Insight into Health Indicators for Diabetes

Group Members: Genevieve Ferguson, Nitin Pagarani, & Cristian Biondi
Mentor: Dr. Giri Narasimhan
Date: Dec 2nd, 2023

Executive Summary:

- **Objective:**
  - Utilize data analytics and machine learning to predict diabetes.
- **Datasets:**
  - Employ the Diabetes Health Indicators dataset from the UCI Machine Learning repository.
  - Use CDC's City Census Data & Health Indicators 18+ for geographical analyses.
- **Methodology:**
  - Implement classification models, exploratory data analysis and geographical analysis.
- **Results:**
  - Exploratory data analysis revealed significant correlations among health indicators.
  - Geographical analysis revealed diabetes prevalence across the different states and cities in the United States.
  - Classification models demonstrated good accuracy of diabetes prediction.
  - SHAP values helped us gain insight into feature importance in the ANN model.
- **Key Features Identified:**
  - Physical health, BMI, Age, mental health, and high blood pressure all emerged as crucial predictors of diabetes.
- **Recommendations:**
  - Prioritize identified factors in diabetes research and prevention.
  - Advocate for lifestyle modifications, regular screenings, and preventive program participation.
- **Conclusion:**
  - The project aims to predict diabetes, providing insight into key factors and geographical variations across the USA.
- **Future Works:**
  - Explore global factors for more comprehensive understanding into diabetes.

Abstract

Diabetes is a prevalent chronic condition in the United States with severe health implications. This research uses data analytics and machine learning while utilizing the *UCI Machine Learning Repository Diabetes Health Indicators Datasheet* and the *CDC (Centers for Disease Control) City Census Data*. The overall aim of this study was to leverage these datasets to identify the most influential health indicators for diabetes, analyze state prevalence, and provide recommendations for risk reduction. Various data analytics techniques such as handling imbalanced data, feature selection, and classification models (Random Forest, ANN, etc.) were used to accomplish this task. In addition, Explainable AI methods using SHAP values were applied to understand feature importance. This paper explores the relationships between diabetes and several other factors that are often screened for in a diabetes analysis. An interactive US map is also visualized to find nationwide diabetes concentrations. Results showed highly accurate classification models with Random Forest and ANN performing with the most accuracy. The research contributes valuable insight into diabetes prediction and risk factors, aiding in preventive healthcare strategies. This study was an equal contribution between all the sections of our project.

Introduction

Diabetes is a chronic condition that occurs when the pancreas cannot produce enough insulin or when the body cannot regulate it. Insulin is the hormone produced to regulate blood glucose in the body ("World Health Organization," n.d.). Hyperglycemia or raised blood sugar is a common effect of uncontrolled diabetes and can lead to other health conditions such as heart disease, vision loss, and kidney disease. In 2019, 1.5 million deaths were reported because of diabetes and 48% of these individuals were less than 70 years old ("World Health Organization", n.d.). In the United States, approximately 38 million people (about twice the population of New York) have diabetes and 20% of these individuals are unaware that they have the condition ("Centers for Disease Control", 2023)

The two most common forms of diabetes are type 1 and type 2, though there are several other rare forms that occur far less often. Type 1 diabetes results from the autoimmune destruction of the pancreas's beta cells which produce insulin. Individuals diagnosed with type 1 diabetes require insulin for survival. Insulin is given as a daily shot or continuously with an insulin pump. Type 2 diabetes is caused by the body gradually producing less insulin and becoming insulin resistant. Other types of diabetes include gestational diabetes and latent autoimmune diabetes in adults. Of all diagnosed cases, it is estimated that 6% are type 1 diabetes, 91% type 2 diabetes, and the remaining 3% are other types of diabetes ("Centers for Disease Control-Prevalence", 2023). In diagnosing individuals with diabetes, patients experiencing symptoms such as urinating frequently at night, blurry vision, numb or tingling hands and feet, and dry skin are recommended to get their blood sugar levels tested ("Centers for Disease Control," 2023). Known risk factors for developing diabetes include weight, age, lack of physical activity, non-alcoholic fatty liver, ethnicity, and family history ("Centers for Disease Control," 2023).

Considering the severity of the disease, its risk factors, and its widespread impact on individuals domestic and globally, our goal is to leverage accessible data and its relevant features to predict diabetes. For our research project, we will be utilizing data analytical tools to explore diabetes data, identify correlated features, and develop classification models to predict diabetes. Next, we will identify the most notable features or health indicators of the dataset. Then, we will analyze diabetes data by state and city to understand the prevalence of diabetes across different cities of the United States. Lastly, we will be making recommendations that could potentially reduce the risk of developing diabetes. Our inspiration and motivation for this project include the fact that diabetes causes many health complications and its prevalence across the world is significant. We believe that by studying and analyzing features of our dataset we can make data driven recommendations to reduce the risk of diabetes among individuals. Understanding diabetes and its impact is crucial for developing a long and healthy life. Furthermore, individuals can alter their lifestyles to better manage their health and prevent the risk of diabetes.

Methodology/Approach

Diabetes Health Indicators Data

Preprocessing & Cleaning

In analyzing our dataset, we first cleaned and preprocessed the data by removing duplicated data and checking for unique values in our dataset. Additionally, we added categorical labels to the data to describe the features of the dataset.

Exploratory Data Analysis

To explore the data and understand the features of our dataset, we developed visualizations to understand trends and gain further insight. Utilizing this methodology, we developed point plots, count plots, heat maps, box plots, violin plots, bar plots, and histograms. While exploring our data, we analyzed relationships between several features of our data such as smoking, alcohol consumption, heart disease, stroke, etc. Furthermore, we were able to understand age distributions, income distributions, and education distributions.

Addressing Imbalanced Data

To combat the imbalanced distribution within our dataset, we decided to use the Near Miss algorithm, an under-sampling technique. This method reduces the abundance of a majority class by retaining instances that align with the minority class. This will help the dataset by preventing the model from being skewed toward the majority class during the training phase. The outcome of this process gives us a balanced dataset to improve our model's observations across all classes and enhance the performance where the minority class is significant.

Feature Selection

To select the most prominent features to build our classification models, we analyzed correlations in data to analyze the most highly correlated features. We developed a heat map to display these results and a bar chart showing the features in descending order. Secondly, we utilized Random Forest to find the most notable features and listed their scores. Lastly, we used Select K Best to find the most important feature scores and listed the features in ascending order by their respective scores. Furthermore, using Select K Best, we selected the top 16 features of the dataset based on their chi-squared scores to use in our classification models. Determining the optimal value of k for selecting the best features is essential to our model's ability to discern the most informative feature to enhance its predictability.

Building Classification Models

We employed a classification model to leverage the power of machine learning in predicting diabetes to find patterns within the health indicators. To classify an individual as having diabetes or no diabetes based on the most salient features, we built 10 classification models. These models included KNN, Random Forest, Decision Tree, SVM, XGBoost, MLP Classifier, Ridge Classifier, Logistic Regression, Passive Aggressive Classifier, and an ANN.

