






# PREDICTING PREDIABETES AND DIABETES WITH MACHINE LEARNING

**WHO'S AT RISK**  
for prediabetes or type 2 diabetes?

You could have prediabetes or type 2 diabetes and not know it—there often aren't any symptoms. That's why it makes sense to know the risk factors:

-  45+ years old
-  Physically active less than 3 times/week
-  Family history of type 2 diabetes
-  High blood pressure
-  History of gestational diabetes\*
-  Overweight

\*Diabetes during pregnancy. Giving birth to a baby weighing 9+ pounds is also a risk factor.

## DID YOU KNOW...

African Americans, Hispanic/Latino Americans, American Indians/Alaska Natives, Pacific Islanders, and some Asian Americans are at higher risk.



**If you have any of the risk factors, ask your doctor about getting your blood sugar tested.**

Tracy A. Cardwell

[tracy.a.cardwell@gmail.com](mailto:tracy.a.cardwell@gmail.com)

<https://www.linkedin.com/in/tacardwell/>

# CONTENTS

Executive Summary	3
Introduction	4
Data	4
Data Wrangling	5
Data Exploration	5
Statistical Analysis	7
Summary of Statistical Tests	8
Machine Learning Approach	9
Preprocessing	9
Part 1	10
Part 2	11
Conclusion	14
Future Work	15
References	16

## EXECUTIVE SUMMARY

Diabetes is the 7<sup>th</sup> leading killer in the U.S. and the leading cause of end-stage kidney disease and new cases of adult blindness. The Centers for Disease Control and Prevention estimate about a quarter of those with diabetes are undiagnosed. One third of U.S. adults have prediabetes but only about one in ten knows they have it.

Machine learning models were developed to predict prediabetes and diabetes using the National Health and Nutrition Examination Survey (NHANES). The goal was to identify new risk factors to improve screening.

The developed models predicted prediabetes with 60% precision and diabetes with 80% precision.

The only novel risk factor identified was osmolality in the prediction of prediabetes.

# INTRODUCTION

The Centers for Disease Control and Prevention (CDC) estimates that 10.5% of the U.S. population had diabetes as of 2018 and about a quarter of those were undiagnosed. Just over a third of U.S. adults had prediabetes and only 15.3% of them had been told by their doctor.<sup>1</sup> That leaves over 8 out of 10 people with prediabetes in the dark!

The vast majority (over 90%) of these diabetes cases are type 2 diabetes, in which the body becomes resistant to insulin and eventually can no longer regulate blood sugar levels. Elevated blood sugar, over time, causes increased risk of heart attack, stroke, kidney failure, blindness, amputation of lower extremities and nerve damage. Diabetes is the 7<sup>th</sup> leading killer in the U.S. as well as the leading cause of blindness in working-age adults, amputation, and end-stage kidney disease.<sup>1,2</sup>

New research is showing that damage to body systems begins with prediabetes. There are huge benefits to the population and to the health care system to find ways to increase the number of people with prediabetes and diabetes who become aware of their condition and start treatment.

The estimated cost of *diagnosed* diabetes in 2017 is \$327 billion, including \$237 billion in direct medical costs and \$90 billion in reduced productivity.<sup>3</sup> Actual costs are most likely significantly higher as only about 75% of those with diabetes were diagnosed.

Machine learning is being increasingly used in health care to aid with diagnosis and treatment of many conditions. Machine learning tools can help health care providers predict who is at high risk for prediabetes or diabetes and allow for earlier detection. It can also be used to tailor disease management plans to individuals to improve outcomes. Machine learning can be especially helpful in areas where demand for healthcare resources outstrips availability as it can allow scarce resources to be leveraged.

The goal of this project is to use machine language in Python to predict prediabetes and diabetes using the National Health and Nutrition Examination Survey (NHANES). The models created will be evaluated to determine whether additional factors can be identified to improve existing screening tools.

## DATA

This study uses NHANES data collected between 2007 and 2016, which are the ten most recent years for which data is available. NHANES comprises demographic, socioeconomic, dietary and health related information, as well as physical examination measurements and laboratory test results. NHANES is a long-term health survey conducted by the National Center for Health Statistics, which is a part of the CDC. The survey collects information from about 5,000 persons each year. Ten years of data yielded observations for just over 50,000 people.

The data was downloaded from the CDC NHANES website at [wwwn.cdc.gov/nchs/nhanes/default.aspx](https://www.cdc.gov/nchs/nhanes/default.aspx).

NHANES oversamples persons over 60 years of age, African Americans and Hispanics to produce reliable statistics for these subsamples.

## DATA WRANGLING

The NHANES survey is not static. Survey data and questions are reviewed and changes made every two years to address new health topics. Data elements and categories are added, removed, and sometimes moved from one file to another. For this analysis, data elements were limited to those available during the entire ten-year span.

