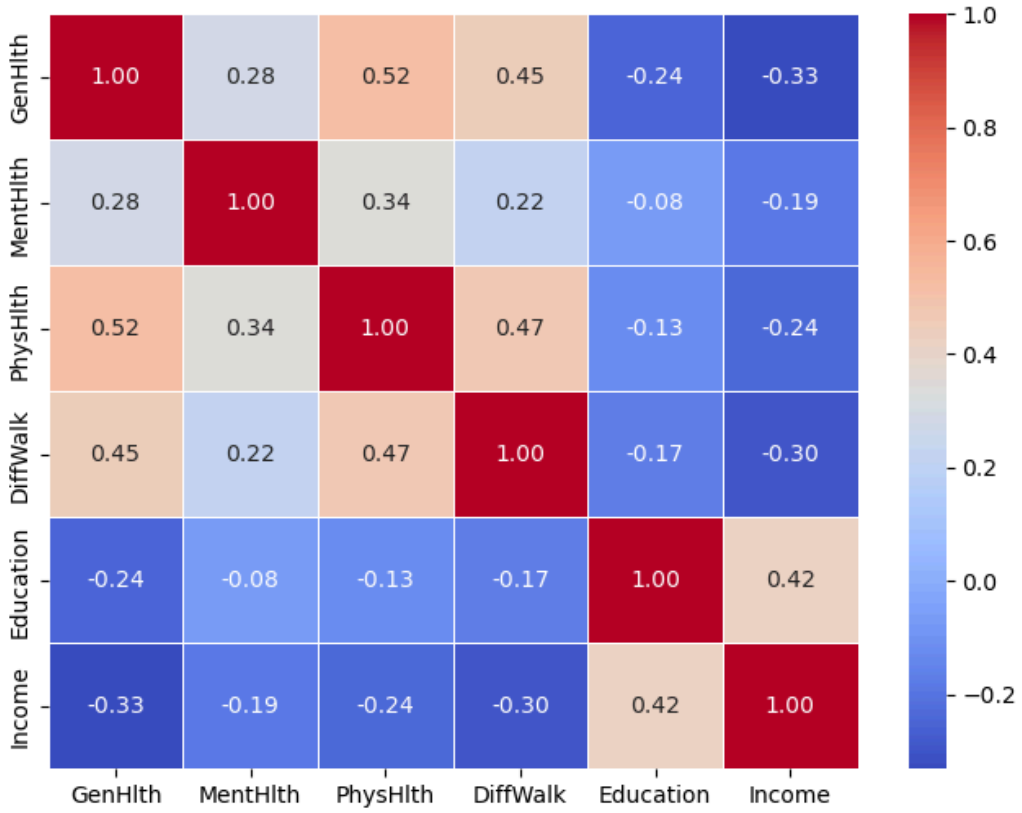
Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump
0.000000	0.40	0.40	0.95	28.03	0.46	0.04	0.08	0.75	0.62	0.80	0.07
1.000000	0.63	0.62	0.99	30.73	0.49	0.06	0.14	0.68	0.60	0.77	0.04
2.000000	0.75	0.67	0.99	31.96	0.52	0.09	0.22	0.63	0.58	0.75	0.02

Diabetes_012	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0.000000	0.94	0.09	2.46	3.30	4.02	0.15	0.43	7.82	5.03	6.03
1.000000	0.95	0.13	2.98	4.53	6.35	0.28	0.44	9.08	4.78	5.35
2.000000	0.96	0.11	3.30	4.49	8.01	0.37	0.48	9.38	4.74	5.20

Taking a look at these averages, there are a few that are quite distinct in their differences among classes. HighBP and HighChol, which represents whether the patient has high blood pressure or has high cholesterol, have significantly higher averages in he diabetic and prediabetic class compared to the non-diabetic class. This is also true for GenHlth, MentHlth, and PhysHlth. It is important to note that these features are on a distinct scale where 1.0 represents "Good" or

"No issue" and higher ratings represent "Bad" or "Lots of issues." So we can see there are higher ratings present in the diabetic and prediabetic class compared to the non-diabetic class. These averages are not necessarily surprising, but they do give insight into which features are likely going to be most indicative of a diabetic patient.

Next we decided to take a look at multicollinearity between features in the dataset. To do this, we created a full correlation matrix (see Appendix B). However, with 21 features, the matrix is quite large, so below we can see a subsection of the matrix with the most relevant results.



