Brian Yi

Dr. Pottenger, Dr. Nakamura, Dr. Nelson

Fundamentals of Analytics

6 September 2020

A Study in Diabetes Detection

Introduction

In the past thirty years, there have been many advancements in the field of biology as a result of new technologies. Whether it is CRISPR genetic sequencing, Magnetic Resonance Imagery, or the breakthroughs in stem cell research, there has been an explosion in the wealth of data created within the biology field, which some have referred to as the phenomenon of "Big Data." Due to the complexity and abundance of data, scientists have been struggling to efficiently analyze "Big Data" with legacy analytical tools and statistical methods. This is why many data-oriented methods such as machine learning and artificial intelligence have risen in popularity. The application of such methods has been seen in the diagnosis of many life-threatening diseases, including that of diabetes[1].

Diabetes Mellitus is a metabolic disorder caused by defective insulin secretion. This insulin deficiency leads to higher levels of glucose in the blood and impaired carbohydrate metabolism. Diabetes Mellitus is divided into two main types, type 1 diabetes (T1DM) and type 2 diabetes (T2DM). T1DM is an autoimmune disease where the immune system attacks and destroys the cells that produce insulin in the pancreas. T2DM, the much more common form of diabetes that is responsible for 90% of diabetic patients, causes patients to have some level of insulin resistance that leads to higher blood sugar levels. The focus of this study will be using the risk factors of T2DM to predict and diagnose a diabetic individual[1].

There are many well-researched diabetes risk factors such as weight, dietary habits, and physical activity[2]. In this study, these risk factors will be considered "non-invasive" because the data can be collected without drawing blood or taking urine samples. There are also many risk factors for diabetes that can be measured through "invasive" measures such as blood glucose, high-density lipoproteins (HDL), and creatinine levels[5]. Research has also found serum selenium concentration[6] and urinary cadmium levels to be abnormally high in diabetic patients[7]. These risk factors that have direct correlation to diabetes will be the core features to this study's diabetes detection models.

What is novel about this study when compared to other diabetes detection models is the inclusion of features pertaining to mental health. Mental wellness is a hot issue in today's society, and there have been many recent research studies that have drawn connections between diabetes and poor mental health[3]. However, the use of mental wellness features has yet to be seen within the context of diabetes detection. The purpose of this study is to see whether mental health features are meaningful enough within predictive modeling such that they can improve model accuracy.

Motivation

Obesity has become a chronic issue in the United States and there have not been definite solutions to this problem. As a result, many diseases that a poor diet can be responsible for have become more prevalent in the past decade. Obesity is one of the leading risk factors for T2DM and it is to no surprise that the incidence for T2DM has increased as well. Especially in this current pandemic, most gyms and exercise clinics have been shut down for safety precautions. With most of the population stuck at home, routine daily physical activity is harder to come by for the individual. Furthermore, many restaurants have also closed down in the past months, forcing many to look for other unhealthier sources of nourishment such as fast food. A lack of exercise along with a poor diet is a very poor pairing within the context of a growing diabetes health dilemma in this country.

Depression has also become a major topic of conversation in the beginning of the 21st century. Especially among the population's youth, the worries of unstable mental health have become very apparent as more studies come out citing its negative short-term and long-term effects. The consequences of poor mental health include fatigue, lack of sleep, bad memory, loss of appetite etc[4]. Many cultures look down on mental wellness and view these ramifications as only a temporary problem that will go away with time. However, long-term depression within today's youth can grow to become a substantial thorn in society's future workforce. Even currently, when looking at this issue from a larger scope, mental wellness has already been tied with pressing concerns such as unemployment, negative social relationships, and drug abuse.

This study ties in the themes of obesity and depression through the development of a novel diabetes detection machine learning model. Even though many models used in clinical settings are very effective at detecting the onset of diabetes, the purpose of this study is to develop a model that could be used more for analytical purposes. Disease detection models are useful in that users can examine them in order to look at trends and make predictions for a given population. As huge trends like obesity and depression fluctuate and change in the future, this model will allow one to also predict shifts for diabetes as well in the United States. With the discovery of more risk factors of diabetes in the future, and as more streams of data become available, disease detection models will become ever more so effective.

Models such as these also serve a greater purpose of validating how beneficial medical data can be in the context of advancements in population health. Diabetes is but one disease that researchers are trying to combat with greater efficacy in this advent of artificial intelligence and big data. In the years to come, studies like these will hopefully aide in the acceptance of a data-driven culture so medical information is more accessible to researchers that can help benefit society's population health.

Literature Review

Google scholar was the main search engine used for literature research for this study. Most of the research done was based on machine learning applications within diabetes, although some research papers will be solely focused on diabetes risk factors. The core eight papers found and utilized were the following: *Machine Learning and Data Mining Methods in Diabetes Research* [1], *Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes*[2], *Diabetes Care and Research: What Should Be the Next Frontier?*[3], *Socially isolated individuals are more prone to newly diagnosed and prevalent type 2 diabetes mellitus – the Maastricht study –*[4], *Type 2 diabetes risk forecasting from EMR data using machine learning*[5], *Serum Selenium Concentrations and Diabetes in U.S. Adults: National Health and Nutrition Examination Survey (NHANES) 2003-2004*[6], *Urinary Cadmium, Impaired Fasting Glucose, and Diabetes in the NHANES III*[7], *Predicting Diabetes Mellitus With Machine Learning Techniques*[8]. The three most impactful papers to this study are reviewed in greater detail.

The first paper, *Machine Learning and Data Mining Methods in Diabetes Research* [1], serves as a useful introduction to preliminary literature research by reviewing the following concepts in this order: background machine learning knowledge, explanation of diabetes mellitus (DM), and machine learning applications within diabetes. The background machine learning knowledge covers both supervised and unsupervised modeling as well as other useful information like feature selection. The third section provides numerous papers on various machine learning method uses in DM research segmented between the following five major areas of application: biomarker identification and prediction of DM, diabetic complications, drugs and therapies, genetic background and environment, and health care management systems.