NHANES data are made available in small files, each containing data relating to one topic for one two-year survey cycle. 314 files were downloaded for this analysis. The files are SAS transport format.

Data for survey cycles were combined by topic. Several required special handling due to elements being moved during the ten-year period. String data also required special handling as the Python package used to convert SAS transport format to CSV format did not wrap strings in quotes, causing unpredictable results. Null heatmaps were generated to visualize missing data. Columns that did not span the entire ten years were dropped. Data was combined into six categories and saved in CSV format. The categories are nearly identical to those on the NHANES website with one addition:

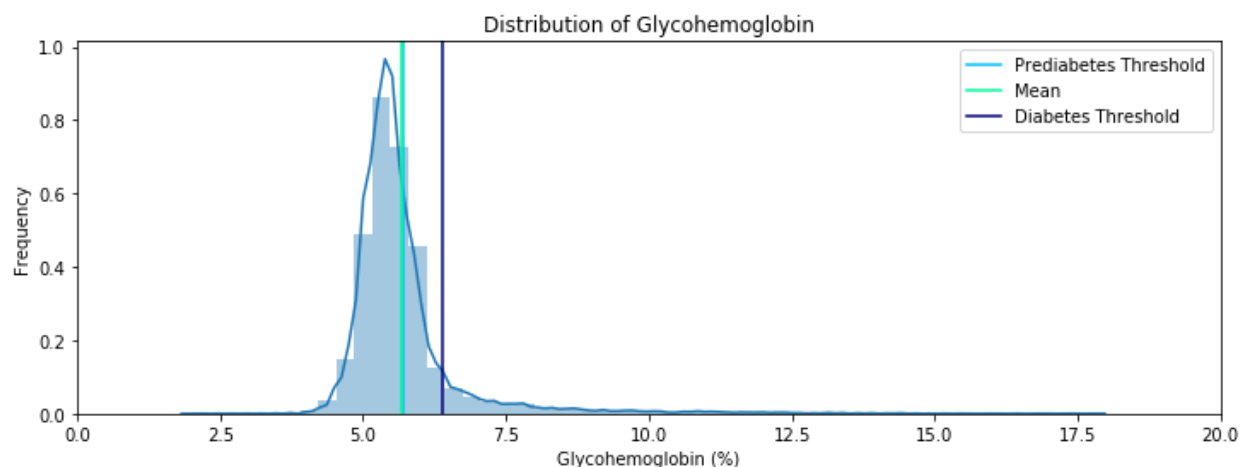
- Demographics: demo.csv
- Dietary intake: diet.csv
- Physical examination: exam.csv
- Laboratory test results: labs.csv
- Questionnaire answers: ques.csv
- Prescription medications (from Questionnaire category): meds.csv

The prescription data was split from the questionnaire data as there is one row for each prescription medication taken by an individual. Separating it simplified processing.

## DATA EXPLORATION

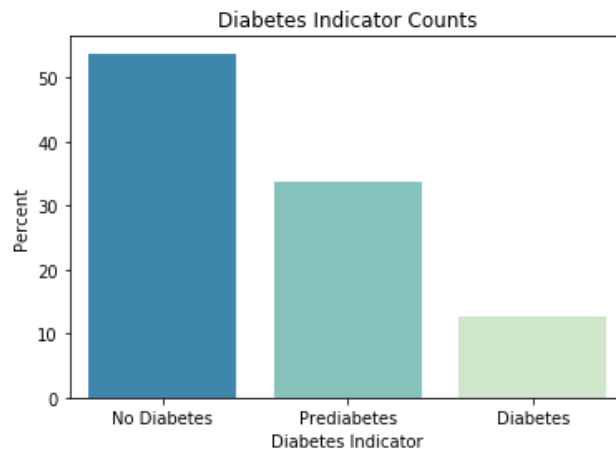
Initial exploration was targeted at statistics and known risk factors for prediabetes and diabetes. Would the study data reflect the statistics and trends reported by the American Diabetes Association (ADA)?<sup>4</sup>

The graph below shows the distribution of glycohemoglobin or HbA1c, a lab value used to diagnose prediabetes and diabetes. Interestingly, the mean is equal to the threshold for diagnosing prediabetes!