Looking at the correlation matrix, it consists of features that would generally be expected to have some correlation. For example, PhysHlth (Physical Health Rating) and DiffWalk (Difficulty Walking Boolean) would be expected to have some correlation. If a patient has difficulty walking, then they are likely going to rate their physical health as being worse. This applies to

several other combinations of features; however, there were no feature correlations that were actually significantly strong correlations. The highest correlation of 0.52 was between PhysHlth and GenHlth, which again is not surprising.

Ultimately the dataset was determined to be unbalanced with a few weakly correlated features. The distinct differences in several feature averages across the classes looked promising for achieving an accurate model in the next phase of the project. Using this better understanding of the dataset, we decided upon the best next steps in order to give our models the best dataset possible with the goal of achieving an actually accurate model. Once we had investigated the contents of the dataset, it underwent a preprocessing phase in order to refine and organize the raw data. During this phase we combined the prediabetes (1) and diabetes (2) classes into one class, deleted duplicate rows, and non-categorical columns were normalized, resulting in 69,057 rows left to use.

Approaches and Results

Once we had finished the data exploratory phase, the next step was to begin creating the various predictive models. The models we selected to use were Logistic Regression, Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Tree Method with Decision Tree, Bagging, XGBoost, and Random Forest, and finally a Neural Network. The chosen models encompass a diverse set of machine learning algorithms, each with their own unique strengths and capabilities. Logistic Regression is a fundamental classification model with high interpretability commonly used for binary classification tasks. Linear Discriminant Analysis (LDA) is chosen for its ability to handle multiclass classification problems and its emphasis on class separability. K-Nearest Neighbors (KNN) is a non-parametric method effective in capturing relationships within the data due to its flexibility. The Tree Method with Decision Tree, Bagging, XGBoost, and Random Forest are ensemble methods known for their ability to enhance predictive performance by combining multiple weaker learners. Finally, a Neural Network, representative of deep learning, is included for its ability to model complex relationships in the data.

In order to try and use the models possible, we implemented hyperparameter optimization. To do this, we used gridsearch-5CV to find the best parameters for each model. You can see the grids below, which contain the initial possible parameter values. From there, optimal hyperparameters were selected by the gridsearch. Since we implemented numerous models and had hyperparameter tuning results, these classification reports and full model evaluations are included at the end of the report in the appendices. Ultimately the results were very similar among all the models with accuracies ranging from 0.73 to 0.75.

a.

Parameter	Search Space
C	0.001, 0.01, 0.1, **1**, 10, 100
Penalty	**L1**, L2
Solver	Saga, **Liblinear**

b.

Parameter	Search Space
K	7, 9, 11, 13, **15**
Weights	**Uniform**, Distance
P	**1**, 2

c.

Parameter	Search Space
Criterion	**Gini**, Entropy
Max Depth	None, **10**, 20, 30
Min Samples Split	2, 5, **10**
Min Samples Leaf	**1**, 2, 4

d.

Parameter	Search Space
Estimators	50, **100**, 200
Criterion	Gini, **Entropy**
Max Depth	None, **10**, 20, 30
Min Samples Split	2, 5, **10**
Min Samples Leaf	1, **2**, 4

e.

Parameter	Search Space	Parameter	Search Space
Estimators	100, 200, **300**	Min Child Weight	1, **2**, 3
Learning Rate	**0.01**, 0.1, 0.2	Gamma	**0**, 0.1, 0.2
Max Depth	3, **6**, 9		

Figure 1: Hyperparameter search spaces for machine learning models in diabetes classification. (a) Logistic Regression Regularization strength (C), penalty types, and solver. (b) K-Nearest Neighbors: Number of neighbors, weight function, and power parameter. (c) Decision Tree: Split criterion, maximum depth, minimum samples required to split a node, and at a leaf node. (d) Random Forest: Number of trees (estimators), split criterion, tree depth, and minimum samples for split and leaf. (e) XGBoost: Number of boosting stages, learning rate, tree depth, Minimum Child Weight, and regularization term gamma. Values selected for the final model are indicated in bold.