Of the five sections, this modeling study primarily coincides with the first application of prediction and diagnosis of DM. Currently, DM diagnosis is done through the A1C test, blood sugar test, or a fasting sugar test. It is crucial that DM is detected early on in the process in order to slow its progress, treat symptoms, and increase life expectancy. This section provides papers that use different machine learning approaches including decision trees, random forests, support vector machines, neural networks etc. Some authors even built disease progression models that could track the typical trajectory from hyperlipidemia to hypertension, impaired fasting glucose, and eventually T2DM. It also explains the use of ensemble approaches, which is the development of multiple machine learning algorithms to improve classification accuracy. An example of an ensemble approach is Ozcift and Gulten's combination of 30 different machine learning algorithms to create the Rotation Forest algorithm. Association rule mining has also been an avenue of research to discover different DM risk factors that might be correlated. This study provides a great introduction to machine learning applications in DM and relevant resources for further exploration of the prediction and diagnosis of T2DM[1].

The second paper, *Predicting Diabetes Mellitus With Machine Learning Techniques*[8], also provides a general introduction to DM and discusses the various machine learning applications within the field. The primary focus of the study is to build decision tree, random forest, and neural network models for the prediction and diagnosis of T2DM. The data set provided is collected from examination data in Luzhou, China and Pima Indians. The study made sure to correct for class imbalances when splitting the data into test and training sets. Furthermore, it uses principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce dimensionality. When validating models, the holdout method is compared to k-fold cross validation; since k-fold cross validation avoids the variance that can be seen in data, this study chooses a 5-fold cross validation. In terms of model evaluation, the study uses recall rate, precision, accuracy, and Mathew's Correlation Coefficient (MCC).

The paper also introduces each algorithm used and explains the concepts behind each model. A decision tree classifies instances by repeatedly dividing the total data set into partitions by features until each partition is as distinct as possible. The decision tree determines the feature that it splits upon at each node based on the amount of information gain of that feature. The decision tree algorithm this paper uses is the J48 tree from WEKA. The random forest classification algorithm uses multiple decision trees to perform either prediction or regression. Essentially the random forest creates multiple decision trees for each instance such that each tree is factored into the final classification of that instance. A neural network is a mathematical model based on the neural networks of a brain that adjusts the relationship between internal neurons (nodes) to make a prediction. The neural network is created with multiple layers of neurons that are divided between the following main layers: input, output, and hidden layers. The input layer handles the data, the hidden layers vary in number and are hidden from the user, while the output layer shows the results.

Overall, this study did not see much difference in performance between the three different algorithms. The best result for the Luzhou data set is 0.8084 compared to the best of the Pima data set being 0.7721; this again validates how data quality and feature availability are two important factors in model accuracy. The final note of the study is how it could not predict the exact type of DM, which is something that can be looked into in the future[8].

This third research paper, *Socially isolated individuals are more prone to newly diagnosed and prevalent type 2 diabetes mellitus – the Maastricht study –*[4], evaluates the relationship between social isolation and DM by assessing various structural and social aspects through multinomial regression. The data for this study is taken from the Maastricht Study, a cohort study that focuses on etiology, pathophysiology, and complications of T2DM. The dependent variable, glucose metabolism, is measured by a standardized glucose tolerance test after an overnight fast. Social networking features are collected in a questionnaire format, and many new features are calculated from this collected data. These calculated features include the number of unique network members, percentage of members the patient interacted with weekly, percentage of network that is family or friends etc. There are also general features collected in the questionnaire that include employment status, alcohol consumption, smoking status, and history of disease.

In terms of analysis, the study first compares the various participant groups (non-diabetic, pre-diabetic, and diabetic) through chi-square, ANOVA, and Kruskal-Wallis tests. This is followed by logistic regression analyses between each social network variable and diabetes status by finding the odds ratio (OR) and the 95% confidence interval for each variable. Each variable is judged separately and adjusted depending on age, BMI, employment status, education level, alcohol consumption, smoking status, hypertension, and disease history. Since a majority of the social network variables shows a significant interaction with sex at a 1% significance level, the analysis is also separated by gender. The results show that both men and women who are more socially isolated and receive less emotional support are more frequently diagnosed with T2DM.

Approach

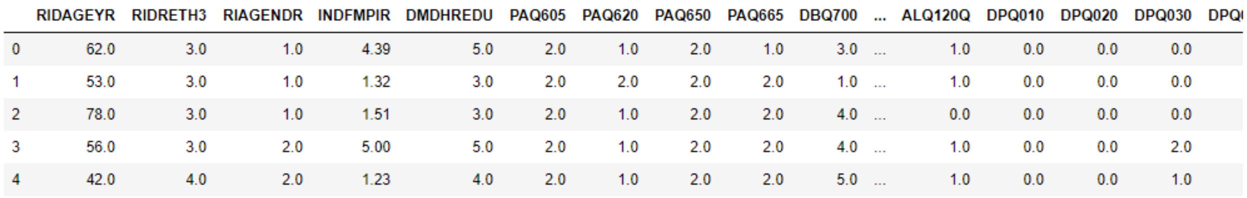
The data was extracted from the 2015-2016 National Health and Nutrition Examination Survey (NHANES) provided by the Center for Disease Control and Prevention (CDC). The NHANES data set contains 124 individual data sets that are partitioned between demographics, dietary, examination, laboratory, and questionnaire data. Since the risk factors for diabetes cover a wide range, there are features selected from all five of these partitions. The NHANES webpage has a variable search engine that was used to search and locate potential features. For example, when looking for variables related to smoking, CDC's search function generates all smoking-related variables, and the data sets they are contained in. Each data set contains all relevant information on how the data is collected as well as the individual variable metadata.

Since many features were from different data sets, usually the inclusion of a new feature involved merging an entirely new data set. Luckily, all of the data are within the NHANES survey such that every instance has a unique identifier that two different data sets can merge on. Consider a demographics data set within the NHANES data that contains 47 features detailing some common social statistics of the survey participants. Now consider a potential feature, participant weight, that belongs to a separate examination data set. The data set with the weight variable is imported into the same file as the demographics data and the two sets are then merged upon that unique instance identifier. This process was repeatedly done each time a potential new feature from a different data set needed to be explored.