Explainable AI (Artificial Intelligence) using SHAP Values

To gain deeper insight into our features and their importance we utilized SHAP. SHAP refers to Shapley Additive Explanations which calculates a feature's impact on the target variable's value. By utilizing this methodology, we gained insight into which features were most important in predicting diabetes. SHAP values not only provided us with a quantitative measure of feature impact but also a comprehensive understanding of the direction and magnitude of each feature contribution.

City Census Data & Health Indicators 18+

Preprocessing & Cleaning

To preprocess and clean this dataset we removed null values and conducted missing value imputation to fill in values with their average. We also checked for unique values in the data.

Exploratory Data Analysis

We created count plots and bar plots to view the distribution of Diabetes by state. Next, we wanted to analyze various cities in various states to see which cities had the highest distribution of diabetes. Hence, we analyzed cities in Florida and California to gain further insight.

US Map

Utilizing the geolocations available on our dataset, we were able to build a map of the US to show the distributions of diabetes by city and state. This interactive map was built using the Folium library.

Data & Experiments

Diabetes Health Indicators Dataset

Data

The Diabetes Health Indicators dataset was retrieved directly from the UCI Machine Learning repository. The dataset has a size of 253,680 x 22. The features of the dataset include Diabetes_binary, Age, Income, Smoker, Education, Sex, High BP, Stroke, Fruits, Veggies, Physical Activity, Mental Health, etc. Furthermore, 128,715 females and 100,759 males participated in this survey. Participants were aged 18-75 and had varying ranges of income levels from 10,000-75,000+.

Experiments

To build predictive models based on the health indicators dataset, we first balanced the dataset using the Near Miss algorithm. The parameters set for this algorithm included version 1 and 15 neighbors. Using this methodology, the size of the majority class was reduced from 194,377 samples to 35,097 samples. Hence both class 0 and 1 now had 35,097 samples to train on. Next, we imported a train test split function from scikit learn to split our resampled data into X_train, X_test, y_train, y_test. We established 30% for testing and 70% for training. To train the Decision Tree Classifier, we imported the classifier from sklearn and fit the training data into the model. We then predicted on the model using X_test. Finally, we evaluated the model using the classification report, confusion matrix, and accuracy score.

To build the KNN classifier, we used 10 neighbors and kd_tree as the algorithm to fit the training data. We also predicted the test data and displayed its evaluation metrics. For the random forest classifier, we used 20 estimators, max_depth of 15, and a random state of 100. For the XGboost classifier, we used a learning rate of 0.1 and error as the evaluation metric. Next, we built an SVM, MLP classifier, Ridge Classifier, and Passive Aggressive Classifier by fitting the training data into the classifiers and evaluating the model. To build a logistic regression model we used the solver as liblinear and multiclass as ‘ovr’ with a random state of 0. For each of these models, their accuracy and classification reports were reported.

To build an ANN using Keras, we created a Sequential Model and added 3 Dense Layers. The input and second layer (hidden layer) were built with 16 neurons and the activation functions of ReLU. The output layer was one neuron with an activation function of Sigmoid. The model was compiled using the Adam optimizer, loss function of binary cross-entropy, and accuracy as the metrics. Early stopping was implemented, and validation accuracy was monitored by splitting the data into 20% for validation while training. The model was fitted into the training data with batch sizes of 10 and early stopping as callbacks.

In evaluating the most notable features that have predictive capability, we utilized SHAP to retrieve their SHAP scores. We used 25 samples of our X_train data in our background summary. Additionally, we used an explainer to retrieve the SHAP values and print the results. We then developed the summary plot to show the features by their SHAP scores. Understanding the SHAP scores will allow us to gain insight into feature importance and its relevance to the entire dataset.

Data: City Census Data & Health Indicators 18+

The City Census Data & Health Indicators 18+ dataset was derived from the CDC. This dataset has a size of 29,006x 24. The features that we used for this dataset included State, City name, Geolocation, Data Value, and Population Count. Using this dataset, we conducted city and state analysis and built a map using the Folium Library.

To build the map, we set the location coordinates to the USA to ensure that we are plotting in the correct region. Next, we plotted the clusters and regions of interest based on the geolocation that was included in our dataset. Finally, we added the respective city names for the geolocations plotted.

Results & Discussion

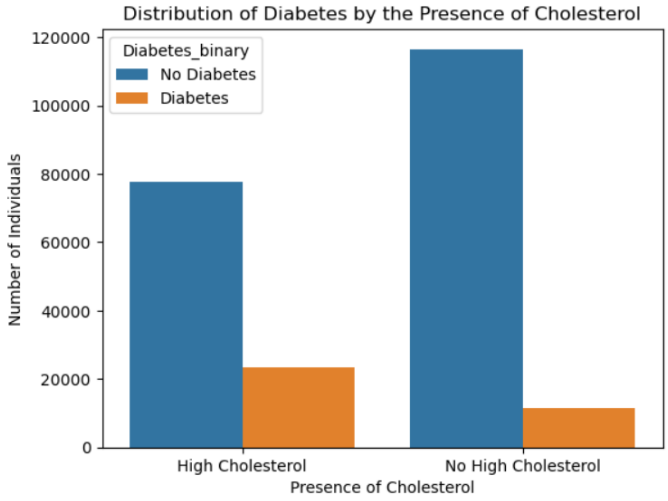

Figure 1: Distribution of Diabetes by the Presence of Cholesterol

To understand our features and its ability to predict our target variable, we first analyzed whether High Cholesterol had an impact on Diabetes. We developed a count plot to find the number of individuals that have High Cholesterol and Diabetes. We noted that the number of people with high cholesterol was higher in individuals with Diabetes. This signifies that this variable has an impact on the target variable and influences our classification model. Also, more individuals had no diabetes or high cholesterol.

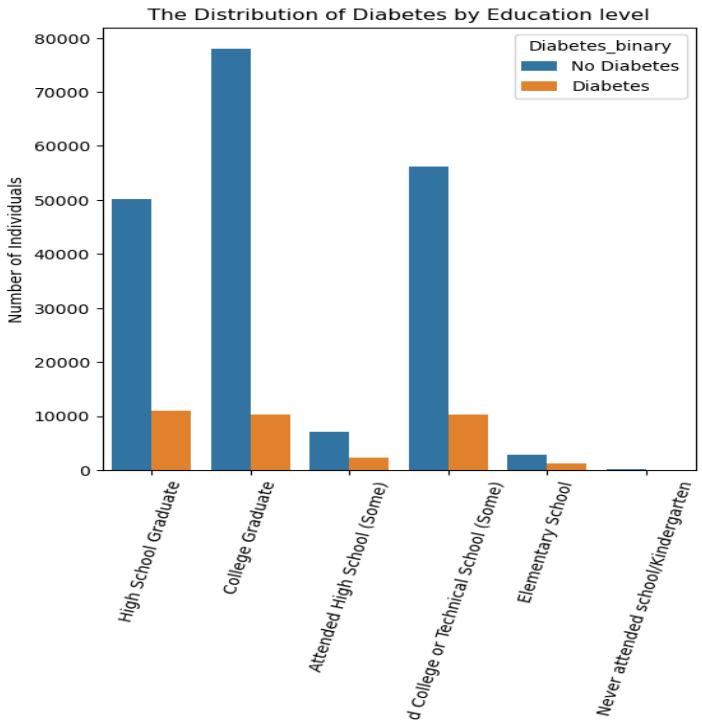

Figure 2: Distribution of Diabetes by Education Level

In analyzing the distribution of diabetes by education level, Figure 2 shows the number of individuals that have diabetes by their highest level of education attained. Individuals who were college graduates were the highest to have no diabetes. Furthermore, individuals who were high school graduates, attended college, or some colleges had the highest count of diabetes. In analyzing this feature, we understand that education also could influence our model in predicting diabetes.