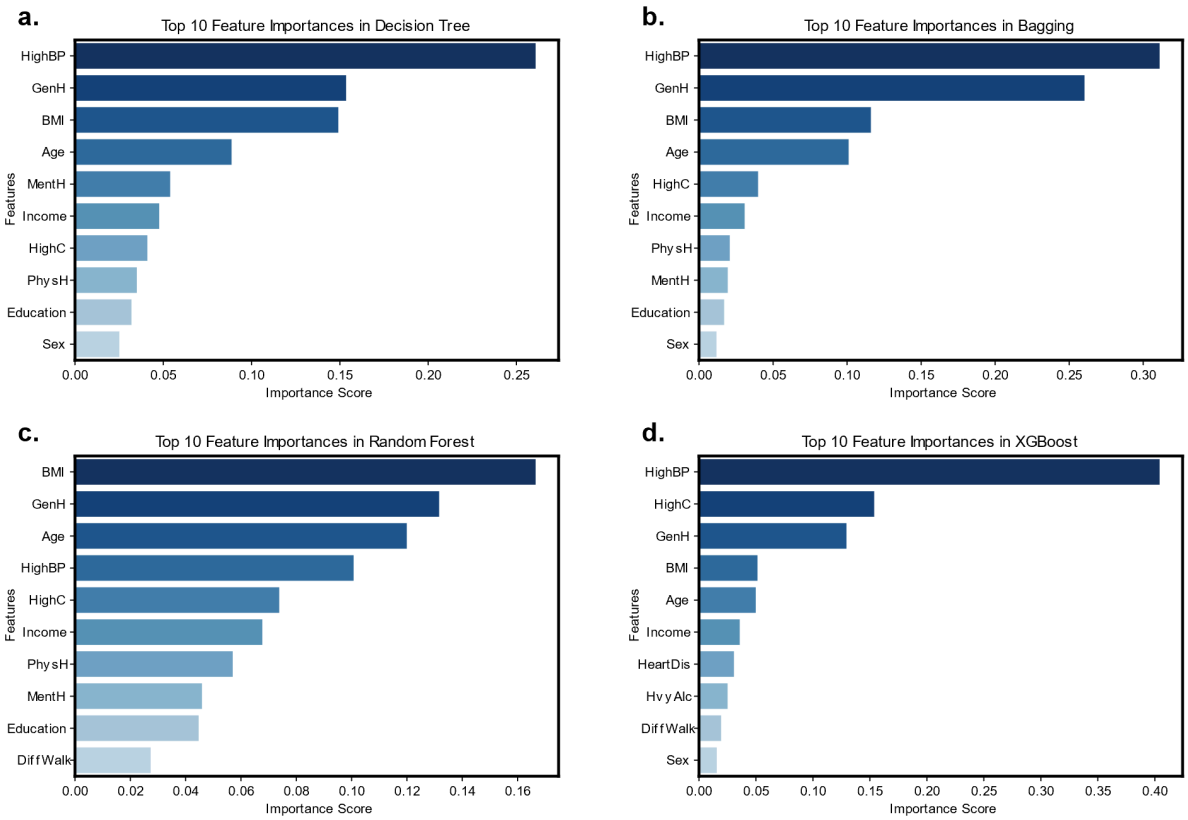


Figure 2: Feature importance scores for tree-based models in diabetes classification. (a) Decision Tree shows HighBP as the most influential feature, followed by General Health, BMI and age. (b) Bagging identifies the same features as the decision tree as prominent. (c) In Random Forest, BMI leads, with General Health, Age, and HighBP also critical. Importance scores for Random Forest are much lower compared to the other models. (d) XGBoost highlights the significance of High Cholesterol. These plots elucidate the varying impact of genetic and clinical features on the models' decision-making processes.