Since features were gathered from a variety of data sets, null values started aggregating quickly when more features were merged into one data set. This was because not every participant in the NHANES survey completed all the questions or examinations across the 124 individual data sets. Since the diversity of the modeling features in this study ranges from invasive urine samples to mental health questions, the majority of the data needed belonged to separate data sets and needed to be merged. When cleaning the data, handling null values became a very real challenge since there was a tradeoff between including an important feature and losing many data instances. If too many instances with null values were removed, then little data was left over for training and testing. To compensate for this conflict, multiple training and testing sets were created with differing balance of features and data points and tested to determine which combination was best. Note that many null values could not be replaced with a mean or median value since most of the features are categorical in nature. As a result, features with many null values were generally not included and any instances with null values were removed from the final data set.

Before a potential feature can be merged with the final data set, the feature must go through a preliminary data exploration step to evaluate whether it is even usable. This feature selection process can be broken down into three distinct steps. Since most of the data is categorical, consider a potential categorical feature. Step one, the variable is visualized as a bar chart where the X axis contains the unique values of the feature and the Y axis is the number of instances for each unique value. This step allows for easy detection of obvious irregularities in the data that should be eliminated. Many of the questionnaire variables had answers such as "I don't know" or "I am unwilling to share this information," which are effectively null values. At this step, these useless values for a variable are removed since the machine learning algorithm should not relate this value to other values because some of these categorical variables were ordinal. The next step is to plot the potential feature against the dependent variable that describes whether a survey participant is diabetic. This second bar chart easily reflects whether the potential feature follows the trend proven in literature research. Furthermore, this second bar chart can also tell whether certain values of the variable should be grouped together to improve feature efficacy. For example, for one of the mental health features, people who answered A, B, or C all had a similar and very low chance of having diabetes in comparison to people who answered D. As a result, responses A, B, and C were relabeled as one distinct categorical value. The final step is to filter the variable due to reasons from literature research. For example, this study opts to remove diabetics that are 17 or younger because most cases of diabetes are above the age of 18. This three-step process details the selection, cleaning, and transformation process of any potential features. Note that statistical tests were not used in step two to evaluate the correlation between the potential feature and the dependent variable. This is because any non-novel feature that was used had already been well researched, and the purpose of this study was not to prove these trends wrong. Furthermore, mental health features were usually included as long as a trend existed, since the connection between diabetes and specific mental health features has not been firmly established as of yet.

Since there were many different potential features, each with many variables that might contain similar information, this process of finding the best feature and exploring its validity resulted in this data extraction process taking up a good portion of this study. This merging, transforming, cleaning, and visualizing process was repeated until a subset of features have been selected for the modeling step (cleaned data set screenshot below).



	RIDAGEYR	RIDRETH3	RIAGENDR	INDFMPIR	DMDHREDU	PAQ605	PAQ620	PAQ650	PAQ665	DBQ700	...	ALQ120Q	DPQ010	DPQ020	DPQ030	DPQ(
0	62.0	3.0	1.0	4.39	5.0	2.0	1.0	2.0	1.0	3.0	...	1.0	0.0	0.0	0.0	
1	53.0	3.0	1.0	1.32	3.0	2.0	2.0	2.0	2.0	1.0	...	1.0	0.0	0.0	0.0	
2	78.0	3.0	1.0	1.51	3.0	2.0	1.0	2.0	2.0	4.0	...	0.0	0.0	0.0	0.0	
3	56.0	3.0	2.0	5.00	5.0	2.0	1.0	2.0	2.0	4.0	...	1.0	0.0	0.0	2.0	
4	42.0	4.0	2.0	1.23	4.0	2.0	1.0	2.0	2.0	5.0	...	1.0	0.0	0.0	1.0	

Instead of testing a multitude of classification algorithms, the major focus of the modeling process was to select one algorithm and increase model performance through two key elements: feature selection and model tuning. Since the decision tree algorithm innately has a feature selection process, it is not always necessary to do an extra step of feature selection prior to model building. However, since potential features in this study were gathered from a variety of data sets, there were more null values as more features were merged into one data frame. As mentioned earlier, if too many features from different data sets are merged together and any instances with null values are removed, there may only be a little data left over for training and testing. As such, it was needed to limit the number of features dropped into the CART algorithm, so some preliminary feature exploration was conducted to remove the features that do not have blatant correlations with diabetes. The model tuning process will be detailed later on in the discussion section.

Since the decision tree is one of the most popular machine learning algorithms in the medical field, the chosen method is the Classification and Regression Tree (CART)[8]. Decision trees have a tree-based structure where data instances are divided based on model features. Starting with the entire data set of survey participants at the root of the tree, the decision tree algorithm splits the data set at each node by a designated feature. Each node will further partition individuals by features until the leaf nodes ideally have distinct groups of either diabetics or non-diabetics. The CART chooses features at each node based on which feature results in the greatest information gain. In other words, at each node of the decision tree, the CART chooses a feature that will best classify individuals in order to get partitions that consist of either diabetics or non-diabetics. As information gain increases, the resulting partitions become more “pure” with only one class of instances.

The primary goal of this study was to build a CART that used features not collected through invasive measures such as urinary or blood sample tests. The benefit of having features that do not need invasive methods is that analysts using this model can gather the needed data with greater ease. However, the downside to only using non-invasive features in the context of diabetes detection is that the performance of the model will likely be worse. To compensate for this, this study also aimed to build a second model that included invasive features to see the degree of improvement to overall model performance. The bulk of the project analysis will be detailing the process of tuning the non-invasive model, evaluating novel feature performance, and comparing the two models based on the following evaluation metrics.

For the purpose of evaluating this diabetes detection model, it was assumed that the model would not be used for analytical purposes, but rather used to detect whether an individual has diabetes. Three common metrics of evaluation are accuracy, precision, and recall rate. The accuracy of the models will be measured based on the percentage of people that are properly classified as diabetic or non-diabetic. Even though overall accuracy is important, the recall rate of the minority class is usually the main metric of importance in rare disease detection models. Recall rate is of such importance because it represents how well the model avoids classifying a true diabetic as a non-diabetic. The worst-case scenario is when a rare disease detection model predicts a sick patient as healthy. Precision is also a useful metric to see the proportion of the diabetic classifications that are correctly predicted. Even though recall rate is the critical measurement for making sure diabetic patients are identified, precision is more important when it comes to evaluating the monetary cost of a model. Improperly classified non-diabetics will need to go through further testing, which is a waste of expenses. The F1 score is a metric that ties precision and recall rate and provides a more balanced view on model efficacy. Accuracy, recall rate, precision, and F1 score will all be used to evaluate overall model performance throughout the non-invasive model tuning process as well as during the comparison process between the invasive and non-invasive model.