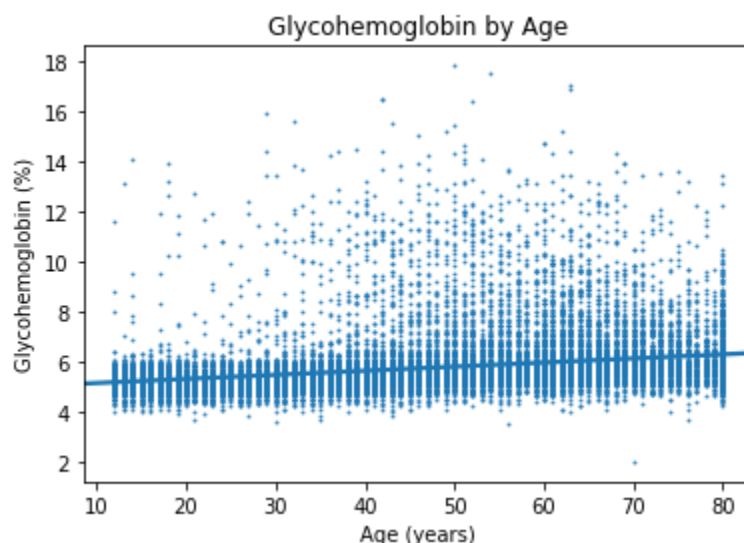


Also of note is the right skew of many of the lab results, with long right tails. This pattern is clearly illustrated in the glycohemoglobin plot. These outliers are very high lab test results and are valid values.

Below is the distribution of diabetes categories in the study data. Lab values used to diagnose prediabetes and diabetes were evaluated according to ADA diagnostic criteria to categorize study participants.<sup>5</sup> The rate of prediabetes was very close to that reported by the ADA. The rate of diabetes was a bit higher than the reported rate, most likely due to NHANES oversampling of higher risk categories (over 60, African American, Hispanic). This chart shows unweighted proportions so does not adjust for oversampling.

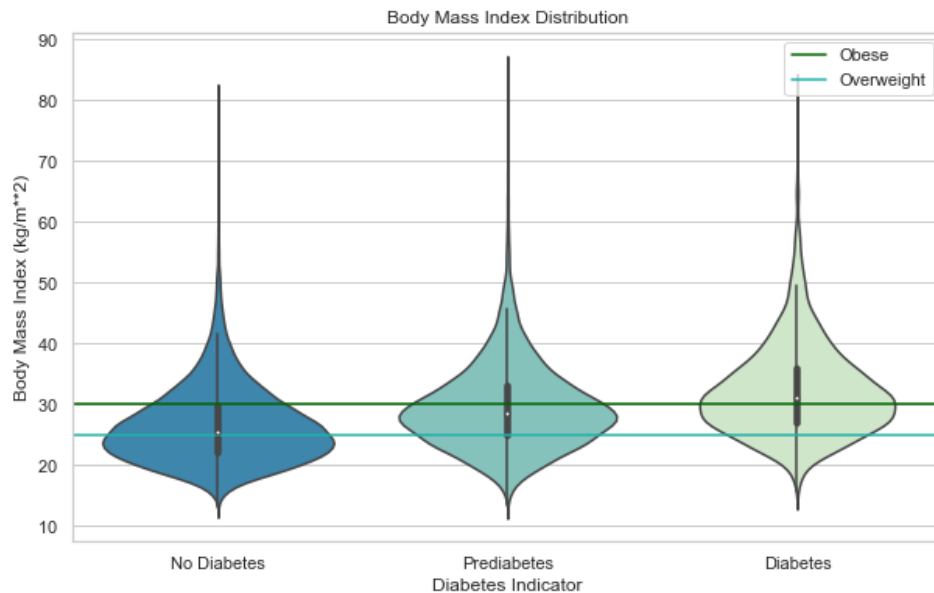


Below is a chart showing the relationship between glycohemoglobin and age. There is a clear trend upward with increasing age, which is in sync with the increased risk of diabetes. Type 2 diabetes was originally called adult-onset diabetes but is now frequently being diagnosed in teens and children. The chart below includes ages 12-80 and shows several teens well into diabetic range.



Finally, the following chart shows the distribution of Body Mass Index (BMI) for the three study categories. There is a clear trend toward higher BMI as we progress from healthy blood sugar to diabetic

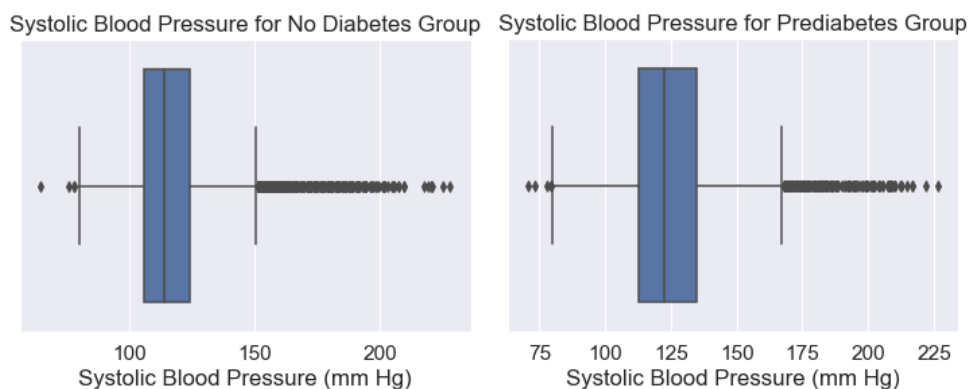
range. Note that the median BMI of the diabetes category is just above the obese BMI threshold. Even the non-diabetic category median is at the overweight threshold.