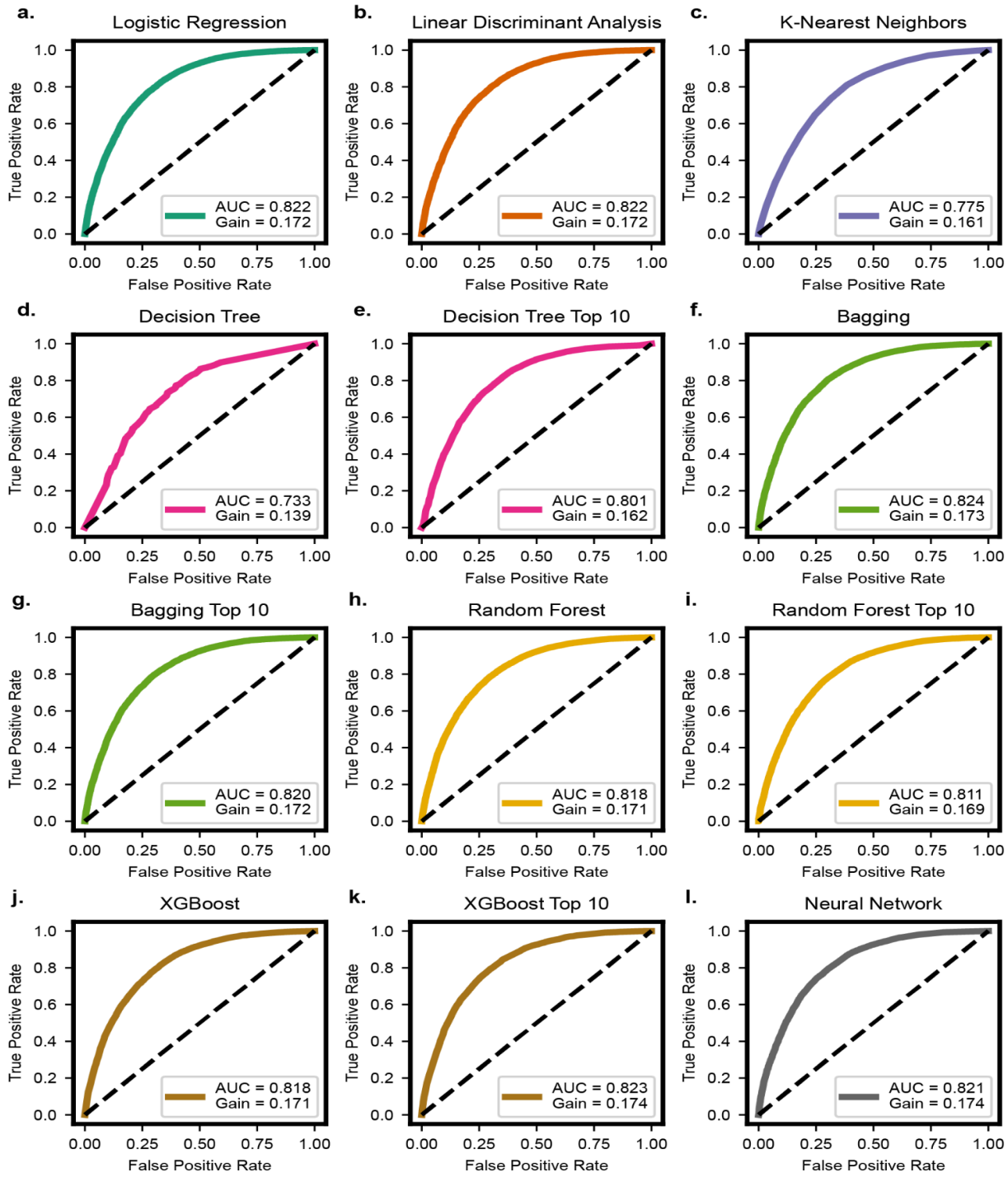


Figure 3: ROC curves for machine learning models in diabetes classification. Panels (a) through (l) display the true positive rates versus false positive rates for Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Decision Tree, Decision Tree with top 10 features, Bagging, Bagging with top 10 features, Random Forest, Random Forest with top 10 features, XGBoost, XGBoost with top 10 features, and Neural Network, respectively. The AUC and Gain values are annotated for each model. Notably, most models perform similarly on each task, indicating the dataset potentially sufferers from high noise as even the more complex models fail to learn a different decision boundary compared to the simpler models.

Conclusion

After analyzing the results of this investigation, we ended up with the highest accuracies being 0.75 from the methods: logistic regression, LDA, decision tree w/ bagging, and random forest and the highest AUC resulted in 0.824 from bagging. The primary challenge we encountered during the investigation was the noise in the dataset. To fix this, we attempted to use only the top features and dropped the rest, however, the results did not differ much.

Based on the results from extensive modeling methods and tuning, we ultimately came to the conclusion that this dataset alone was not sufficient in providing significant predictors for diabetes, which traditionally is not laborious to predict. Considering there was minimal if any differences in the results between our simple models and our most complex, using this dataset by itself is simply not sufficient. Due to the nature of survey results and minimal numerical lab data, this dataset contained much noise and nonideal predictors for achieving very high accuracies. For future investigation into a diabetes prediction model, getting more specific continuous data rather than broad generalizations of survey results would likely lead to more accurate results. Additionally, other models for predictive symptoms of diabetes could also be another route to go down, such as predicting blood glucose levels after certain consumption scenarios given a patient's other relevant health indicators.