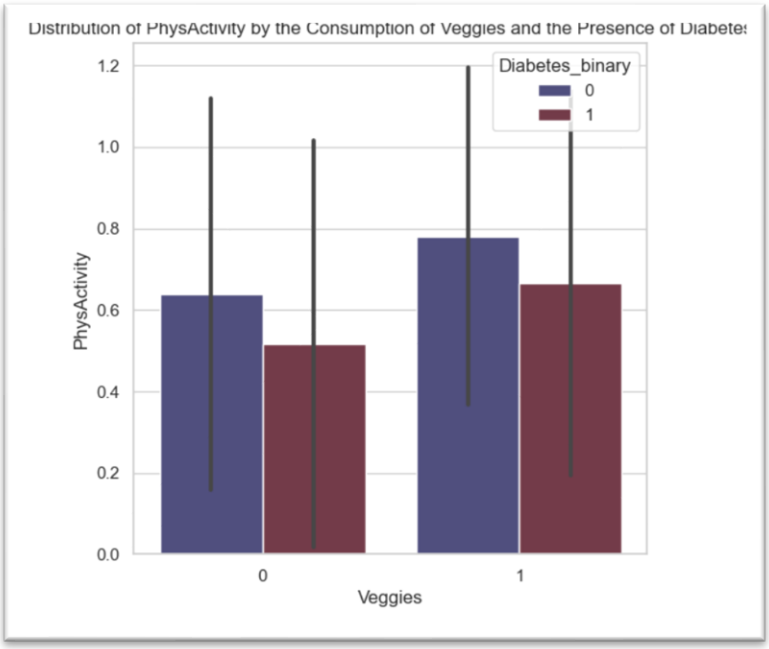

Figure 3: Physical Activity by the Consumption of Veggies & the Presence of Diabetes

Figure 3 depicts that people who consume veggies and are more physically active are less likely to have diabetes. Additionally, it provides insight that people who consume veggies and are physically active also have a high chance of developing diabetes. In further analyzing this data, we understand that variables such as veggies may not have a significant impact in predicting diabetes. Furthermore, the bar plots display error bars which represent the variation in the data.

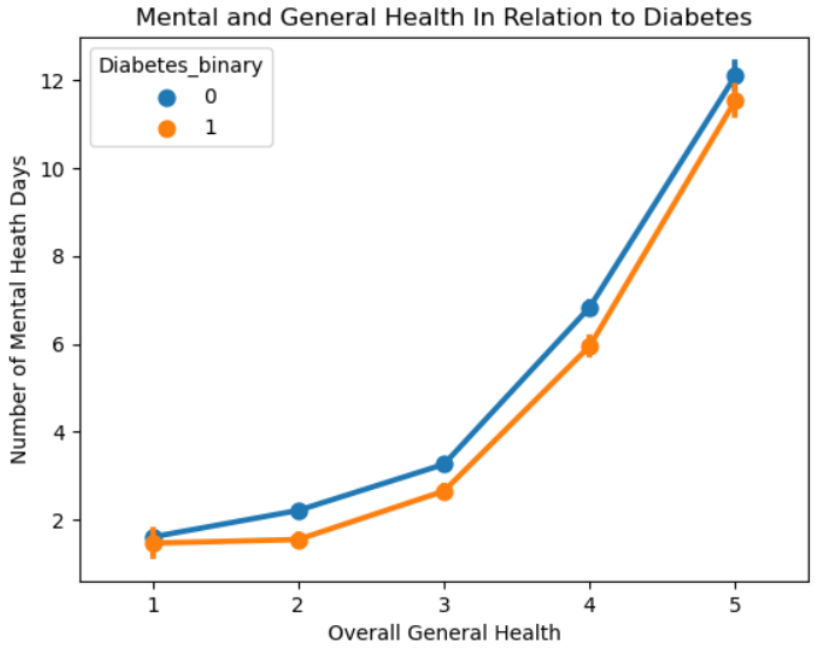

Figure 4: Mental & General Health in Relation to Diabetes

Figure 4 shows that mental health and general health impact Diabetes. As the number of mental health days increases overall general health also increases. 1 represents excellent general health while 5 represents poor overall general health. Hence individuals in excellent health and lower mental health days have a lesser chance of developing Diabetes.

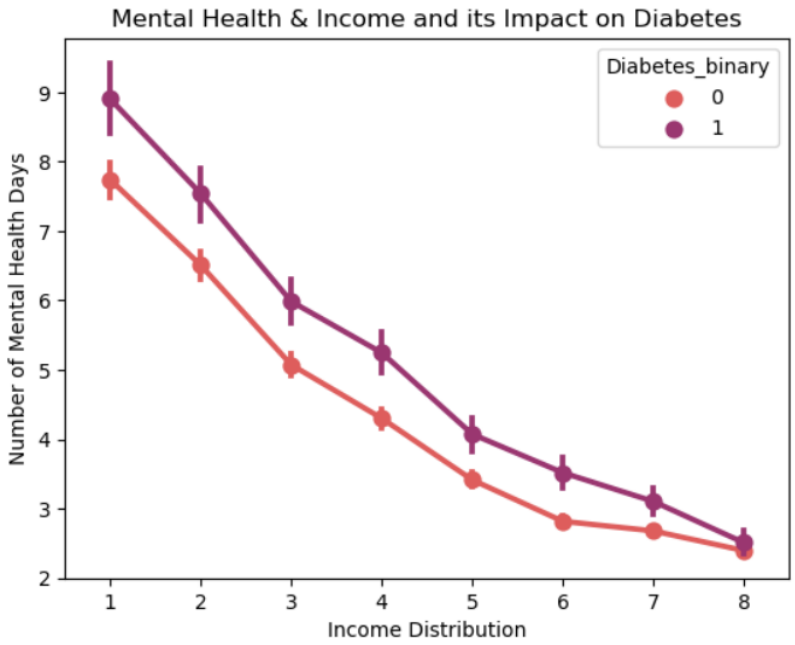

Figure 5: Mental Health & Income and its Impact on Diabetes

In analyzing Figure 5, as income increased the number of mental health days also decreased. It was also noted that individuals are less likely to develop diabetes if they have greater income and lowered mental health days. Individuals with lower income and greater mental health days are more susceptible to developing diabetes.