## STATISTICAL ANALYSIS

Various statistical tests were run to determine whether relationships exist in the data between the study categories and known risk factors as well as some factors not recognized as risk factors. All tests resulted in finding for a statistically significant relationship between the tested factor and study categories. It should be noted that due to the complex design of NHANES, adjustments to standard statistical tests are recommended. Python statistics packages do not have the capability to account for such complex survey designs so these statistical analyses may report p-values that are smaller than would be calculated were we able to account for the survey design.

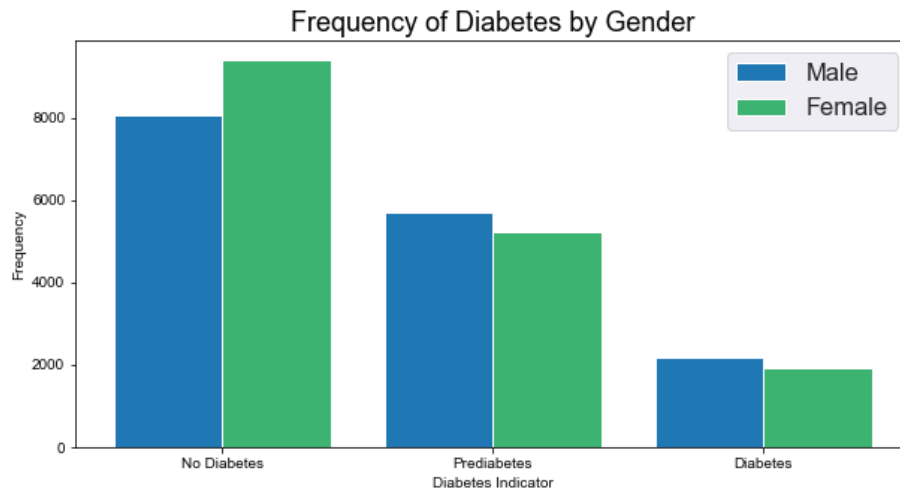
A student's t-test was run comparing the mean values of systolic blood pressure for the no diabetes and prediabetes groups. High blood pressure is a known risk factor for diabetes. The null hypothesis was no difference between the means, the alternative a statistically significant difference.



The t-statistic was -40, the p-value 0.0. The null hypothesis was rejected and we conclude that systolic blood pressure means are statistically different between non-diabetic and prediabetic categories.

A chi-squared test for independence was run on gender and diabetes category. The null hypothesis was no relationship, the alternative a statistically significant relationship. Gender is not a known risk factor for diabetes, although some studies show different risk of complications and different rates of diabetes.<sup>8</sup>

The  $\chi^2$  value was 131 with a p-value near zero ( $10^{-29}$ ). This gave evidence against the null hypothesis, and we conclude that gender and diabetes category are related.



## SUMMARY OF STATISTICAL TESTS

Chi-square Tests	Result
Gender relationship to diabetes category	Statistically significant relationship ( $\chi^2 = 131$ ; $p < 0.0001$ )
Hypertension diagnosis relationship to diabetes category	Statistically significant relationship ( $\chi^2 = 3269$ ; $p = 0.0$ )
Moderate activity relationship to diabetes category	Statistically significant relationship ( $\chi^2 = 381$ ; $p < 0.0001$ )
Marital status relationship to diabetes category	Statistically significant relationship ( $\chi^2 = 17$ ; $p = 0.0002$ )

T-Tests	Result
Means of triglycerides of no diabetes and prediabetes categories	Statistically different ( $t = -21$ ; $p < 0.0001$ )
Means of systolic blood pressure of no diabetes and prediabetes categories	Statistically different ( $t = -40$ ; $p = 0.0$ )
Means of body mass index of no diabetes and prediabetes categories	Statistically different ( $t = -36$ ; $p < 0.0001$ )
Means of serum iron of no diabetes and prediabetes categories	Statistically different ( $t = 10$ ; $p < 0.0001$ )
Means of aspartate aminotransferase (AST) of no diabetes and prediabetes categories	Statistically different ( $t = -7$ ; $p < 0.0001$ )



# MACHINE LEARNING APPROACH

Two groups of machine learning models were developed:

- The first group were multiclass classifier models trained to predict one of three categories: no diabetes, prediabetes, or diabetes. These categories were defined as described previously by comparing the NHANES diagnostic lab values to the ADA diagnostic criteria.
- The second group were binary classifier models trained to predict no diabetes or diabetes. These categories were defined by noting which participants were taking medication for diabetes.