References

"CDC Diabetes Health Indicators." UCI Machine Learning Repository, 2015, archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators..

Teboul, Alex. "Diabetes Health Indicators Dataset." Kaggle, 8 Nov. 2021, www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

Team Member Contributions

Elishel Sason

- Initial modeling with Logistic Regression, KNN, and Tree Methods for the 50/50 split binary dataset and the 3 class dataset
- Data processing
- Worked on presentation and final report

Jacob Slack

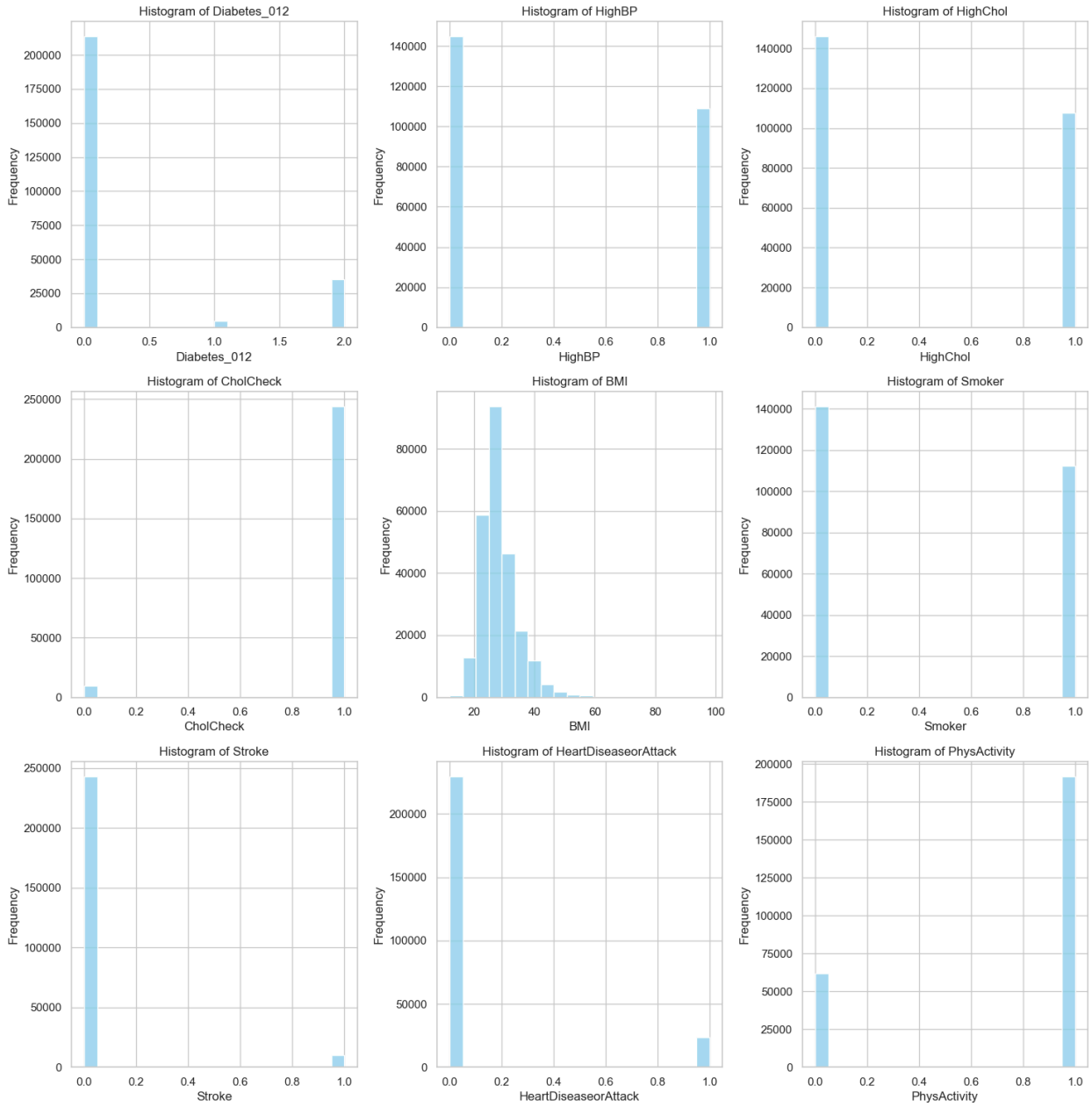
- Dataset Exploratory Analysis Visualizations and Explanations
- Initial investigatory modeling with Logistic Regression, KNN, and Tree Method with full binary target variable dataset
- Helped with explanation of model approaches and results
- Helped with Presentation/Report material for relevant sections and overall formatting