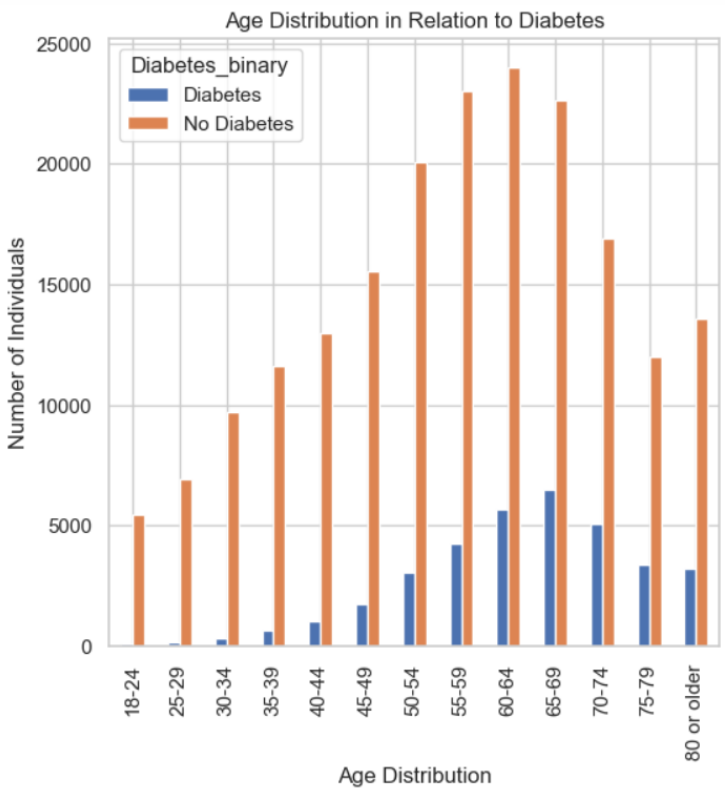

Figure 6: Age Distribution Concerning Diabetes

Figure 6 shows the age distribution of individuals with Diabetes. Individuals aged 65-69 had the highest count of individuals with diabetes. Furthermore, the distribution of diabetes aged 70-74 is lower than age group 65-69. It is noteworthy to analyze that age is only one factor that determines if individuals are more likely to develop diabetes.

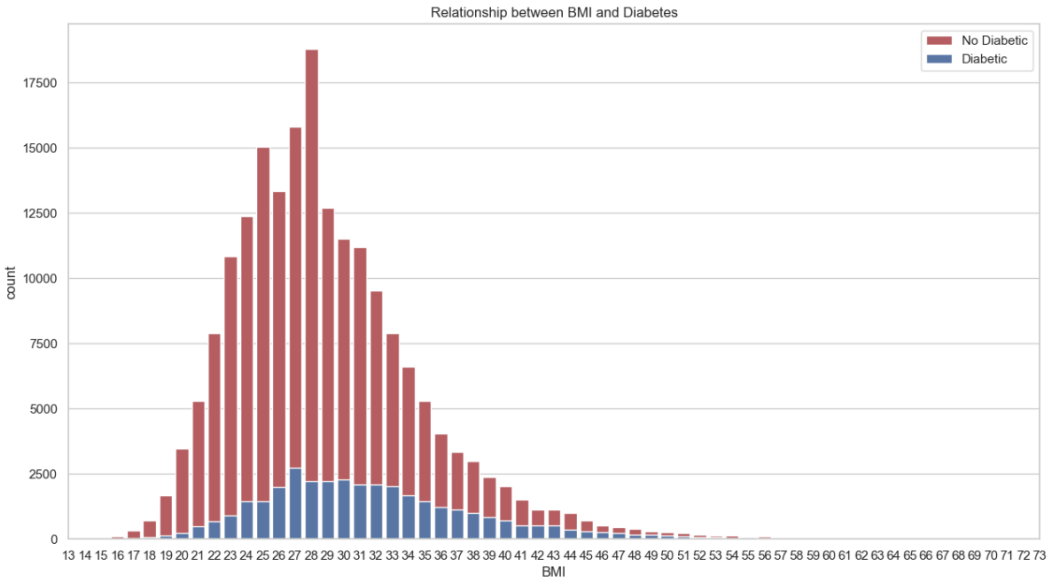

Figure 7: Relationship between BMI & Diabetes

Figure 7 shows people with diabetes have a BMI between 20-45. BMI plays a significant role in diagnosing diabetes as people that are considered overweight and obese have a higher risk of developing diabetes. This is shown in Figure 8 below.

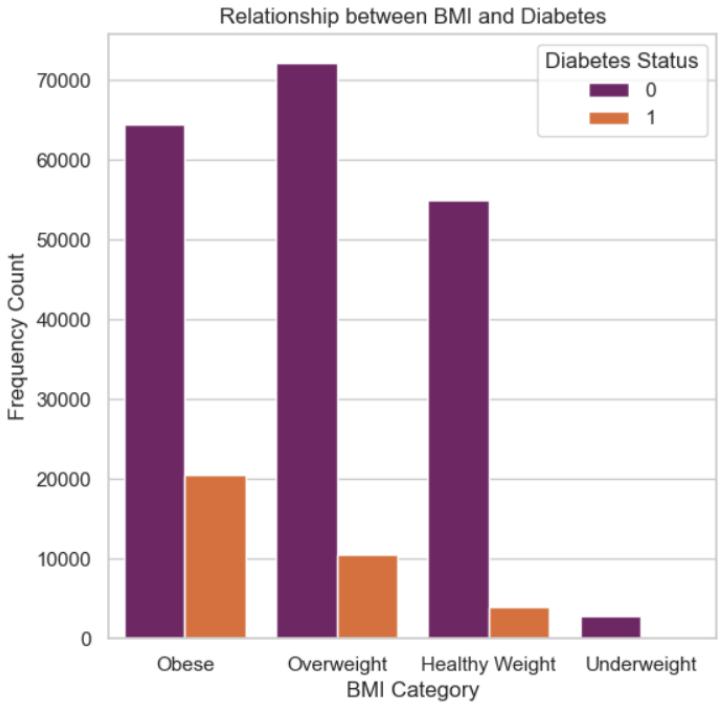

Figure 8: BMI Description & Diabetes

Figure 8 shows that weight plays a key role in developing Diabetes. Furthermore, the healthier you are and the more in control of your weight, the less likely you will develop diabetes.

