Onset Diabetes Prediction

Programmers/Authors:
Jacob Manangan - 190629720
Gautam Sood - 190417060

Professor:
Sukhjit Singh

Class:
CP322 - Machine Learning, Winter 2023

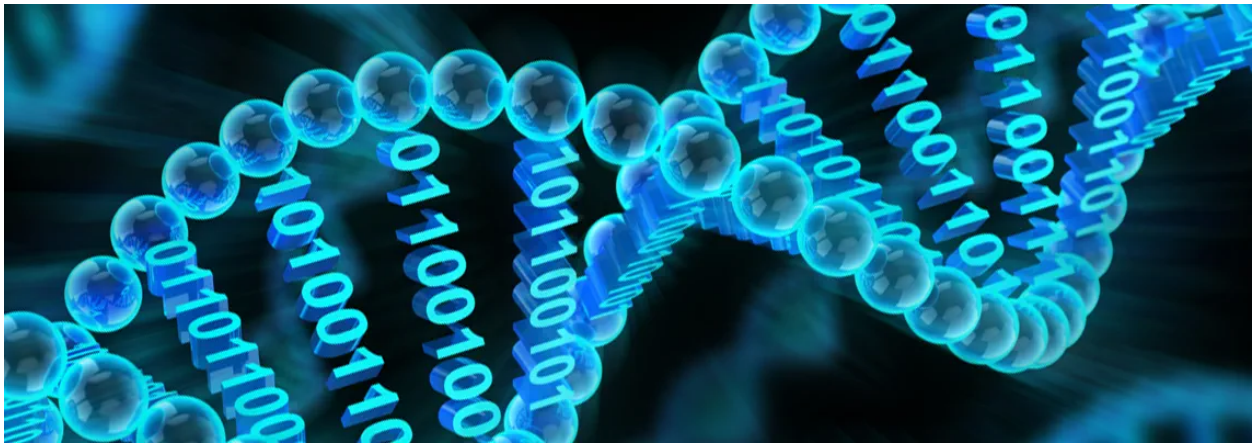


Table of Contents

Abstract

Diabetes causes significant impacts on individuals and healthcare systems making it one of the most prevalent chronic illnesses worldwide. While cases of diabetes are inevitable, early detection and prevention of the diseases are crucial for improving outcomes and reducing healthcare costs.

In the bioinformatics field, machine learning has emerged as a favourable tool for predicting the onset of these diseases based on numerous health and demographic variables.

Our research aims to analyze a wide range of variables such as cholesterol level, BMI, Blood Pressure, Alcohol consumption, eating habits and many more to identify key trends and patterns that correlate with the disease. We also trained, developed, and deployed cutting edge machine learning models such as logistic regression, Gaussian Naive Bayes, Decision Trees, and Random Forest Classifier in order to predict the onset signs of diabetes.

Overall, our study demonstrates the potential of how statistics and machine learning could be incorporated within the healthcare industry, which could inform premature treatment to improve patient results and reduce healthcare costs.

Introduction

Diabetes is a disease that affects over 537 million people globally (CDC, 2020). That means roughly 1 in 10 people have some form or may develop some form of diabetes in their life. Due to unhealthy eating habits and lack of exercise this number is said to rise to 637 million by 2030 (CDC, 2020). Scientists have been trying to develop methods to better predict people who will develop diabetes to catch the disease early.

As machine learning has grown over the past decade many researchers and computer engineers have ventured into the space of training ML models to help doctors with this very problem. Early detection is crucial in preventing or delaying further complications in a patient due to the illness.

Many times, a few changes in a person's lifestyle such as a better diet and regular exercise can make a huge impact on the progression of Diabetes and how severely it impacts its host. With over half a billion people already having some form of the disease, reducing the impact it has by catching it early can also benefit hospitals and other health care providers by lessening the load on the resources they need to employ to treat patients with critical cases.

In this paper we explore how novel ML models can be trained to predict early onset diabetes using a large dataset of people's medical information. We use the latest models such as logistic regression, Naive Bayes, Decision Trees, and Random Forest to develop a product that performs with a high level of accuracy. We evaluate the performance of our model using the most common evaluation metrics demonstrating its potential for early detection and prevention of diabetes.

Project Methodology:

Our research followed a general methodology of training machine learning models. The following steps outline a detailed view of our steps:

1. **Data Collection:** we collected our data through Kaggle which is an online resource that includes datasets and code. The dataset contains 253,680 survey responses with 21 feature columns and 1 target variable.

2. **Data Cleaning:** This step consisted of data wrangling in order to properly clean the data for future use. This step included handling missing values, dropping unneeded columns, and handling duplicated data.

3. **Exploratory Data Analysis (EDA):** Here, we visualized aspects of the dataset in order to gather better information needed for modeling. Our goal was also to find the correlation of data between the features and our target binary variable.

4. **Feature Engineering:** We performed a chi-squared analysis to find the top 75% of features that best predicted our target diabetes variable.