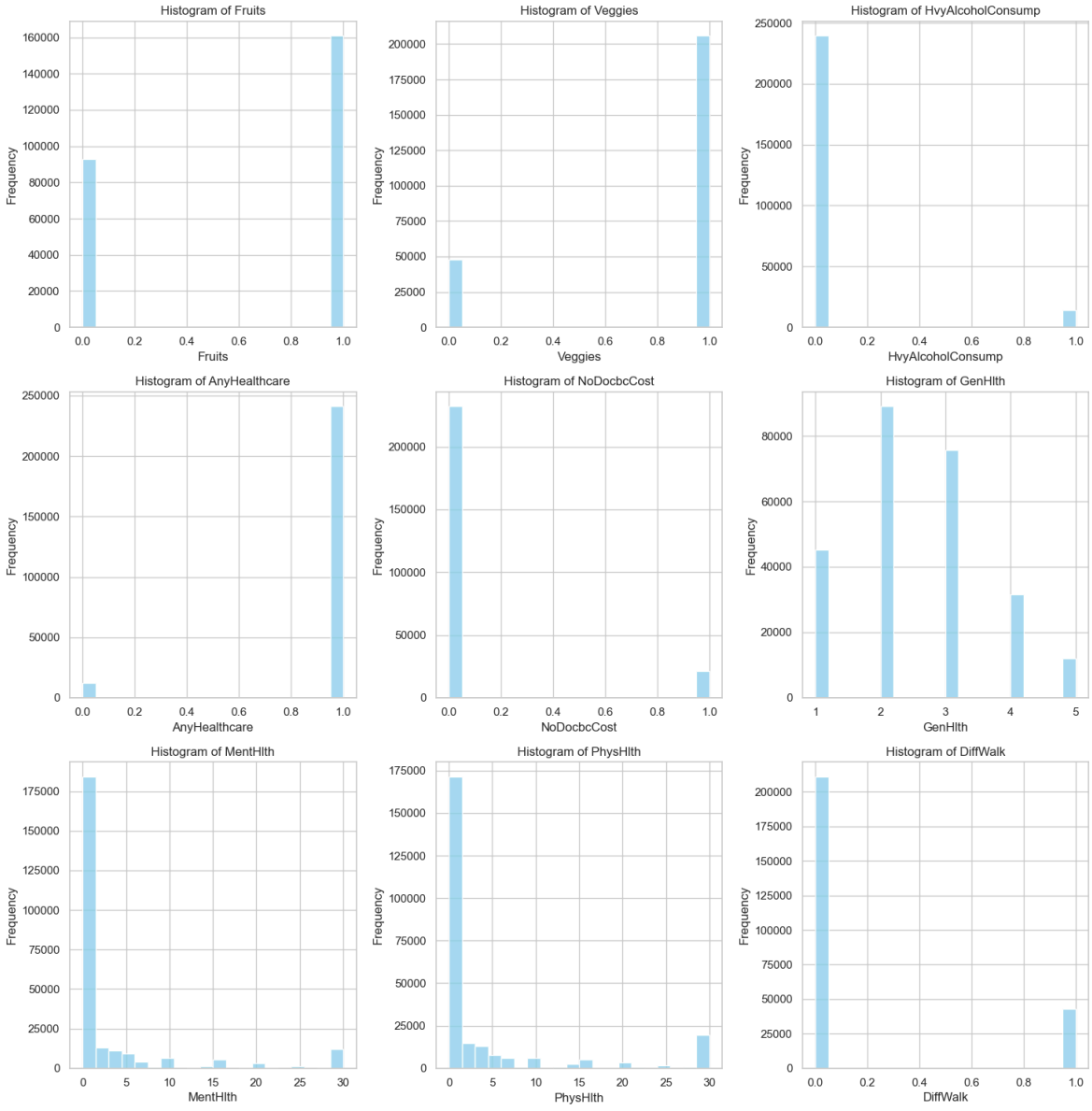
Matthias Rathbun

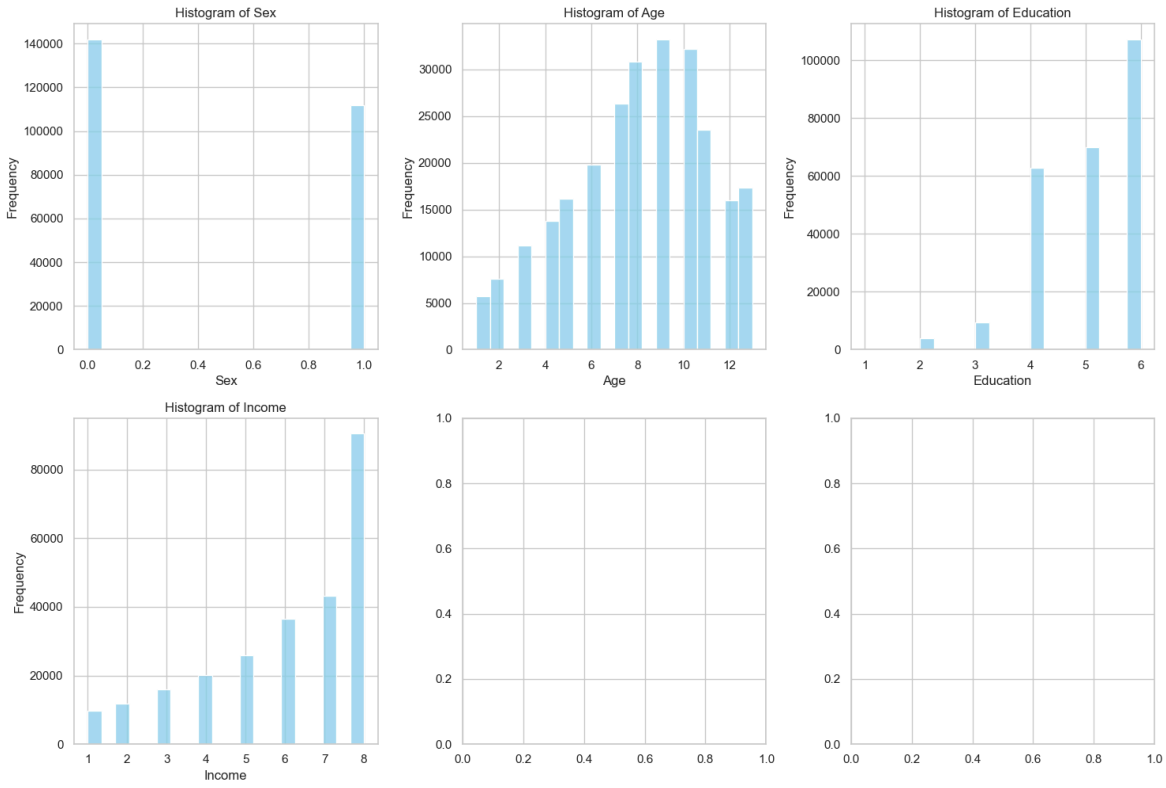
- Hyperparameter Optimization
- Acquired Feature Importances from tree based models
- Model Evaluation using classification report and AUC-ROC
- Neural Network Design
- Figure Making. Edited Figures produced in python with Adobe Illustrator. Wrote Figure Captions

Appendix A

As seen below, these are the resulting feature distribution plots. These give insight into the exact distributions of features that the dataset originally contained.




Histogram of Fruits
Frequency
Fruits
Histogram of Veggies
Frequency
Veggies
Histogram of HvyAlcoholConsump
Frequency
HvyAlcoholConsump
Histogram of AnyHealthcare
Frequency
AnyHealthcare
Histogram of NoDocbcCost
Frequency
NoDocbcCost
Histogram of GenHlth
Frequency
GenHlth
Histogram of MentHlth
Frequency
MentHlth
Histogram of PhysHlth
Frequency
PhysHlth
Histogram of DiffWalk
Frequency
DiffWalk


Histogram of Sex
Frequency
Sex
Histogram of Age
Frequency
Age
Histogram of Education
Frequency
Education
Histogram of Income
Frequency
Income

Appendix B

Below is the full correlation matrix created during the data investigatory phase.