As the goal for this project is classification, the algorithms chosen were Logistic Regression and tree-based algorithms, including Decision Tree, Random Forest, and various Boosting algorithms including Adaboost, Gradient Boosting, and XGBoost. Support Vector Machine and k-Nearest Neighbors were ruled out due to the very high dimensionality of the data. The chosen algorithms were first used with default parameters, then results were reviewed and selected models were tuned by optimizing hyperparameters with `RandomizedSearchCV()` using `StratifiedKFold()` with 5 folds.

Evaluation criteria had to span both regression and tree-based models. Initial criteria chosen were ROC AUC, Precision and Recall. Accuracy, Specificity and F1-score were also reported. For the multiclass models, a one vs all approach was taken to plotting ROC curves and calculating AUC values.

## PREPROCESSING

A significant amount of preprocessing was necessary to prepare the data.

- Data features with more than 40% null values were eliminated.
- Data that was administrative in nature was eliminated (e.g., interviewer id number).
- Several features were duplicates of others, but in different units. These were eliminated.
- Several features were so strongly correlated they were nearly identical (e.g., self-reported weight and actual measured weight). The self-reported values were eliminated.
- Features used to derive the targets were eliminated.
- NHANES contains much self-reported data in the questionnaire and demographics sections. Many continuous self-reported features have code values high above the range of the data. Some code values indicate “don’t know” or “refused to answer”. Some also indicate the reported value is above a threshold and not indicative of a specific number. These code values were stripped off several continuous features so the models would not interpret them numerically.
- 30% of the data was split off and saved to test the models.
- Missing continuous data was imputed with `scikit-learn`’s `IterativeImputer()` using Bayesian Ridge regression and 5 neighbor columns as features.
- NHANES categorical data is nearly all integer codes. Missing values for these integer categorical features were imputed the same as for continuous data, then rounded to integer values. This worked well as long as the integer categories were contiguous.
- All imputation was fit on training data, then used to transform training and test data.

- Categorical features were converted to dummy variables. The first category was dropped to prevent collinearity problems with the regression algorithm. Categories were defined from the training data. Any new categories found in test data were ignored.
- For Logistic Regression, the continuous features were normalized with StandardScaler() so large valued features would not dominate the regression. Tree-based algorithms used unscaled features.

## PART 1

The target was defined by comparing the three diagnostic labs to the ADA criteria.<sup>5</sup> As glycohemoglobin, plasma fasting glucose and oral glucose tolerance test results were used to define the targets, they were eliminated as features. Serum glucose was also eliminated due to its correlation with fasting plasma glucose. This made predicting prediabetes and diabetes more challenging, but reflects the real-life challenges of trying to determine who is high risk and should be tested.

### Summary of pre-tuning results:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Logistic Regression	.715	71.7 (.274)	.58	.63	.77	.60	.74
Decision Tree	.597	60.1 (.381)	.45	.46	.72	.46	.60
Random Forest	.674	67.3 (.540)	.55	.47	.80	.50	.74
Adaboost	.705	70.5 (.317)	.59	.53	.81	.56	.73
Gradient Boosting	.717	72.0 (.262)	.60	.56	.81	.58	.76
XGBoost	.712	71.7 (.457)	.60	.55	.81	.57	.77

### Post-tuning:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Random Forest	.702	70.7 (.584)	.57	.60	.78	.59	.76
XGBoost	.728	72.8 (.529)	.61	.59	.81	.60	.77

**Note: Accuracy is for overall model. All other measures are for Prediabetes class.**

**As this is medical data, reported Sensitivity & Specificity in addition to Precision & Recall.**

Tuned XGBoost performed the best, although tuning did not make a significant difference to performance.

Top features important to this model are:

DIQ010\_2.0 = Doctor ever said you have diabetes = NO

SMAQUEx\_2.0 = Smoking recent use questionnaire flag = >= 18 yrs old

RIDAGEYR = Age (yr)

DIQ050\_2.0 = Taking insulin now = NO

SMAQUEx2\_2.0 = Smoking cigarette use questionnaire flag = 12-17 yrs old

DIQ010\_3.0 = Doctor ever said you have diabetes = BORDERLINE

PHAFSTHR = Total fasting time before blood draw for labs (hr)

PHDSESN\_2.0 = Examination session = afternoon

PHDSESN\_1.0 = Examination session = morning

RIAGENDR\_2.0 = Gender = Female

BMXWAIST = Waist Circumference (cm)

DIQ160\_2.0 = Doctor ever said you have prediabetes = NO

The untuned model, with very similar performance, has slightly different feature importance:

DIQ010\_2.0 = Doctor ever said you have diabetes = NO

RIDAGEYR = Age (yr)

PHAFSTHR = Total fasting time before blood draw for labs (hr)

BMXWAIST = Waist circumference (cm)

DIQ010\_3.0 = Doctor ever said you have diabetes = BORDERLINE

LBXSGTSI = Gamma glutamyl transferase (U/L)

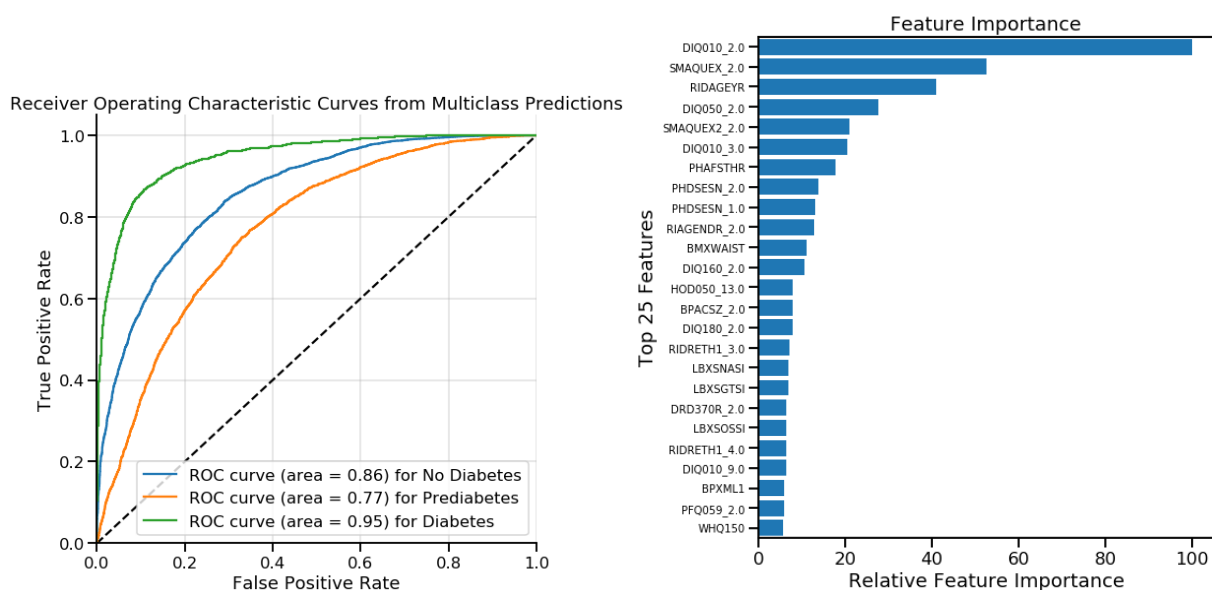
DIQ050\_2.0 = Taking insulin now = NO

BPACSZ\_3.0 = Blood pressure cuff size = Adult

LBXRDW = Red cell distribution width (%)

PHDSESN\_1.0 = Examination session = morning

The ROC curve for the tuned model shows it is a much better predictor of diabetes than prediabetes.



## PART 2

The goal of this portion of the study was binary classification of diabetes/no diabetes with a slightly different target. NHANES contains self-reported prescription medication. The NHANES interviewer looked at medication containers where possible. When the containers were not available, the interviewer took the participants' reports of their medications. Participants were asked the reason for each medication. Those answers were used to select an ICD-10-CM code from a reference list built into the survey. ICD-10-CM is the *International Classification of Diseases, 10th Revision, Clinical Modification* (ICD-10-CM).<sup>6</sup>

This ICD-10-CM code data was used as the target for the second half of the analysis. There are only 7990 participants with this data as it was tracked starting in 2013 and not everyone takes prescription medication.

The ICD-10-CM code prefix E11 was used to identify the Diabetes category per a CDC ICD-10-CM code lookup tool.<sup>6</sup> There was insufficient data to create a prediabetes category, so this was a binary classification problem.

The lab values used to determine the target for part 1 were included as features in this analysis. Plasma fasting glucose and oral glucose tolerance test were both dropped due to excessive nulls. Serum glucose (also fasting) was included.

XGBoost once again performed best. Tuning made no difference in its performance.