This entire study is coded in python within Jupyter Notebook, mainly utilizing the NumPy, pandas, scikit-learn, and Matplotlib frameworks. Pandas is the main package used to provide the DataFrame data structure for any data, as well as provides most of the tools needed to clean the raw data. Matplotlib is the tool used to visualize and explore individual features when selecting potential features for modeling. Sklearn is used for the entirety of the modeling process.

Results

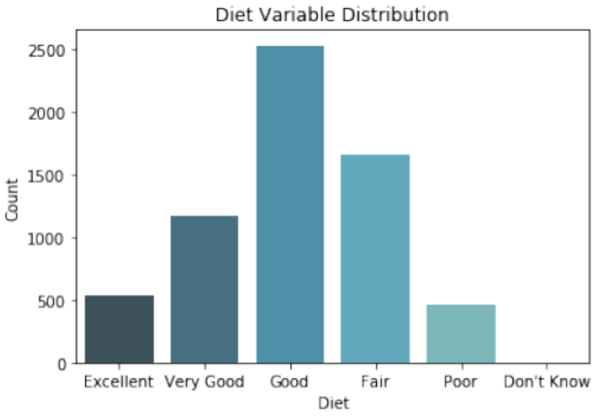


Figure 1

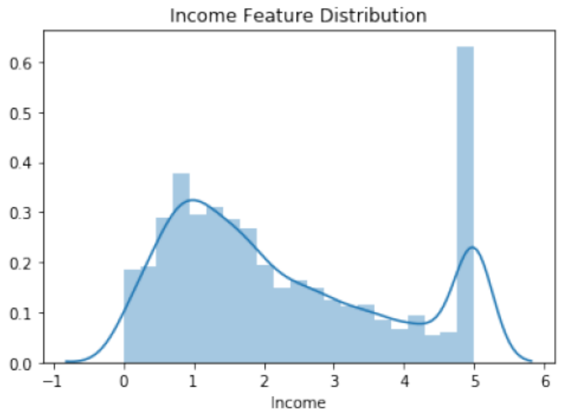


Figure 2

Figure 1 and 2 are the first exploratory analysis charts that visualize the overall distribution of the potential feature. Bar charts are used for categorical variables while histograms are used for numerical variables. For Figure 1, the "Don't Know" bar has a few instances that are effectively null values, which can be identified and removed after viewing this visualization. Figure 2 visualizes a spike of participants with a max income at the right tail of the distribution; a closer look into the metadata of the income variable shows that participants above a certain salary are all given a salary level of five.

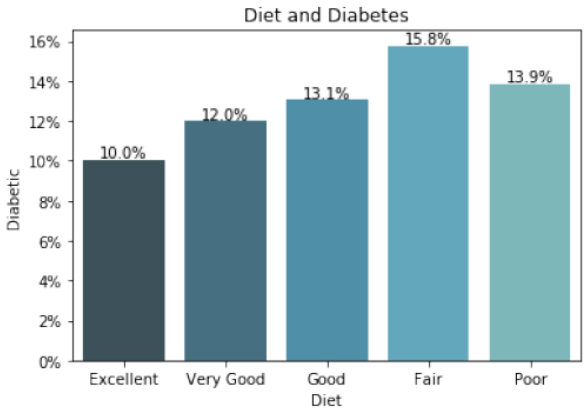


Figure 3

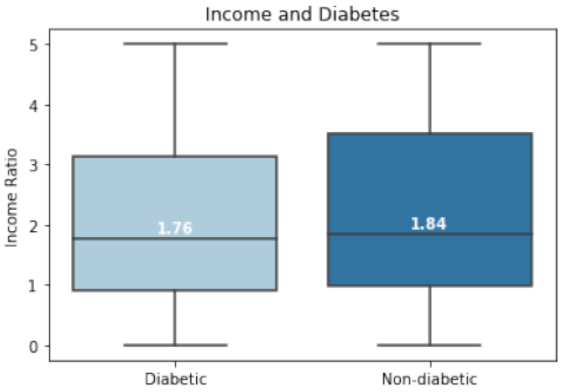
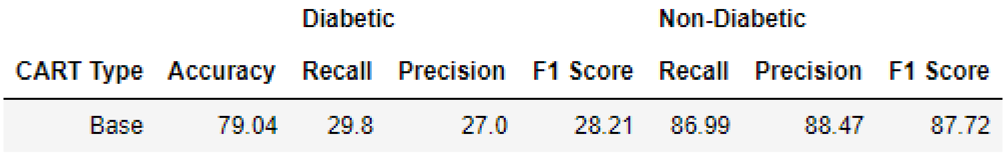


Figure 4

Figure 3 and 4 are the second exploratory analysis charts that visualize the correlation between potential features and the dependent variable of whether an individual is diabetic. Bar charts are again used for categorical variables while boxplots are used for numerical variables. If the correlation between the potential feature and the dependent variable does not align with research, this variable is discarded. This is also the step where individual values for a categorical feature are binned together if it is more reasonable.

Sklearn has a good implementation of CART that is used to build the first non-invasive model. Training and testing splits were created with 80/20 splits, respectively, and used to train

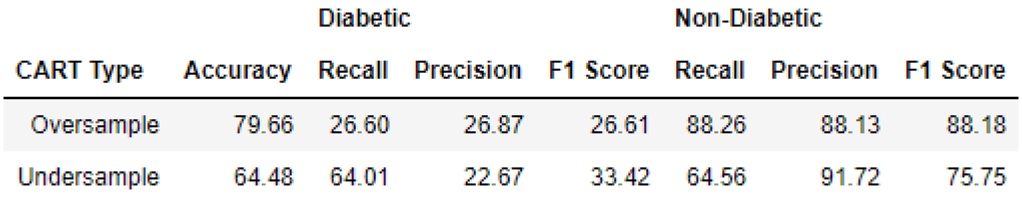
1000 CART models. The following thirteen non-invasive features were included: age, race, gender, income, education, physical activity, diet, BMI, waist circumference, hypertension, smoke, alcohol, and mental health. The average accuracy, recall rate, precision, and F1 score were taken across 1000 runs and shown in Figure 5.