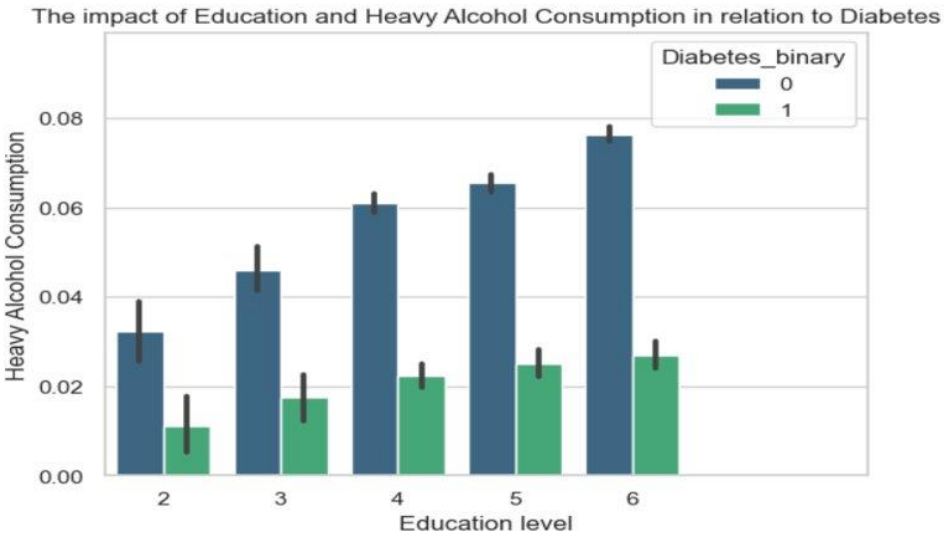

Figure 9: Education & Heavy Alcohol Consumption in relation to Diabetes

Figure 9 shows that a person with a higher education level is more likely to drink alcohol in higher concentrations and is more likely to develop diabetes. Education and alcohol consumption become variables that are likely to impact diabetes classification.

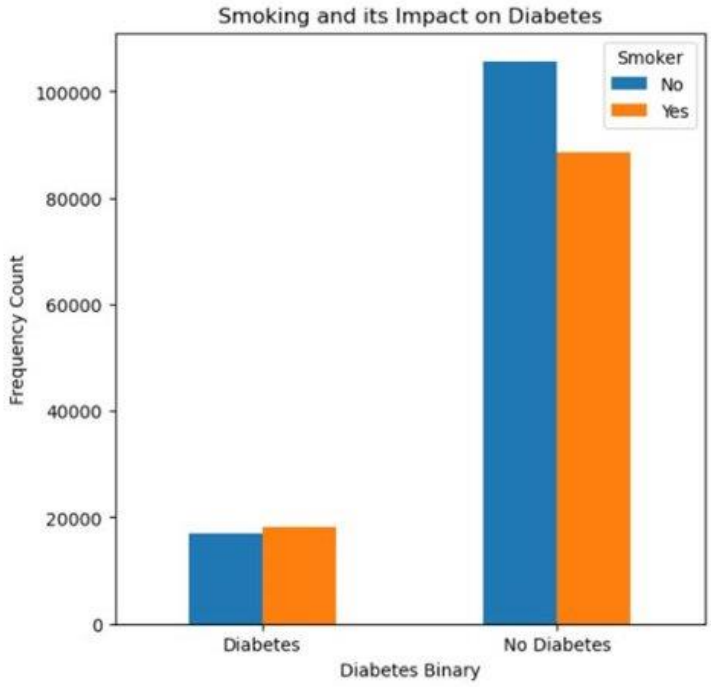

Figure 10: Smoking and its impact on Diabetes

Figure 10 shows that smoking may not necessarily have a substantial impact on Diabetes. A person who smokes has a greater chance of developing diabetes. However, it is not the only factor that determines whether a person will develop diabetes.

To conduct a city and state analysis, we used our city data to analyze diabetes distributions across different states and cities. Figure 11 below shows the diabetes distribution across the top ten cities in California. Figure 11 shows that Hemet, San Bernadino, and El Monte have the highest concentrations in California.

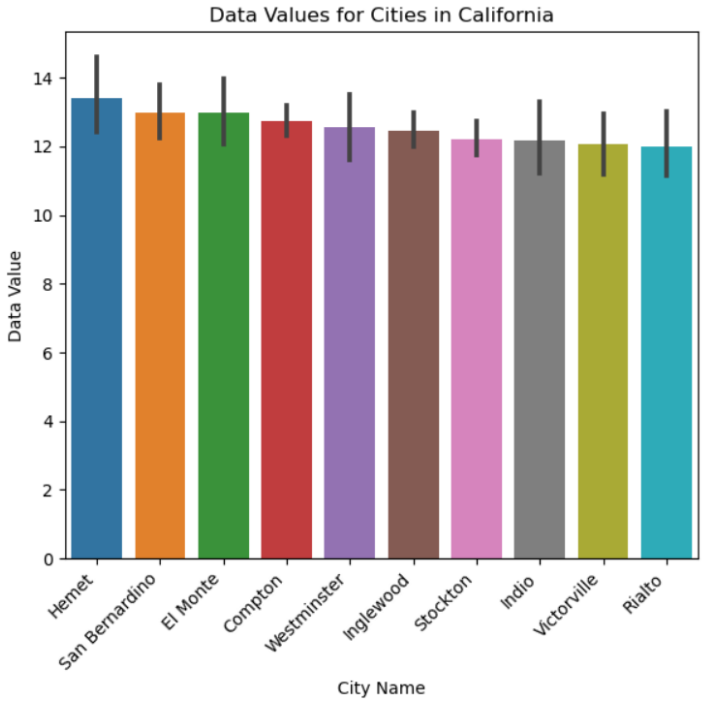

Figure 11: Top 10 Cities in California

Figure 12 below shows the top 10 cities in Florida. Hialeah, Miami Gardens, and Miami have the highest concentrations of Diabetes in Florida.

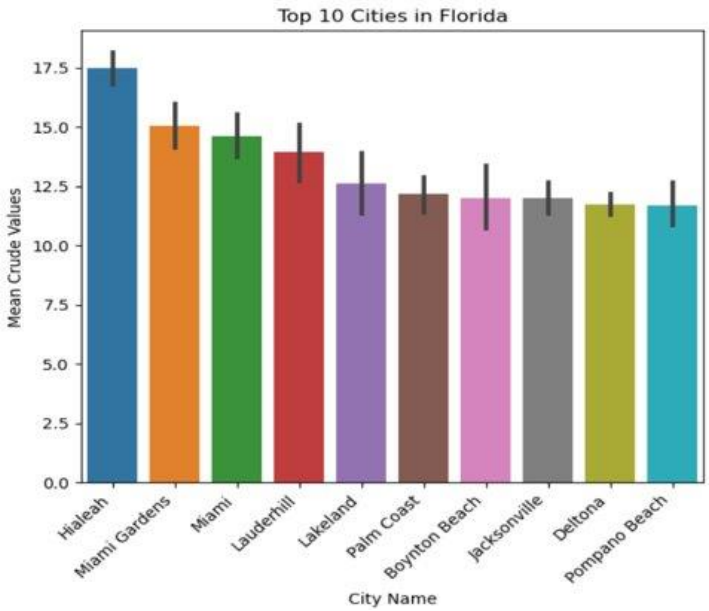

Figure 12: Top 10 Cities in Florida

Using the Folium library, we were able to build an interactive map that displayed the concentrations across the United States. The map is interesting because it provides an analysis of concentrations across the entire US. This is shown below in Figure 13.