5. **Train/Test Split:** We utilized a test/train split of 80-20 to properly train and evaluate our machine learning models

6. **Model Implementation:** We implemented 4 different models that specialized in binary classification. The models we chose were logistic regression, Gaussian Naive Bayes, Decision Trees, and Random Forest Classifier

7. **Model Evaluation:** We used different metrics such as accuracy, precision, recall, F1-score, and confusion matrix to properly evaluate our models

8. **Model Deployment:** Our models can be run through your computer's terminal.

Data Collection:

Our dataset was collected through Kaggle which is a platform that provides machine learning and data science specific content. It provides free access to thousands of data sets and allows data scientists to collaborate on projects and learn from each other (Kaggle, 2022).

The data set includes medical information along with lifestyle information which was collected through various surveys and laboratory tests. It has 253,680 survey responses and 21 features collected directly from the CDC (center for disease control). Along with the demographic information and clinical measurements, the lifestyle responses such as physical activity and dietary intake were self-reported metrics (Kaggle, 2022).

Below is a breakdown of our dataset along with an explanation of each variable:

Column breakdown and feature variable explanation of data set:

Feature variable name	Description
Diabetes Binary	Binary response (0 or 1) to determine whether you've already been diagnosed with diabetes.
High Blood Pressure	Binary response (0 or 1). For adults who have been told they have high blood pressure by a doctor, nurse, or other health care professional.
High Cholesterol	Binary response (0 or 1). For adults who have been told that their blood cholesterol is high by a doctor, nurse, or other health care professional.
Cholesterol checkup	Binary response (0 or 1). Has the individual had their cholesterol checked in the past 5 years.
BMI	Body Mass Index. The individual's weight in kilograms divided by the square of their height in meters. Used to determine a person's obesity levels.

Smoker	Binary response (0 or 1). Have you smoked at least 100 cigarettes in your entire life. 5 packs is roughly 100 cigarettes.
Stroke	Binary response (0 or 1). Have you ever had a stroke.
Heart disease or attack	Binary response (0 or 1). Respondents that have ever reported of having a coronary disease or attack of any kind.
Physical Activity	Binary response (0 or 1). Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
Fruits	Binary response (0 or 1). Does the individual consume fruits at least once a day.
Veggies	Binary response (0 or 1). Does the individual consume vegetables at least once a day.
Heavy alcohol consumption	Binary response (0 or 1). Heavy drinkers, adult men having more than 14 drinks a week and adult women having more than 7 drinks per week.
Any healthcare	Binary response (0 or 1). Do you have any kind of health care coverage, including health insurance, prepaid plans, or government plans such as Medicare, or Indian Health Services.
No doctor cost	Binary response (0 or 1). Was there a time in the past 12 months when you needed to see a doctor but could not because of cost.

General health	(1-5 scale answer). How would you rank your general health. 1 being generally horrible and 5 being generally fantastic.
Mental health	Thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good. (0-30 scale) each value representing the number of days where your mental health was not good.
Physical health	Thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good. (0-30) scale each value representing the number of days where your mental health was not good.
Difficulty walking	Binary response (0 or 1). Do you have serious difficulty walking or climbing stairs.
Sex	Binary response (0 or 1). Indicates sex of respondents (Female is 0 and male is 1).
Age	Fourteen-level age category (1-14).
Education	What is the highest grade or year of school you've completed (1-6). Each representing how many years of education you've done.
Income	Is your annual household income from all sources above the countries median. (1-8) 1 being below the median and 8 being above the median.

Data Cleaning

We used a process of data cleaning protocols in order to sufficiently clean our dataset. Our dataset included 24206 rows of duplicated data that we dropped. In this phase, we also detected and dropped any null values.

We used two different data wrangling libraries in order to clean our data:

NumPy: NumPy is an extremely useful open-source python library meant for computing numerical operations on arrays and matrices (NumPy, 2022). It has built in functionality to support functions like linear algebra, Fourier transform, and random number generation. It also has more abstract functions such as its multidimensional array objects which are the foundation for many numerical computations in python. NumPy allows for fast and memory-efficient manipulation of large tabular data sets and gives the developer the ability to slice, reshape, and index the data in whichever way is required for the application (NumPy, 2022).

Pandas: Pandas is another amazing open-source library for data analysis and manipulation. It provides powerful tools for data exploration, data cleaning, and data transformation (ActiveState, 2022). This library is built on the previously mentioned NumPy and has added functionality for specific data structures such as Series and DataFrame. In data science and machine learning pandas plays a crucial role in data preprocessing and feature selection (ActiveState, 2022).

Exploratory Data Analysis (EDA)

We used two different data visualization libraries in order to explore our data:

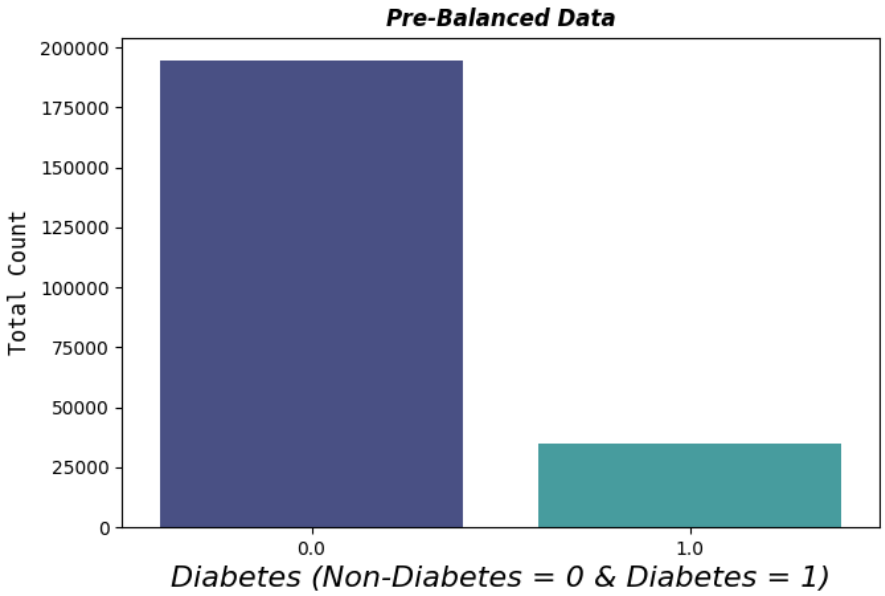
Matplotlib: Matplotlib is an open-source python library that data scientists and machine learning engineers use for visualization, creating charts and plots, and for creating informative graphs (Matplotlib, 2023). This library can be used to visualize multiple data types and formats, including scatter plots, time series, heatmaps, and histograms. Matplotlib is an excellent tool for analyzing data in its early preprocessing and cleaning phase. It allows developers to better identify correlations, patterns, and anomalies in data which can help with model selection and feature engineering (Matplotlib, 2023).

Seaborn: Seaborn is a data visualization open-source library for python. It's built upon the matplotlib library and is an excellent tool for generating graphs, and charts based on tabular data (Seaborn, 2022). It provides a high-level interface for creating statistical graphics that aren't only informative but also visually attractive. This has made it an extremely popular tool in the machine learning community and among other data scientists. Seaborn offers a variety of different charts and graphs such as scatterplots, heatmaps, and line charts that can be customized using various settings to better identify and display relationships and dependencies present in the data. This library is also frequently used to explore correlations between different features,

analyze the performance of the models, and to visualize the distribution of target variables and features (Seaborn, 2022). This tool's excellent ability to help visualize complex data has made it an extremely useful part of the machine learning model development and evaluation process.

Visualizations

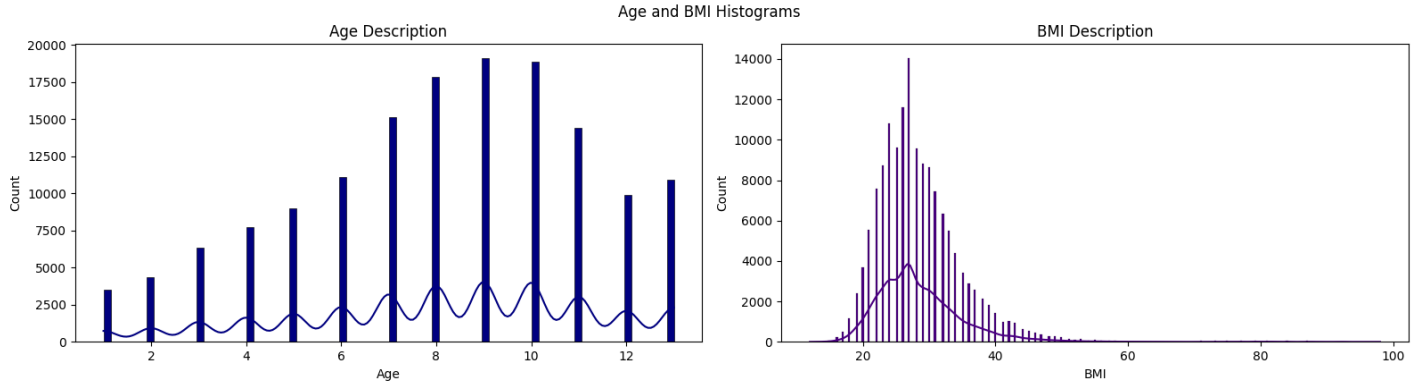
Proportion of Diabetes



Observation:

- Our dataset is imbalanced by containing a larger proportion of people who have diabetes than without diabetes. There are 194377 who do not have the disease while only 35097 who do have the disease.

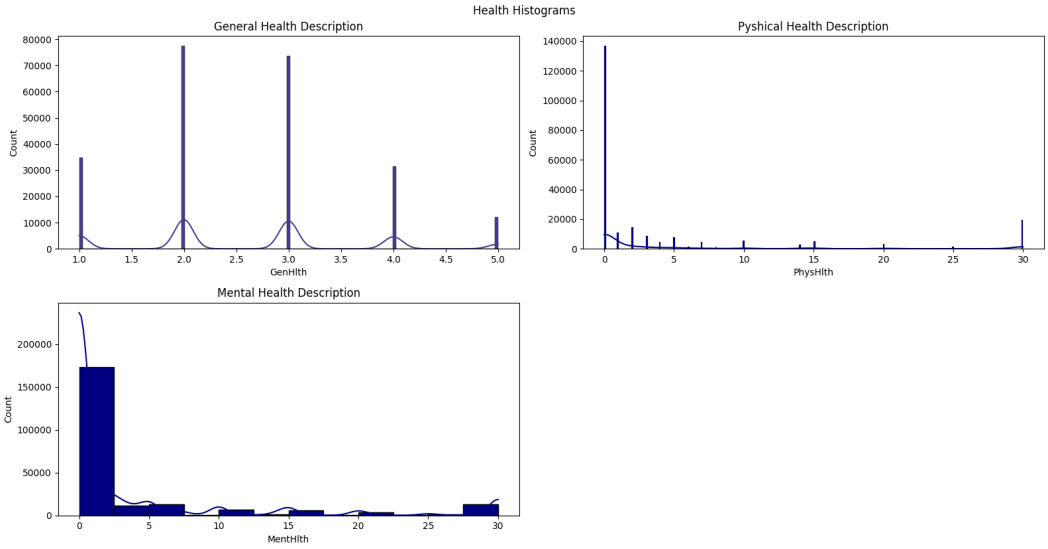
Age and BMI Histograms



Observation:

- Both Age and BMI have a Normal Distribution

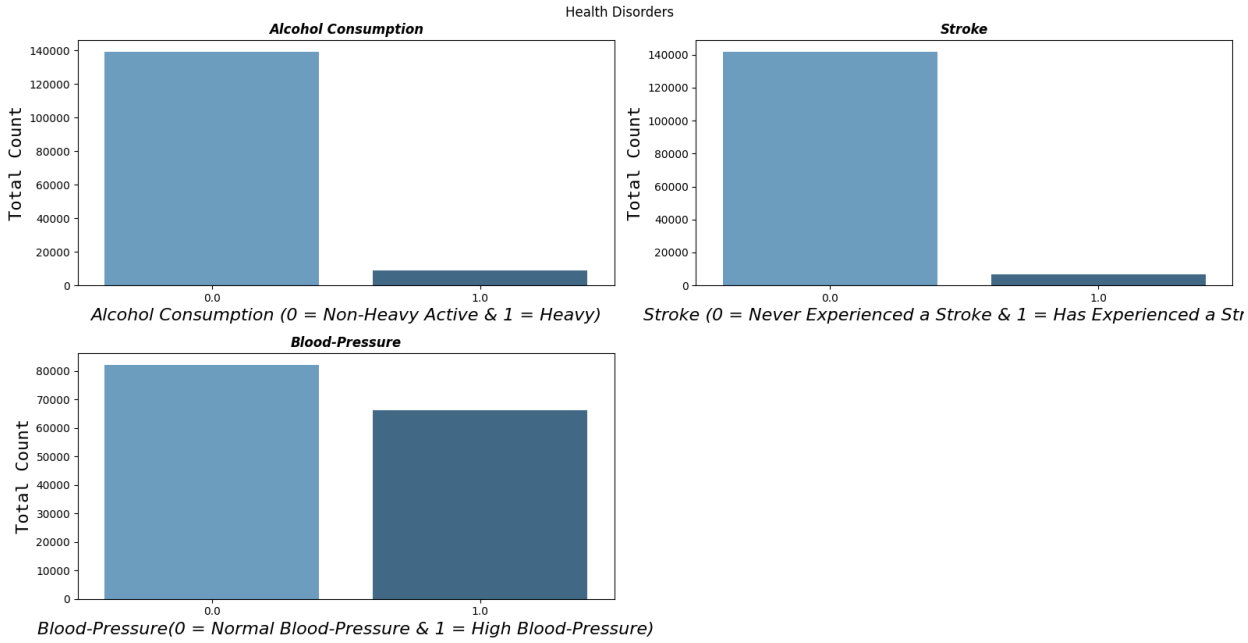
Health Histograms



Observation:

- General Health has a Normal Distribution with a median at 3.0
- Physical Health is left heavy but has an increase at 30
- Mental Health is left heavy meaning more people have lower mental health problems

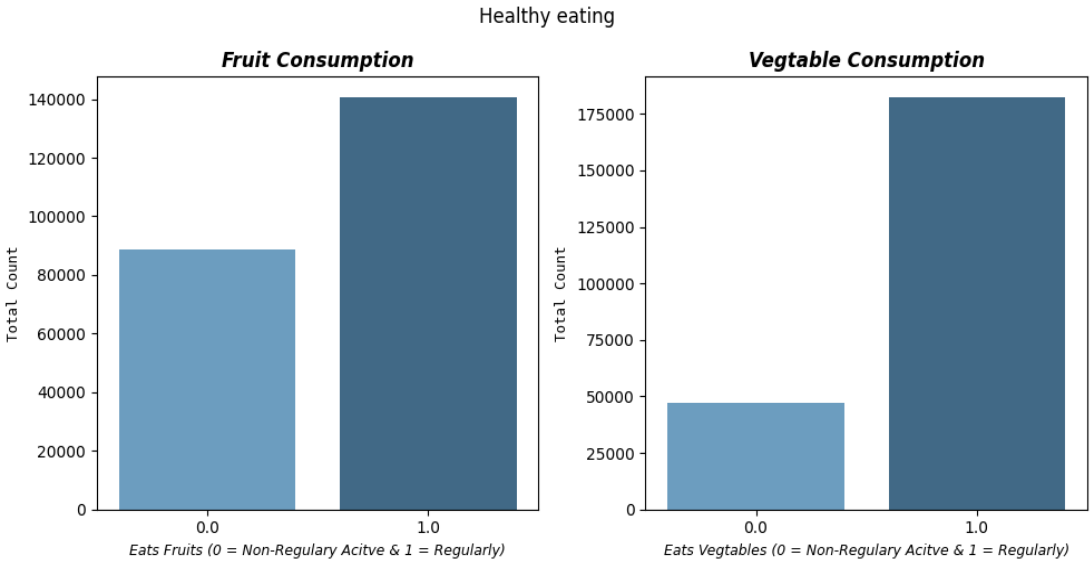
Health Disorders Visualizations



Observation:

- Low amount of alcoholics within this dataset
- A lot more people who never experienced a stroke
- Amount of people who have a high blood pressure is high

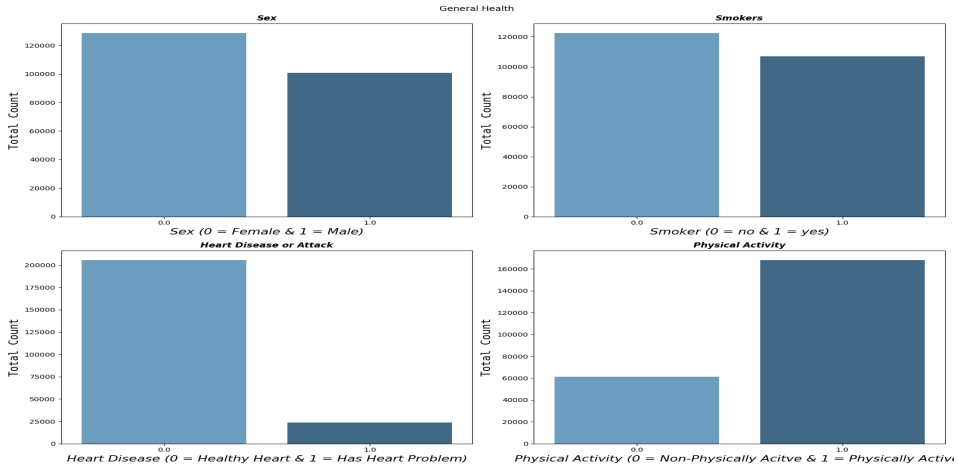
Healthy Eating Visualizations



Observation:

- More people eat fruit regularly
- More people eat vegetables regularly

General Health Visualizations



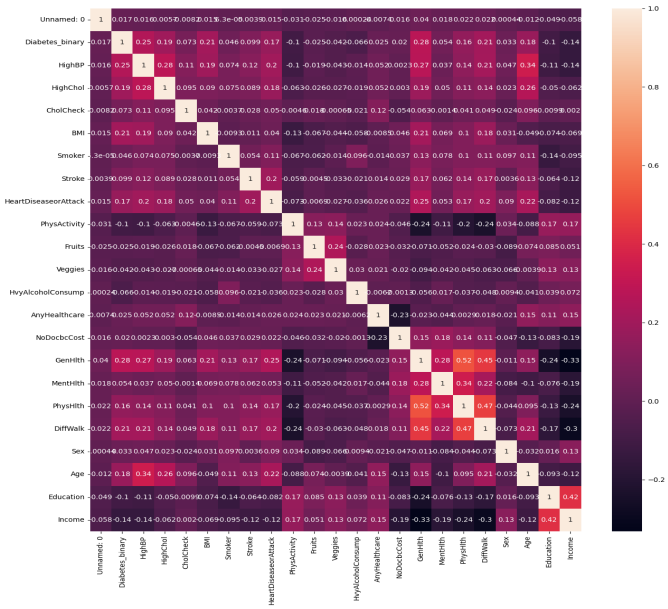
Observation:

- Relatively equal male to female ratio, but slightly female dominant
- Relatively equal smoker to non-smoker ratio, but slightly non-smoker dominant
- More people have healthy hearts in this dataset
- Larger proportion of physically active individuals

Feature Engineering

Feature selection is an important machine learning process that's used to create a subset of relevant features or variables taken from a larger dataset. Its primary goal is to identify the most important features in any given dataset and only use those when training the predictive model. This helps eliminate any outliers or unnecessary noise when working with large data and in result trains a model which has a higher level of accuracy. Feature selection approaches can be categorized into three main buckets: filter methods, wrapper methods, and embedded methods. Filters use numeric measures to rank and select the most relevant features, whereas wrappers test different subsets of features with multiple models to determine which one has the highest accuracy. Embedded methods combine feature selection along with model training to better help in identifying the variables that lead to the most accurate predictions. Feature selection is an important step in the machine learning process and should not be overlooked when trying to develop an accurate mode.

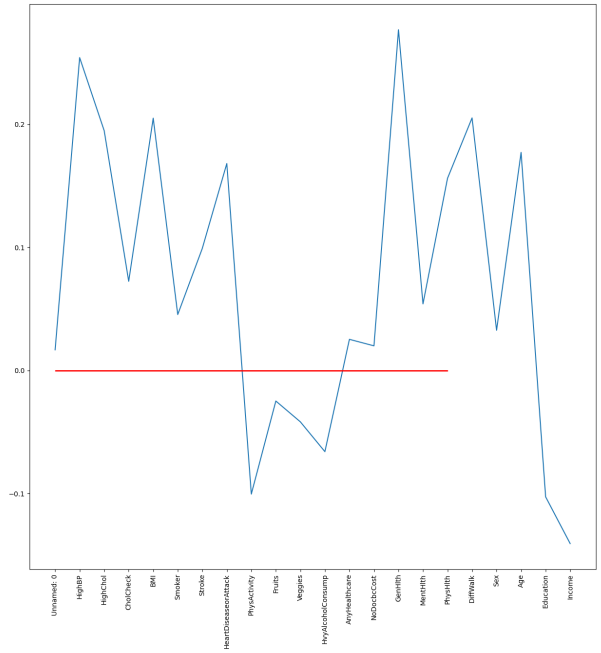
Variable Relationship



Observation:

- Income and Education had a high positive correlation
- Physical Activity and DiffWalk have a negative correlation
- General Health and Physical Activity also have a negative correlation

Correlation With Diabetes



Observation:

- High Blood Pressure, BMI, General Health, Physical Health, and Age had the strongest correlations with Diabetes
- Physical Activity and Heavy Alcohol Consumption had the strongest negative correlations with Diabetes

Chi-squared test

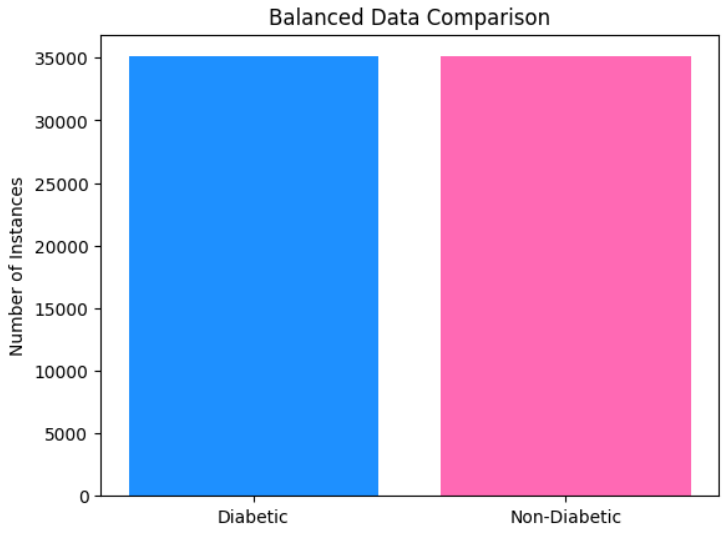
When doing feature selection on our own data set, we chose to use the chi-squared to find the best variables to train our model on. Chi-squared is a statistical method used primarily in feature

selection to determine the strength of association between two categorical variables. Its main goal, just like any other feature selection method, is to determine which features are most relevant when predicting the target variable and training the model (JMP, 2023). The test calculates the difference between the observed frequency and the expected frequency of each category in a feature, and then computes a statistic based on the squared differences. The computed value is then compared to a chi-squared distribution to determine whether the feature should be included in the model (JMP, 2023). A high chi-squared value indicates that there is a strong connection between the feature and target variable, making it an important predictor to keep in the model.

In our analysis, we chose to include the top 75% of features based on their chi-squared metrics.

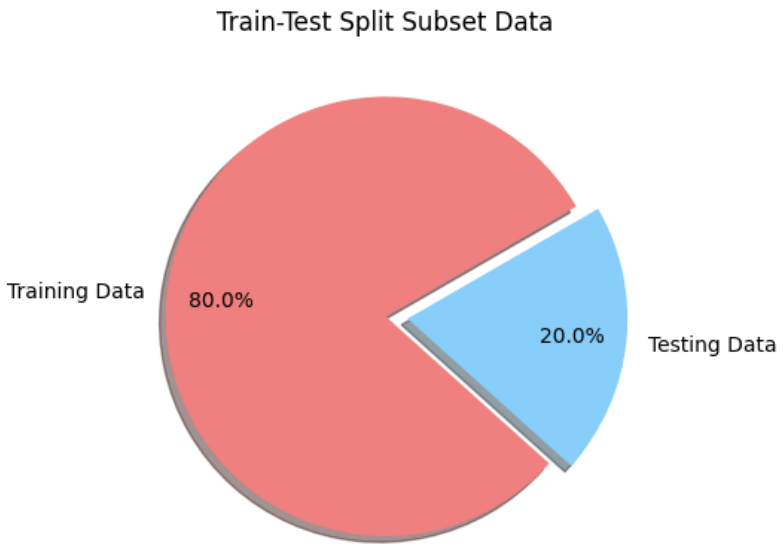
Model Balancing

To combat the imbalance of the dataset, we use a method called Up-Sampling where we take the dataset and artificially reproduce instances of diabetes to create a 50-50 split between diabetic and non diabetic patients. This helps our models to better understand what diabetic instances look like to increase our predictive precision and accuracy.



Train Test Split

We used a train-test split procedure to properly estimate the performance of the machine learning algorithms. To do this, we split the dataset into two subset datasets whereby the Training dataset is used to train the algorithms and the Test dataset is used to evaluate the models performance. For this case, we split the subsets into 80% training data and 20% testing data.



Model Implementation

We used one library to implement our machine learning models

Ski-kit learn: When selecting which models, we wanted to train our data set on. We looked at the most popular and widely used options from Scikit-learn. Sklearn for short is an extremely useful and popular open-source machine learning library for python (SciKit, 2023). It offers a variety of popular algorithms such as regression, classification, dimensionality reduction, and clustering, as well as functions for data cleaning, preprocessing, and model evaluation. These models are used in machine learning by providing a set of functions and classes that allow users to easily implement and train various models on their dataset making it a powerful and flexible library for machine learning tasks (SciKit, 2023).

Logistic regression

Logistic regression is a machine learning model that is used to analyze the relationship between a dependent and one or more independent variables. Logistic regression is one of the most popular

methods of classification in situations where the prediction is more important than the relation. This method is primarily used to predict the probability of an event occurring, in this case whether an individual will develop diabetes at some point in their life. Logistic regression works by fitting a logistic curve to the data that maps the input variables to the predicted probability of the event occurring. This is considered a binary classification algorithm as its prediction is either true/false or yes/no.

Result for Logistic Regression:

Accuracy: 0.7379.

Confusion Matrix:

TP = 2844	FP = 779
FN = 927	TN = 1960

Classification Report

	precision	recall	f1-score	support
0.0	0.75	0.78	0.77	3623
1.0	0.72	0.68	0.70	2887
accuracy			0.74	6510
macro avg	0.73	0.73	0.73	6510
weighted avg	0.74	0.74	0.74	6510

Gaussian Naïve bayes:

Gaussian Naïve Bayes is another important machine learning algorithm that is used primarily for classification problems. This algorithm originates from Bayes Theorem which is another

statistical rule that predicts the probability of an event happening based on prior evidence or knowledge. This algorithm assumes that all features in the data set are independent which makes the runtime more efficient and less computationally expensive. It also allows for easy implementation in a variety of classification-based problems. Naïve Bayes computes the conditional probability of each class given a set of features, and then selects the class with highest probability as the predicted outcome. It's an extremely elegant algorithm due to its simplicity and speed and has become widely used in the machine learning community.

Results for Gaussian Naïve Bayes:

Accuracy: 0.7028

Confusion Matrix

TP = 2790	FP = 833
FN = 1102	TN = 1785

Classification Report

	precision	recall	f1-score	support
0.0	0.72	0.77	0.74	3623
1.0	0.68	0.62	0.65	2887
accuracy			0.70	6510
macro avg	0.70	0.69	0.70	6510
weighted avg	0.70	0.70	0.70	6510

Decision trees

Decision trees like logistic regression are used for classification tasks, they can be used for regression but in our case, we used a classification tree. A tree visually shows all the decisions in order and their possible consequences, where each internal node represents a decision-based feature, and each leaf node represents one of the predicted outcomes. Decision trees work by

recursively splitting the training data based on the statistical significance of the features, with each node having the goal of creating the most informative split. Once the algorithm has run and the tree has been constructed, it can be used to make predictions on new data by following the path from the root of the tree to a specific leaf node.

Results for Decision Tree:

Accuracy: 0.9135

Confusion Matrix:

TP = 3088	FP = 535
FN = 56	TN = 2831

Classification Report

	precision	recall	f1-score	support
0.0	0.98	0.85	0.91	3623
1.0	0.84	0.98	0.91	2887
accuracy			0.91	6510
macro avg	0.91	0.92	0.91	6510
weighted avg	0.92	0.91	0.91	6510

Random forest

Another machine learning model we employed was a random forest. This algorithm builds atop of decision trees in the sense that it combines multiple decision trees to create an even more predictive model. In random forests, each decision tree is trained on a subset of the input data and a random selection of input features. The RF combines the result of the decision trees to

make a final more well-informed prediction. In doing so random forests can reduce overfitting and improve the overall generalization and prediction accuracy of the model.

Results for Random Forest:

Accuracy: 0.9531

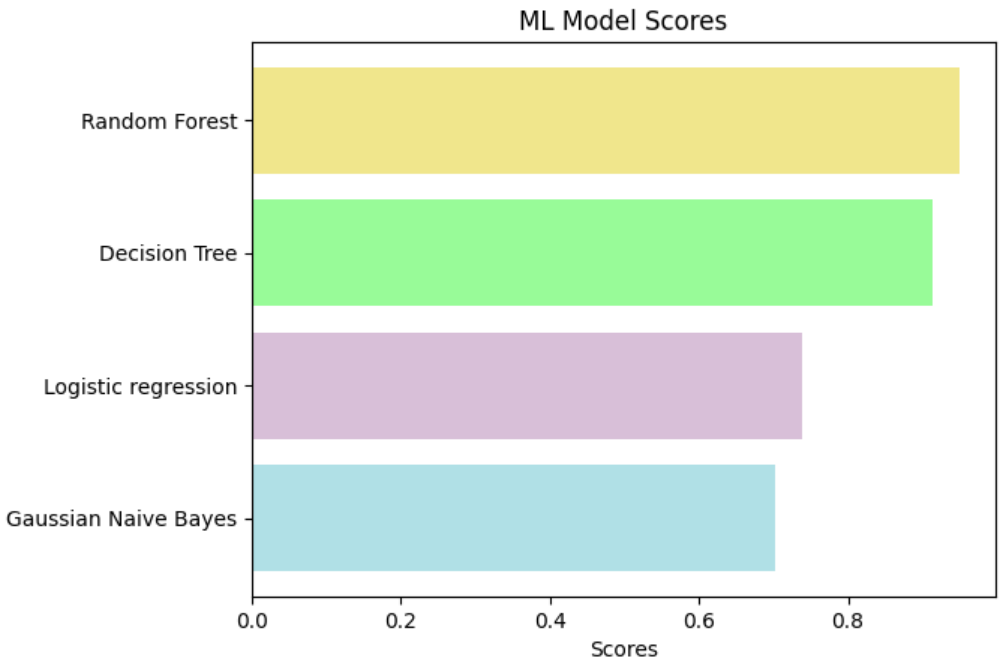
Confusion Matrix:

TP = 3350	FP = 273
FN = 51	TN = 2836

Classification Report:

	precision	recall	f1-score	support
0.0	0.99	0.92	0.95	3623
1.0	0.91	0.98	0.95	2887
accuracy			0.95	6510
macro avg	0.95	0.95	0.95	6510
weighted avg	0.95	0.95	0.95	6510

Results and Experimental Analysis



The results of our research demonstrate that machine learning models are able to accurately predict instances of diabetes with a high degree of accuracy. Our team compared the performance of a variety of binary classification machine learning models such as Logistic Regression, Gaussian Naive Bayes, Random Forest Classifier, and Decision Tree Classifier.

To evaluate our algorithms, we used a variety of metrics such as accuracy, confusion matrix, precision, recall, and F1 score to determine the best performing models.

Our analysis found that the Random Forest Classifier performed the best by achieving an accuracy of 95.31% in correctly predicting diabetes of patients. Our Decision Tree Classifier also performed well with an accuracy of 91.35%.

Taking a deeper dive into our best performing model, the Random Forest Classifier, the results show from class 1 (diabetes present) a precision score of 0.91 and recall score of 0.98. The results from class 0 (non-diabetes present) show a precision score of 0.99 and recall score of 0.92. These highly promising results show that the model is slightly better at classifying class 0 (non-diabetes presents) though by a very small margin.

With all the classification report scores being above 0.91, it shows that our model has performed well in predicting the target variable of diabetes. Also, the high F1-score of 0.95 demonstrates the high balance of precision and recall of our model.

Our team's feature importance analysis showed the most prominent and important features for predicting diabetes were High Blood Pressure, BMI, General Health, Physical Health, and Age. These results provide valuable insight for healthcare and medical professionals in identifying patients' risk of diabetes.

The next steps for improving our model would be to lower the false negative instances (type ii error). While our model only had 51 occurrences of this out of the 6510 sample, in the context of the medical field, false negatives can be alarming due to the concern of the potential delayed diagnoses and treatment for the patient.

Conclusion

Overall, our study and results demonstrate the power that machine learning algorithms have in predicting diabetes and provide information for the use of clinical features as predictors. While our research delivered promising results, further testing with additional diverse patient data would be needed to validate our findings in order to be used in the real world health care industry.

Citations:

CDC Global Health - Infographics - World Diabetes Day (2020) *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/globalhealth/infographics/diabetes/world-diabetes-day.html#:~:text=Today%20415%20Million%20people%20worldwide,a%20billion%20will%20have%20diabetes. (Accessed: April 6, 2023).

The chi-square test (no date) *JMP*. Available at: https://www.jmp.com/en_be/statistics-knowledge-portal/chi-square-test.html (Accessed: April 6, 2023).

Matplotlib 3.7.1 documentation (no date) *Matplotlib documentation - Matplotlib 3.7.1 documentation*. Available at: https://matplotlib.org/stable/index.html (Accessed: April 6, 2023).

S, S. (2022) *What is Pandas in python? everything you need to know*, *ActiveState*. Available at: https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/ (Accessed: April 6, 2023).

Ski-Kit Learn (no date) *scikit*. Available at: https://scikit-learn.org/stable/ (Accessed: April 6, 2023).

Statistical Data Visualization (no date) *seaborn*. Available at: https://seaborn.pydata.org/#:~:text=Seaborn%20is%20a%20Python%20data,attractive%20and%20informative%20statistical%20graphics. (Accessed: April 6, 2023).

Teboul, A. (2021) *Diabetes health indicators dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv (Accessed: April 6, 2023).

What is numpy? (no date) *What is NumPy? - NumPy v1.24 Manual*. Available at: https://numpy.org/doc/stable/user/whatisnumpy.html (Accessed: April 6, 2023).