		Diabetic			Non-Diabetic		
CART Type	Accuracy	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Base	79.04	29.8	27.0	28.21	86.99	88.47	87.72

Figure 5. Base CART Model

Even though model accuracy is almost 80%, the focus is on the recall rate of the diabetic class, which is unfortunately very low at 29.8%. Meanwhile, the recall rate and precision are both much higher for the non-diabetics at close to 90% each. Note that this disparity is caused by the huge class imbalance between diabetics and non-diabetics in the data, where there are 490 diabetic instances and 3030 non-diabetic instances. In order to account for class imbalances, oversampling and undersampling techniques were used to build two more CART models. The CART algorithm from Sklearn was encapsulated in an additional method that adjusts the data sets with these data resampling techniques. Note that oversampling and undersampling are integrated before the training step of the model.

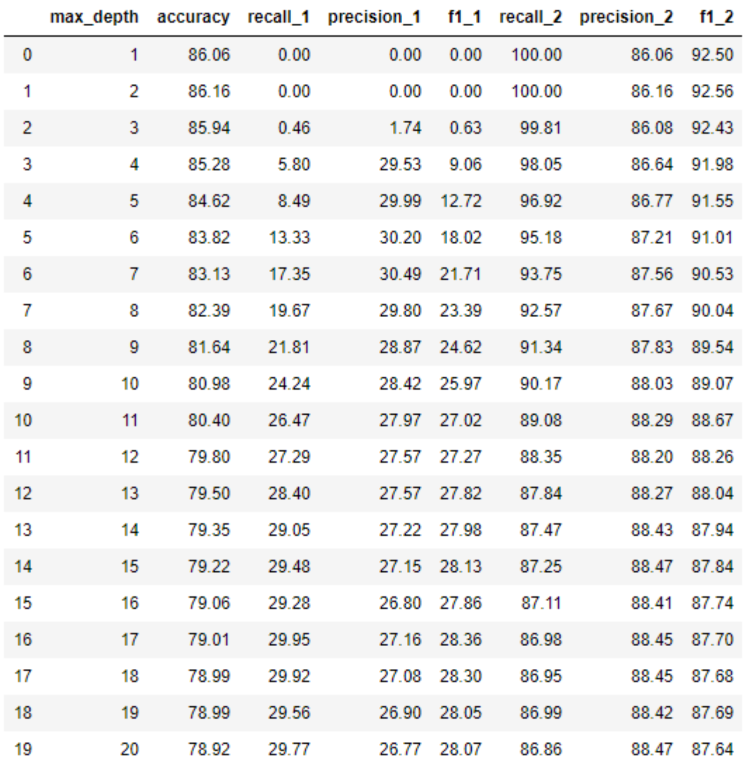


		Diabetic			Non-Diabetic		
CART Type	Accuracy	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Oversample	79.66	26.60	26.87	26.61	88.26	88.13	88.18
Undersample	64.48	64.01	22.67	33.42	64.56	91.72	75.75

Figure 6. Oversample and Undersample CART Models

After oversampling, model accuracy remained the same at about 79%, while the recall rate of diabetics decreased by 3% to 26.6%. After undersampling, model accuracy decreased 15% to 64.48%, but recall rate of diabetics improved by 34% to 64.01%. The F1 scores for the diabetic class were 26.6% and 33.42% for the oversample and undersample models respectively, with the latter being slightly higher with its increase in recall rate.

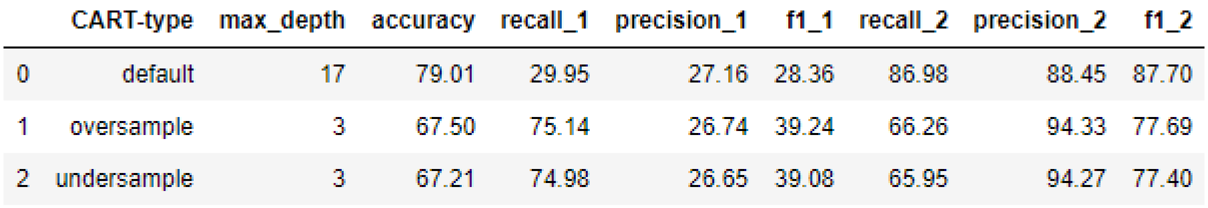
The CART models were then optimized in a second manner by tuning decision tree hyperparameters. This study focused on adjusting the maximum depth of the tree to increase the recall rate for the diabetic class. A method was created to adjust the maximum depth for all three decision trees to find the optimal depth. This method builds 1000 CART models at each depth and returns the average evaluation metrics used previously. This simply approach was chosen over a hill-climbing algorithm variation because some simple testing revealed that the optimal recall rate for all three model types appeared between max depths of 1 through 20. The output for the base CART model is shown in Figure 7, and the outputs for the oversampled and undersampled trees can be found in the report code.



	max_depth	accuracy	recall_1	precision_1	f1_1	recall_2	precision_2	f1_2
0	1	86.06	0.00	0.00	0.00	100.00	86.06	92.50
1	2	86.16	0.00	0.00	0.00	100.00	86.16	92.56
2	3	85.94	0.46	1.74	0.63	99.81	86.08	92.43
3	4	85.28	5.80	29.53	9.06	98.05	86.64	91.98
4	5	84.62	8.49	29.99	12.72	96.92	86.77	91.55
5	6	83.82	13.33	30.20	18.02	95.18	87.21	91.01
6	7	83.13	17.35	30.49	21.71	93.75	87.56	90.53
7	8	82.39	19.67	29.80	23.39	92.57	87.67	90.04
8	9	81.64	21.81	28.87	24.62	91.34	87.83	89.54
9	10	80.98	24.24	28.42	25.97	90.17	88.03	89.07
10	11	80.40	26.47	27.97	27.02	89.08	88.29	88.67
11	12	79.80	27.29	27.57	27.27	88.35	88.20	88.26
12	13	79.50	28.40	27.57	27.82	87.84	88.27	88.04
13	14	79.35	29.05	27.22	27.98	87.47	88.43	87.94
14	15	79.22	29.48	27.15	28.13	87.25	88.47	87.84
15	16	79.06	29.28	26.80	27.86	87.11	88.41	87.74
16	17	79.01	29.95	27.16	28.36	86.98	88.45	87.70
17	18	78.99	29.92	27.08	28.30	86.95	88.45	87.68
18	19	78.99	29.56	26.90	28.05	86.99	88.42	87.69
19	20	78.92	29.77	26.77	28.07	86.86	88.47	87.64