**Summary of results pre-tuning:**

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Logistic Regression	.935	93.8 (.723)	.75	.88	.95	.81	.97
Decision Tree	.928	92.0 (.576)	.77	.78	.96	.77	.87
Random Forest	.917	91.7 (.703)	.85	.57	.98	.69	.97
Adaboost	.942	94.1 (.575)	.82	.81	.97	.81	.97
Gradient Boosting	.953	95.4 (.324)	.81	.92	.96	.86	.98
XGBoost	.952	95.5 (.537)	.82	.89	.96	.85	.98

**Post-tuning:**

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Random Forest	.948	95.4 (.448)	.78	.93	.95	.85	.98
XGBoost	.951	95.7 (.596)	.81	.89	.96	.85	.98

**Note: Accuracy is for overall model. All other measures are for positive class: Diabetes.**

**As this is medical data, reported Sensitivity & Specificity in addition to Precision & Recall.**

Top features are:

DIQ010\_2.0 = Doctor ever said you have diabetes = NO

DIQ010\_3.0 = Doctor ever said you have diabetes = BORDERLINE

LBXGH = Glycohemoglobin (%)

DIQ050\_2.0 = Taking insulin now = NO

DIQ170\_2.0 = Ever told you had a health risk for diabetes = NO

LBXSGL = Glucose, serum (mg/dL)

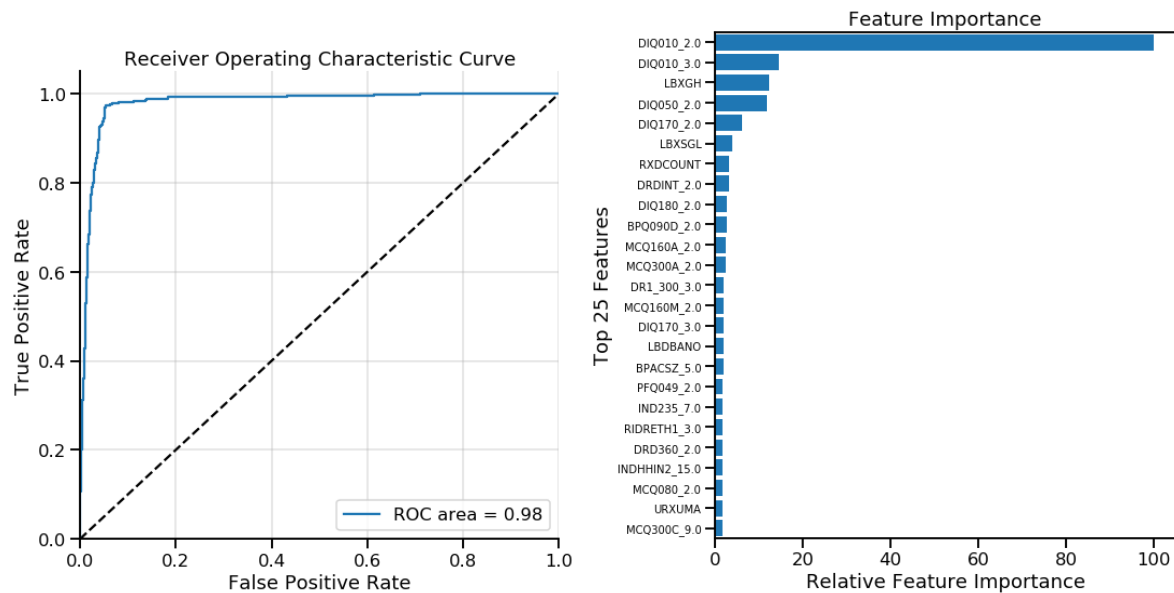
RXDCOUNT = Number of prescription medications

DRDINT\_2.0 = Number of days tracked dietary intake in survey = 2

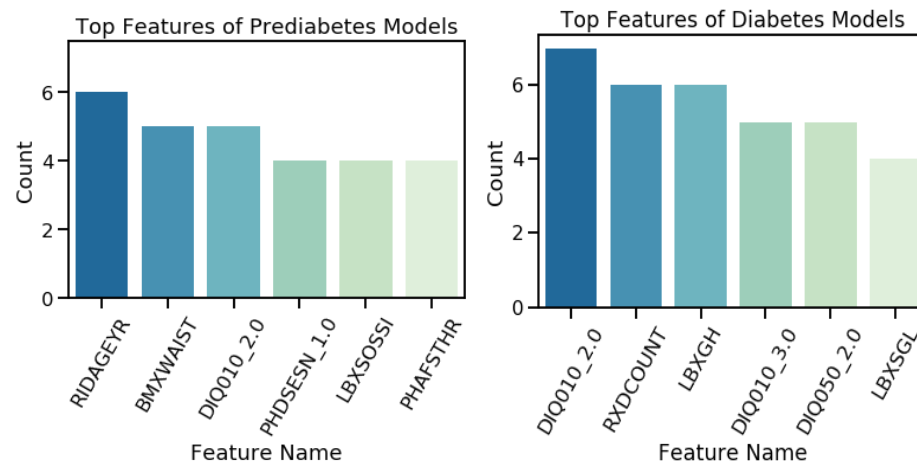
DIQ180\_2.0 = Blood test for high blood sugar in past three years = NO

BPQ090D\_2.0 = Ever told to take a prescription for cholesterol = NO

The ROC curve below illustrates this binary model is much better at predicting diabetes accurately than the best model's diabetes classification in part 1. This is likely due to the addition of glycohemoglobin and serum glucose as features.