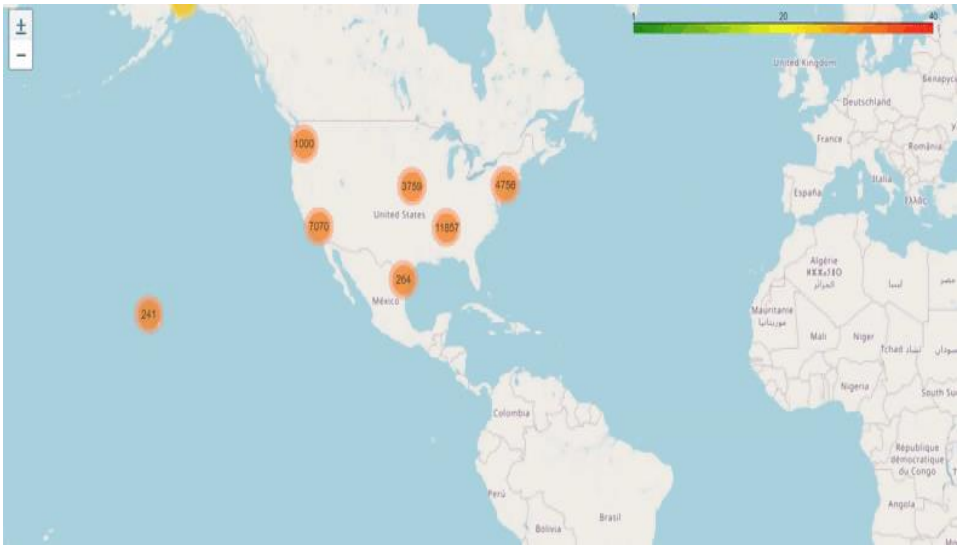

Figure 13: US Map of Diabetes Concentrations

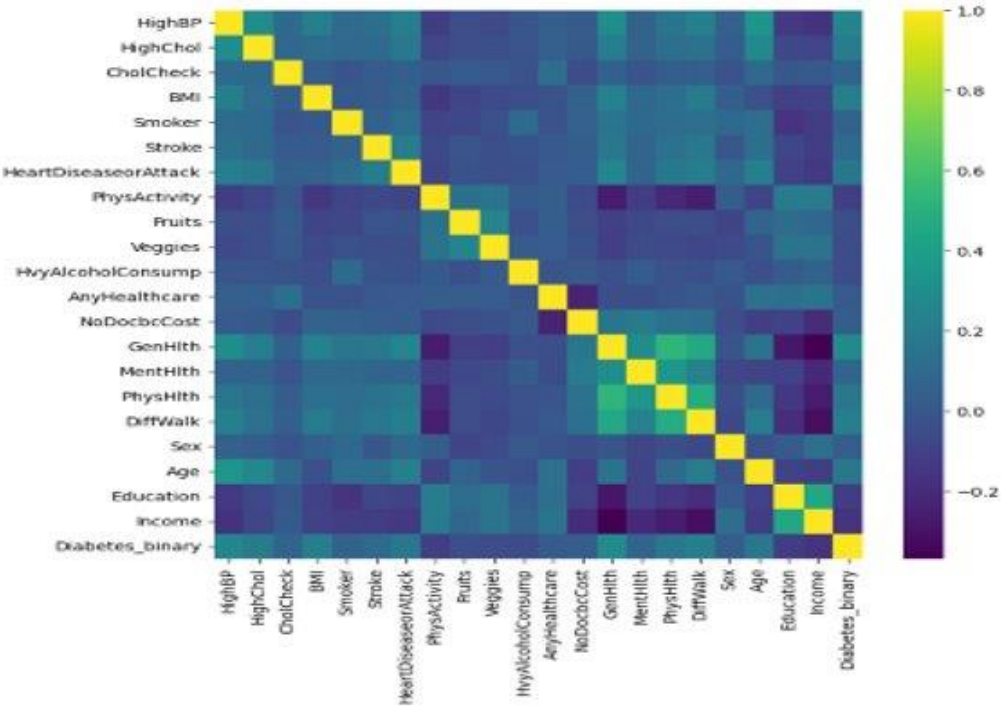

Figure 14: Correlation Matrix

Figure 14 displays the correlation matrix showing the features that are highly correlated with each other. Figure 15 below provides a bar chart showing the correlated features in

descending order of correlations. Variables such as General Health, High BP, Diff Walking, BMI, High Cholesterol, and Age are the highest correlated features.

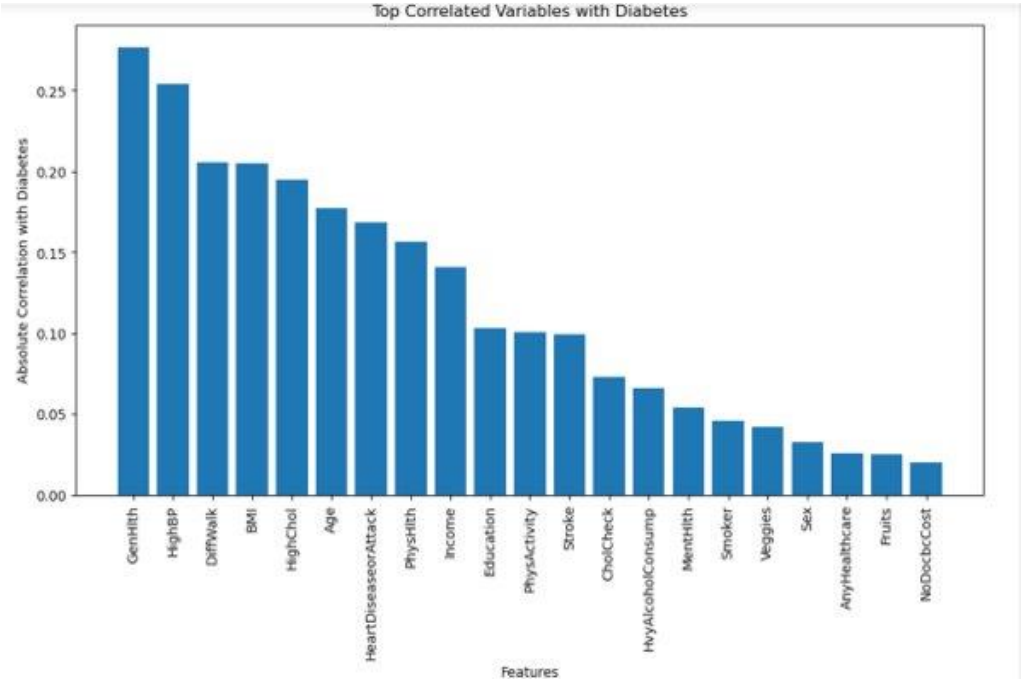

Figure 15: Correlated Features

To select the most salient features for our classification models, we utilized the Select K Best feature method using chi squares. Their feature importance score is listed below in Figure 16.