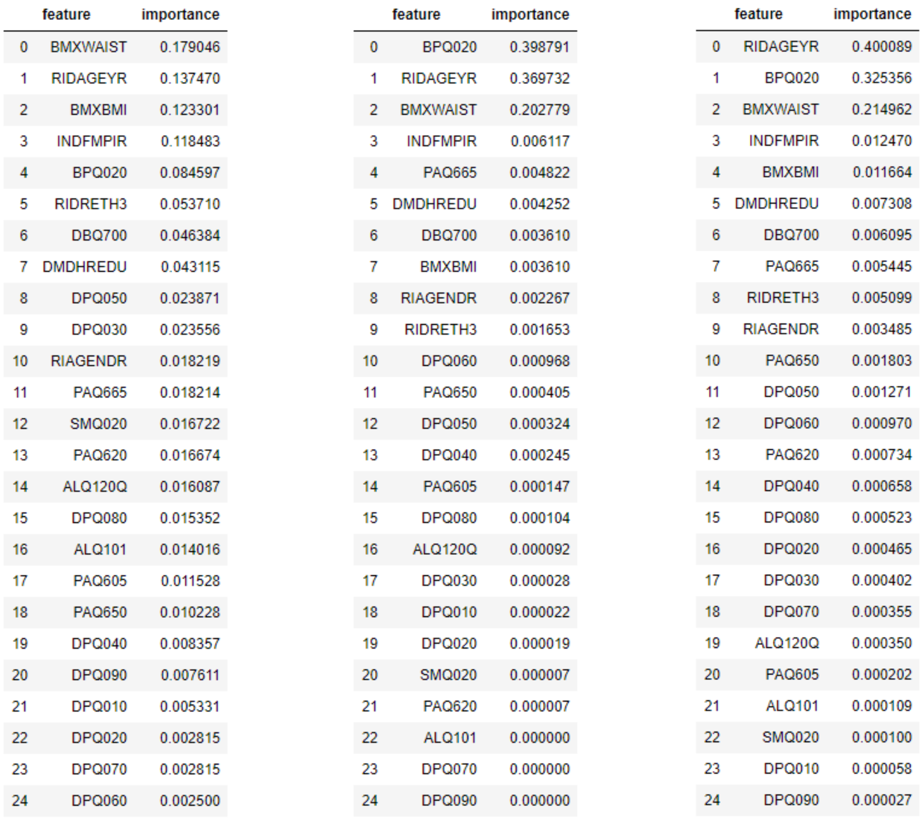
Figure 7. Base CART Model Maximum Depths

Of the base CART trees with varying specified maximum depth, the tree with the highest recall rate is one of maximum depth of 17 with a recall rate of 29.95% for the diabetic class and accuracy of 79.01%. Of the oversampled CART trees, the tree with the highest recall rate for the diabetic class is one of maximum depth of three with a recall rate of 75.14% and accuracy of 67.50%. Of the undersampled CART trees with varying specified maximum depth, the tree with the highest recall rate is one of maximum depth of one with a recall rate of 76.05% and accuracy of 62.37%. Since the overall accuracy is not optimal, this tree is compared to others to see if there is a model with a higher accuracy without much recall rate loss. The undersampled CART with a maximum depth of a three has an increase in accuracy from 62.37% to 67.51% with only a decrease in recall rate from 76.05% to 74.38%. The three final non-invasive CART models are shown in Figure 8 below. Note that the F1 scores for the diabetic class are 28.36%, 39.24%, and 39.08% for the three models, which are improvements to the pre-tuned models.



	CART-type	max_depth	accuracy	recall_1	precision_1	f1_1	recall_2	precision_2	f1_2
0	default	17	79.01	29.95	27.16	28.36	86.98	88.45	87.70
1	oversample	3	67.50	75.14	26.74	39.24	66.26	94.33	77.69
2	undersample	3	67.21	74.98	26.65	39.08	65.95	94.27	77.40

Figure 8. Final Non-Invasive CART Models



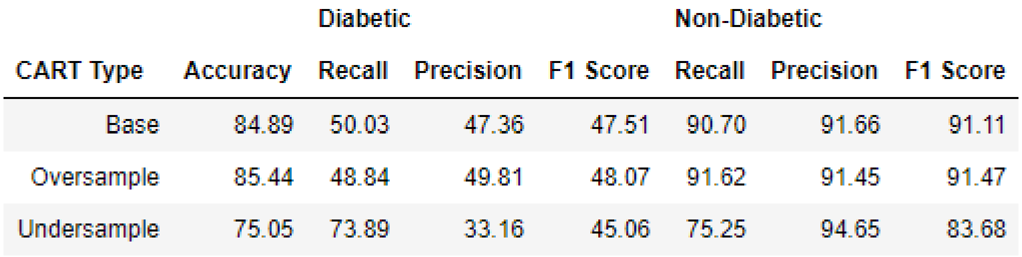
	feature	importance
0	BMXWAIST	0.179046
1	RIDAGEYR	0.137470
2	BMXBMI	0.123301
3	INDFMPIR	0.118483
4	BPQ020	0.084597
5	RIDRETH3	0.053710
6	DBQ700	0.046384
7	DMDHREDU	0.043115
8	DPQ050	0.023871
9	DPQ030	0.023556
10	RIAGENDR	0.018219
11	PAQ665	0.018214
12	SMQ020	0.016722
13	PAQ620	0.016674
14	ALQ120Q	0.016087
15	DPQ080	0.015352
16	ALQ101	0.014016
17	PAQ605	0.011528
18	PAQ650	0.010228
19	DPQ040	0.008357
20	DPQ090	0.007611
21	DPQ010	0.005331
22	DPQ020	0.002815
23	DPQ070	0.002815
24	DPQ060	0.002500

	feature	importance
0	BPQ020	0.398791
1	RIDAGEYR	0.369732
2	BMXWAIST	0.202779
3	INDFMPIR	0.006117
4	PAQ665	0.004822
5	DMDHREDU	0.004252
6	DBQ700	0.003610
7	BMXBMI	0.003610
8	RIAGENDR	0.002267
9	RIDRETH3	0.001653
10	DPQ060	0.000968
11	PAQ650	0.000405
12	DPQ050	0.000324
13	DPQ040	0.000245
14	PAQ605	0.000147
15	DPQ080	0.000104
16	ALQ120Q	0.000092
17	DPQ030	0.000028
18	DPQ010	0.000022
19	DPQ020	0.000019
20	SMQ020	0.000007
21	PAQ620	0.000007
22	ALQ101	0.000000
23	DPQ070	0.000000
24	DPQ090	0.000000

	feature	importance
0	RIDAGEYR	0.400089
1	BPQ020	0.325356
2	BMXWAIST	0.214962
3	INDFMPIR	0.012470
4	BMXBMI	0.011664
5	DMDHREDU	0.007308
6	DBQ700	0.006095
7	PAQ665	0.005445
8	RIDRETH3	0.005099
9	RIAGENDR	0.003485
10	PAQ650	0.001803
11	DPQ050	0.001271
12	DPQ060	0.000970
13	PAQ620	0.000734
14	DPQ040	0.000658
15	DPQ080	0.000523
16	DPQ020	0.000465
17	DPQ030	0.000402
18	DPQ070	0.000355
19	ALQ120Q	0.000350
20	PAQ605	0.000202
21	ALQ101	0.000109
22	SMQ020	0.000100
23	DPQ010	0.000058
24	DPQ090	0.000027

Figure 9. Feature Importance (Base, Oversample, Undersample)

Figure 9 shows the feature importance for all features used in the base (left), oversample (middle), and undersample (right) CART models. The features that were one of the top three contributors across all three models are waist circumference (BMXWAIST), age (RIDAGEYR), BMI (BMXBMI), and hypertension (BPQ020). Note that all the novel mental health features begin with the letters DPQ, which will be further analyzed in the discussion.



		Diabetic			Non-Diabetic		
CART Type	Accuracy	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Base	84.89	50.03	47.36	47.51	90.70	91.66	91.11
Oversample	85.44	48.84	49.81	48.07	91.62	91.45	91.47
Undersample	75.05	73.89	33.16	45.06	75.25	94.65	83.68

Figure 10. Invasive Models

After building and tuning non-invasive models, this study also briefly explored the improvement in model performance with the inclusion of invasive features. The invasive model combined the 13 non-invasive features from the non-invasive model with the following 8 invasive features: glycohemoglobin, blood pressure, glucose, creatinine, high-density lipoprotein, triglycerides, selenium, cadmium. Unsurprisingly, every evaluation metric improved across all three models.

Discussion

Since the data exploration process was explained in the study approach, this discussion section will focus on the model building and tuning analysis. After building the initial base CART model, the immediate concern was that the recall rate for the diabetic class was remarkably low because decision trees heavily prioritize splitting on information gain based on the majority class, which is non-diabetics in this case. This trend seems to hold true for the first default CART model since the recall rate for the majority class, the non-diabetics, is high at 87% even though the recall rate for the diabetic class is 29.8% (Figure 5). As such, the 6:1 class imbalance between diabetics and non-diabetics in the data was assumedly causing the recall rate to be low for the diabetic class.

This class imbalance within the dataset was fixed in two ways, by oversampling the minority class and undersampling the majority class. The oversampling method introduces duplicate instances of existing instances in the training data until the number of diabetic instances and non-diabetics are equal in the training set. On the other hand, the undersampling method randomly samples a subset of the majority class from the training set such that the subset is equivalent in size to the minority class. Both methods make sure that the CART algorithm does not prioritize one class over the other during training.

Oversampling the diabetic class did not end up improving model accuracy and ended up decreasing the recall rate of the diabetic class by 3%. Replicating the same instances of diabetics within the training set seems to be ineffective in providing more information to the CART algorithm. Meanwhile, undersampling the non-diabetic class improved the recall rate of diabetics by 34%. Even though oversampling creates a larger dataset, a smaller dataset with unique diabetic instances from undersampling seems to be much more impactful at impr oving the recall rate. Unfortunately, the undersampling method also decreases the accuracy of the model by 15%. Between the two data correcting techniques, the undersampling method holds some improvement over the base model when it comes to identifying dia betic patients, at the cost of overall model accuracy (Figure 6).

After trying out various corrected datasets, the next step to the model tuning process was tuning the hyperparameters of the CART model, specifically maximum tree depth. The final CART models have maximum depths of 17, 3, and 3 for the base, oversample, and undersample models, respectively. Even though the base model had the highest accuracy of 79.01% of the three, the recall rate for the diabetic class was 29.95%. Meanwhile, the recall rate of the oversample and undersample models were about 75% with an accuracy of 67% each. It seems that regardless of the data correction technique used, further tuning the CART algorithm by maximum depth is enough to improve the recall rate by 45% while only losing less than 2% in model accuracy. This new insight reveals that oversampling is still effective in correcting data imbalances as long as model hyperparameters are adjusted accordingly (Figure 8).

Even though most of the model tuning process was dictated by improving the recall rate of the diabetic class, the overall F1 score of the final models should be addressed as well. Since the data has very skewed class imbalances, the minority diabetic class has a high chance of having a low recall or precision rate from the CART algorithm prioritizing the majority non-diabetic class; this is reflected in the default CART tree with a F1 score of 28.36%. Using the oversampling and undersampling methods along with maximum depth tuning, the F1 score improved 10% to just under 40%. Even though recall rates for the diabetic class improved by 45%, the precision rate even decreased by 1% in the oversample and undersample CART models. This discrepancy between the recall and precision rates is responsible for the low F1 score. It is hard to overcome such a class imbalance and improve both recall rate and precision without collecting better data (Figure 8).

After settling on the final three models, analysis on feature importance was critical to see whether the novel features have an impact on model performance. Across the three CART models, the novel depression features were not as important as the proven features such as hypertension, age, waist circumference, BMI, etc. This comes as no surprise since there has been extensive research done for the well-known risk factors of diabetes. However, it was interesting to see that out of the 24 features in the tuned base CART model, two mental health features (DPQ050 and DPQ030) were the 9th and 10th most impactful. DPQ050 explains whether an individual has poor appetite while DPQ030 explains whether an individual has trouble sleeping. Even though these two features were in the top 10 most impactful features, their degree of importance is not to the extent as some of previous features. The top 5 features contributed to about 63% of the model, with the 2 mental health features contributing to only about 5% of the remaining 37%. Note that these two features were still twice as important as some well-known risk factors such as smoking, alcohol, and physical activity. Even though these results do not immediately indicate poor mental health as a top 10 risk factor for diabetes, it does show that some mental health symptoms can be useful in predicting the onset of diabetes when used in conjunction with other well-known risk factors (Figure 9).

Looking at the final oversample and undersample models with a tree depth of 3, the feature importance distribution was very different from the tuned base model with a tree depth of 17. The top 3 features in the oversample and undersample methods contributed to about 97% and 94% of the respective models. The maximum tree depth of three really forced the feature distribution to be very top heavy since the CART algorithm now had to select the most important features and divide the instances to the most pure partitions possible within a limited number of splits. This skewed feature distribution does not mean that the other less impactful features are also less significant when it comes to predicting diabetes; they are just not as important as the primary risk factors when the algorithm is limited by hyperparameter constraints (Figure 9).

The secondary focus of this study was to build an invasive diabetes detection model. Compared to the default non-invasive model, the default invasive model's recall rate for the diabetic class improved from 29.8% to 50.03% for the diabetic class while accuracy improved from 79.04% to 84.89%. The oversample invasive model's recall rate improved from 26.6% to 48.84% while accuracy improved from 79.66% to 85.44%. The undersample invasive model has improved the recall rate from 64.01% to 73.89% while accuracy improved from 64.48% to 75.05%. In all three models, the inclusion of invasive features has improved overall recall rate for the diabetic class as well as overall model accuracy. This makes sense since the invasive features include that of fasting glucose and glycohemoglobin, which are the most important measurements taken when diagnosing cases of diabetes (Figure 10).

Conclusion and Future Work

The results from modeling were pretty expected when it came to the use of non-invasive and invasive features with the latter performing much better without model tuning. In future work, this study would like to optimize the invasive model through hyperparameter tuning. The hyperparameter focus for this study was maximum depth, but many other hyperparameters such as split feature, minimum instances in a leaf node, and max features could further improve model performance. Since extracting, cleaning, exploring, and balancing the data within Python took a good portion of the project time frame, time spent on model tuning was less than expected.

One of the biggest obstacles in this study was also the data quality, specifically referring to the quantity of diabetics. Since many features were included from varying data sets, many data points were lost in the process of cleaning out null values. There were many other data balancing methods that could fix this lack of data such as synthetic minority over-sampling (SMOTE). SMOTE essentially generates new instances in the minority class by creating instances that should be similar to the ones already existing in the class. By creating unique instances, this method holds promise compared to the oversampling method that was used, which only replicated existing data.

Even though this study was focused on building a decision tree and improving it through feature selection, adjusting for class imbalance, and tuning hyperparameters, the exploration of other models is also a natural course of progression. Notable models that would have yielded interesting results would be support vector machines and neural networks due to the common use of these models in the medical field. SVMs are great for classifying a binary class problem such as predicting diabetic and non-diabetic instances. Neural networks have been a very popular modeling technique and are praised for their high model performance within industry applications. Some of the literature research used in this study also used SVMs and NNs, which would be a great source for future guidance and comparison as well.

When looking at the performance of mental health features, it was a pleasant surprise to see two features contributing as the 9th and 10th best out of the 24. However, the effects of poor mental health cannot be summed up in only nine features; there were many variables in the NHANES data set that shed light on the social interactions of individuals, which could have been used as potential mental health features. As mentioned earlier, neural networks would be a great modeling choice here with the increase in number of features. However, this approach is again stymied with the lack of data, which ties in with the earlier discussion of new data augmentation techniques. Overall, it seems that no matter what path is taken to continue this study, the data collection and model tuning processes are intertwined and necessitate careful thinking to build a more useful and powerful model.

Citations

1. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*. 2017;15:104-116. https://www.sciencedirect.com/science/article/pii/S2001037016300733

2. Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Making*. 2010;10(16). https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-16#citeas

3. Marrero DG. Diabetes Care and Research: What Should Be the Next Frontier?. *Diabetes Spectr*. 2016;29(1):54–57. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4755457/

4. Brinkhues S, Dukers-Muijrers NHTM, Hoebe CJPA, et al. Socially isolated individuals are more prone to newly diagnosed and prevalent type 2 diabetes mellitus – the Maastricht study –. *BMC Public Health*. 2017;17(955). https://bmcpublichealth.biomedcentral.com/articles/%2010.1186/s12889-017-4948-6#citeas

5. Mani S, Chen Y, Elasy T, et al. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA Annu Symp Proc*. 2012;2012:606–615. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540444/

6. Laclaustra M, Navas-Acien A, Stranges S, et al. Serum Selenium Concentrations and Diabetes in U.S. Adults: National Health and Nutrition Examinatino Survey (NHANES) 2003-2004. *Environmental Health Perspectives.* 2009;117(9). https://ehp.niehs.nih.gov/doi/full/10.1289/ehp.0900704

7. Schwartz GG, Il'yasova D, Ivanova A. Urinary Cadmium, Impaired Fasting Glucose, and Diabetes in the NHANES III. *Diabetes Care*. 2003;26(2).

8. Zou Q, Qu K, Luo Y. Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*. 2019;9(1664-8021):515. https://www.frontiersin.org/article/10.3389/fgene.2018.00515