The most frequent important features across all models are:



where:

RIDAGEYR = Age at interview (years)  
 BMXWAIST = Waist circumference (cm)  
 DIQ010\_2.0 = Doctor ever said you have diabetes = NO  
 PHDSESN\_1.0 = Examination session = morning  
 LBXSOSI = Osmolality (mmol/Kg)  
 PHAFSTHR = Fasting time before blood draw for labs (hr)

RXDCOUNT = Number of prescription medications

LBXGH = Glycohemoglobin (%)  
DIQ010\_3.0 = Doctor ever said you have diabetes = BORDERLINE  
DIQ050\_2.0 = Taking insulin now = NO  
LBXSGI = Glucose, serum (mg/dL)

## CONCLUSION

The XGBoost model performed best in both parts of this analysis.

- The best prediabetes model has about a 60% rate of correctly predicting prediabetes. This leaves room for improvement but is still much better than random chance as the prediabetes class comprises only 33.7% of the data.
- The best diabetes model has about an 80% rate of correctly predicting diabetes. This class makes up only 15.8% of the original data.

These precision rates may be partly due to the imbalance between the classes as there is a 54%/34%/13% ratio of no diabetes/prediabetes/diabetes observations in the dataset. It would be interesting to test whether balancing the classes would improve the models.

Predicting prediabetes proved more difficult than predicting diabetes. This is most likely because there is more overlap in features between the no diabetes and prediabetes classes than between no diabetes and diabetes. In addition, some in the prediabetes category may have only recently moved into that category and may not yet have developed sufficient complications or elevated laboratory results that would assist the models in classification.

The models in part 2 likely performed better than those in part 1 due to the inclusion of glycohemoglobin and serum glucose as features.

Finally, the NHANES dataset contains a large number of variables, many of which contributed noise to the models. Some of this noise was removed but certainly some remains. Perhaps the models would perform better with more of the noise removed through additional preprocessing of the data.

The only novel feature identified by more than half the models is osmolality in predicting prediabetes. While not part of the ADA risk assessment tool, there is an established relationship between osmolality and elevated blood glucose.<sup>7</sup>

## FUTURE WORK

Based on this study's findings, the following are areas of additional work that may yield improved results.

- Additional Feature Engineering:
  - Many of the laboratory test results have large outliers that were left in the data. These could be removed or modified.
  - Several administrative features were removed. More exist that could be removed.
  - Categorical imputation method worked well for contiguous integer categories. While all categorical features used were integers, not all were contiguous leading to a few new, undefined categories.
  - First category was dropped in creation of dummy variables to avoid collinearity problems in regression. This may result in lost information for tree-based algorithms.
  - Resampling to balance classes may improve results.
  - Several features with known high correlation were removed. There may be others.
- Combining models or other models:
  - Perhaps another algorithm or combination of algorithm would yield better results.

## REFERENCES

1. Centers for Disease Control and Prevention. (2020) *National Diabetes Statistics Report*. <https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf>
2. Brannick, B., Wynn, A., & Dagogo-Jack, S. (2016) Prediabetes as a toxic environment for the initiation of microvascular and macrovascular complications. *Experimental Biology and Medicine*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4950274/>
3. American Diabetes Association. (2018) *Economic Costs of Diabetes in the U.S. in 2017*. <https://care.diabetesjournals.org/content/early/2018/03/20/dci18-0007>
4. American Diabetes Association. (2020) *Statistics About Diabetes*. <https://www.diabetes.org/resources/statistics/statistics-about-diabetes>
5. American Diabetes Association. (n.d.) *Diagnosis*. <https://www.diabetes.org/a1c/diagnosis>
6. Centers for Disease Control and Prevention. (2019) *ICD-10-CM*. <https://icd10cmtool.cdc.gov/?fy=FY2019>
7. Rao GM (1992) *Serum electrolytes and osmolality in diabetes mellitus*. <https://www.ncbi.nlm.nih.gov/pubmed/1293047>
8. Kautzky-Willer, A., Harreiter, J., & Pacini, G. (2016). Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus. *Endocrine reviews*, 37(3), 278–316. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4890267/>