Feature	Feature Score
Physical Health	97988.76
BMI	15507.73
Mental Health	11419.58
Age	8539.90
High BP	8098.54
Difficulty Walking	7875.49
General Health	7671.73
Heart Disease or Attack	5822.14

High Cholesterol	4869.31
Income	3377.09
Stroke	2156.67
Heavy Alcohol Consumption	937.40
Physical Activity	617.56
Education	479.11
Smoker	253.82
Sex	137.83

Figure 16: Features for Classification Model

Using the top 16 features listed in Figure 16, we built classification models based on our resampled data from the Near Miss algorithm. Utilizing a balanced dataset and the top 16 features we were able to build models with high accuracy. The classification models and their accuracy are listed below in Figure 17.

Classification Model	Accuracy
Decision Tree Classifier	80.6%
KNN	83.2%
Random Forest Classifier	86.4%
SVM	85.0%
XGBoost	85.2%
MLP Classifier	85.1%
ANN	86.5%
Ridge Classifier	84.7%

Logistic Regression	84.9%
Passive regression	76.5%

Figure 17: Classification Models & Accuracy Score

The classification model performance chart reveals the efficiencies of diverse machine learning algorithms in predicting diabetes, with accuracy ranging from 76.5% to 86.5%. The Random Forest, XGBoost, MLP Classifier, Ridge Classifier, and ANN exhibited the highest accuracies making them the superior model for this study. These results show the versatility between different algorithms enhancing the adaptability of diabetes prediction models in the healthcare application. Ultimately, the ANN had the highest accuracy and was thus chosen for further analysis for the remainder of the study.

In explaining the architecture of the ANN, we created a Sequential Model with 3 layers. The input layer (1st layer) contains 16 neurons for the 16 features we used in our models. We also used ReLu as our activation function. For the second layer (hidden layer), we continued using 16 neurons and ReLu as the activation function. For the third layer (output layer), we used 1 neuron and sigmoid as the activation function to output 0 or 1 for the classification of diabetes. The model was compiled using the Adam Optimizer and early stopping was implemented. We set the batch size to 10 and conducted a 20% validation split to monitor validation loss and accuracy while training the model. The model trained until it had reached maximum accuracy and loss was minimum to prevent it from overfitting. The results are shown below in Figure 18 and Figure 19.

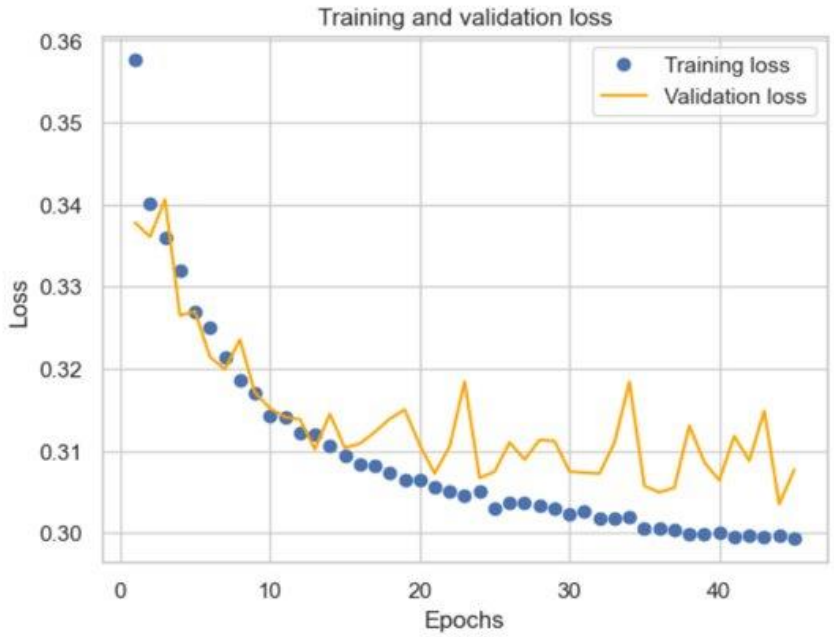

Figure 18: Training and Validation Loss

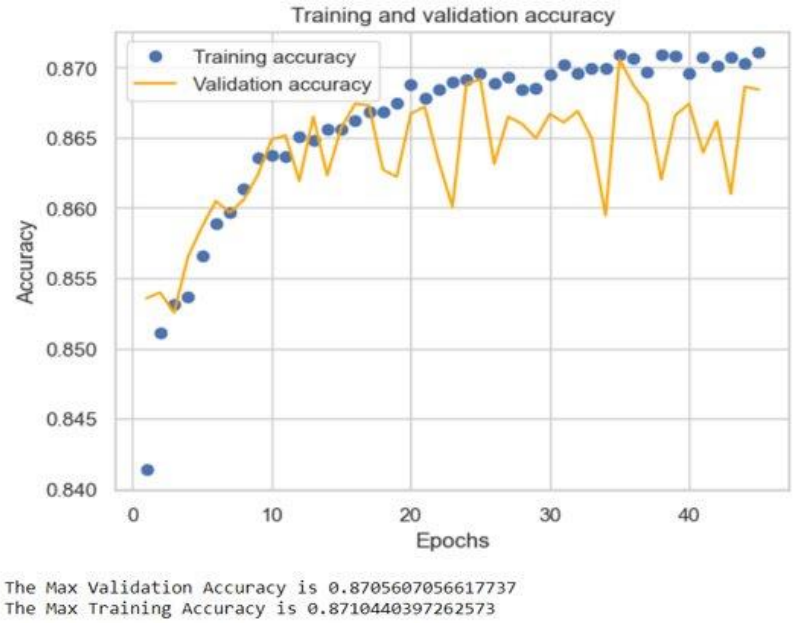

Figure 19: Training and Validation Accuracy

The evaluation metrics of our model are shown below in Figure 20 and Figure 21.

	Precision	Recall	F1-score	Support
0	81%	96%	88%	10454
1	95%	77%	85%	10605
Accuracy			87%	21059
Weighted Avg	88%	87%	86%	21059

Figure 20: Classification Report

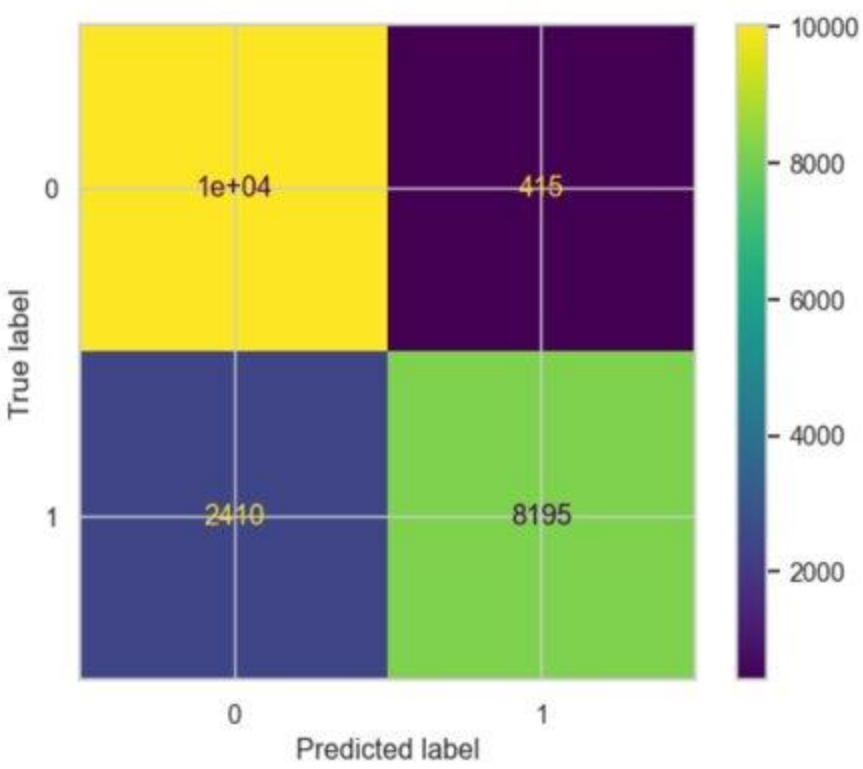

Figure 21: Confusion Matrix

After understanding the labels that it was able to predict correctly, we decided to get insight into the features after modelling to see which features impacted the model's outcome. Using SHAP values, we got the features that had the highest relevance or importance in predicting. The plot of the top 5 features and their respective scores are shown in Figure 22 and Figure 23.

Feature Name	Mean Abs Importance Score
PhysHlth	0.13
Age	0.11
BMI	0.09
MentHlth	0.03
HighBP	0.02

Figure 22: Top 5 SHAP Feature Scores

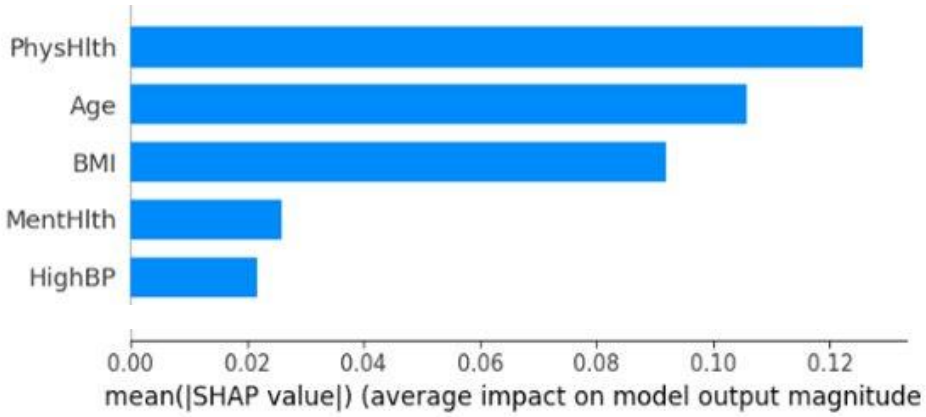

Figure 23: Top 5 SHAP Feature Plot

SHAP refers to Shapley Additive Explanations and calculates the impact of a feature on its target variable. Each feature is considered a player, and the dataset is a team. Each feature contributes to the dataset. Utilizing this methodology, SHAP uses combinatorial calculus to find feature importance. For our ANN model, Physical Health, Age, BMI, Mental health, and High BP were the highest features of relevance that had predictive capability. Using SHAP was beneficial as it allowed us to gain a deeper insight into the most salient features to make more informed decisions.

Recommendations

In making data-driven recommendations based on our analysis, we first recommend focusing on the most prominent features when studying diabetes. These include Physical Health, Age, BMI, Mental Health, and High Blood Pressure. We should also focus less on features such as No Doc Because of Cost, Veggies, Fruits, Cholesterol Check, and Any Health Care.

Secondly, we recommend individuals to improve their physical health by exercising daily and making more health-conscious choices. Changes in their lifestyle to prioritize health such as less alcohol and drug intake, maintaining a balanced diet, more exercise, and a greater focus on any other medical conditions would be recommended to benefit the most important feature of diabetes that we found in the study: physical health. In addition, giving greater importance to mental health conditions would be advised, as it seems to impact diabetes more than one may commonly assume. Increasing awareness of risk factors such as higher age, mental health, and high blood pressure would be recommended to individuals to prevent the risk of diabetes. Individuals that fall into any of these categories are recommended to screen for diabetes regularly. Furthermore, individuals are advised to maintain a healthy BMI via exercise, a healthy diet, and medical evaluations.

Thirdly, we recommend getting involved in programs such as the National Diabetes Prevention Program by the CDC. This program is a partnership between public and private organizations to prevent or delay diabetes. Through this program individuals can learn the risks of diabetes and how best to prevent it. Furthermore, programs that aim to reduce diabetes

should focus on areas where Diabetes is prevalent by utilizing the map tool that was built. Our map shows areas of high concentration where project members can target and conduct awareness campaigns. Through these measures we can reduce the risk of developing diabetes and live healthier lives.

Conclusion

In conclusion, our research aimed to predict diabetes using data analytics and machine learning to gain valuable insights from our data. Through analysis of our two datasets, we developed classification models, explored geographical variations, and identified key factors of diabetes risk. Our classification models all demonstrated a good understanding of predicting diabetes based on health indicators. Exploratory data analysis revealed a significant correlation between certain health indicators, emphasizing the importance of considering multiple factors in diabetes predictions. Features such as education, physical activity, and mental health all emerged as influential in understanding diabetes prevention. Geographical analysis displayed variations in diabetes across different states and cities across the United States. SHAP values offered interpretable insight into the impact of individual features on the model prediction. Knowing these SHAP values allows one to understand the main factors contributing to diabetes risk.

While our analysis covered a wide variety of health indicators, there might be more factors influencing diabetes than we did not consider in this study. Incorporating other factors such as genetics, dietary habits, and environmental factors can improve this study by giving more perspectives. Our study focused on the United States but extending the analysis to a global scale could provide more insight into regional variations and give us more information to understand diabetes prevention. Future works can also be accompanied by the integration of genetic data, fine-tuning hyperparameters, and ensemble modeling to gain more insight into diabetes prevention.

References

CDC diabetes health indicators. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

Centers for Disease Control and Prevention. (2023, March 7). Prevalence of diagnosed diabetes in adults by diabetes type - United States, 2016. Centers for Disease

Centers for Disease Control and Prevention. (2023, August 1). *National Diabetes Prevention Program*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/prevention/index.html

Centers for Disease Control and Prevention. (2023, September 5). *What is diabetes?* Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/diabetes.html

Centers for Disease Control and Prevention. (n.d.). 500 cities: Diagnosed diabetes among adults aged >=18 years. Centers for Disease Control and Prevention. https://data.cdc.gov/500-Cities-Places/500-Cities-Diagnosed-diabetes-among-adults-aged-18/cn78-b9bj

Control and Prevention. https://www.cdc.gov/mmwr/volumes/67/wr/mm6712a2.htm#T1_down

World Health Organization. (n.d.). *Diabetes*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes