Diabetes Prediction & Risk Factor Analysis

Understanding Differences in Prediction of Prediabetes and Diabetes

Taylor Erickson
Applied Computer Science
University of Colorado Boulder
Boulder Colorado US
taer7274@colorado.edu

PROBLEM STATEMENT/MOTIVATION

Type 2 diabetes is a common chronic illness, affecting roughly 1 in 10 American adults. It is caused by a prolonged inability to maintain healthy blood sugar levels. Diabetes is associated with many adverse outcomes, including heart disease, blindness, and kidney disease [1]. While there is no known cure for Type 2 diabetes, it is possible for patients to experience reduction of symptom severity or even reversal of the disease through treatment [2]. Thus, early detection and regular screening of high-risk patients is imperative.

One of the most widely studied risk factors for type 2 diabetes is the presence of prediabetes. Considered an intermediate stage of hyperglycemia, prediabetes is diagnosed when blood sugar levels are higher than normal but are not high enough to be considered diabetes. 25% of patients with prediabetes will develop diabetes within 3-5 years, and up to 70% will develop diabetes over the span of their lifetime. In addition to an increased risk of diabetes, prediabetes is also associated with myriad adverse health outcomes, including increased risk of cardiovascular disease, peripheral neuropathy, and stroke. While a very common condition, affecting roughly 1 in 3 American adults, prediabetes often goes undiagnosed, with 90% unaware they have the disease [3, 4]. Widespread screening of high-risk individuals has been recommended for detection of diabetes [5, 6]. However, universal screening for prediabetes is currently not recommended [7]. This is in part due to inconsistent findings on potential risk factors, making it difficult to identify who to screen and how to justify the cost [8, 9]. However, emerging evidence suggests that prediabetes is a distinct disease with its own set of risk factors, diagnostic criteria and disease progression [10, 11, 12, 13]. Unfortunately, there is still a gap in understanding of risk factors and appropriate screening criteria for prediabetes as a distinct disease.

The goal of this project is to leverage logistic regression techniques to identify the strongest predictors of prediabetes and diabetes as separate outcomes and compare them to current screening standards. This will help identify whether screening approaches could be expanded to improve the likelihood of diagnosis and early detection in individuals who are at high risk for prediabetes.

1 Literature Survey

There are a set of widely accepted risk factors for diabetes. They are commonly grouped as behavioral, cardiometabolic, and non-modifiable attributes. The most significant behavioral risk factors include physical activity, diet, smoking behavior, and alcohol consumption. The most significant cardiometabolic risk factors include obesity, hypertension, and dyslipidemia. Finally, the most significant attribute risk factors include gender, race/ethnicity, age, and genetic factors. [16, 17, 18, 19] When reviewing the available literature, these risk factors are well researched and thoroughly proven to be associated with diabetes. However, there is much less clarity on prediabetes risk factors. The first issue is that many studies combine prediabetes and diabetes into a single outcome, making it difficult to identify the differences between prediabetes and diabetes risk factors [24]. The second issue is that there are differing definitions of prediabetes, which causes inconsistency in modeled risk factor strengths [20]. The third issue is that studies investigating prediabetes often consider prediabetes as a risk factor to progression of diabetes, rather than an outcome of interest [21, 22, 23].

However, if we consider studies that have investigated prediabetes risk factors specifically, we find there are some similar risk factors between diabetes and prediabetes. Examples of overlapping risk factors include gender, age, hypertension, dyslipidemia, obesity, and diet [10, 25, 29, 31, 32, 36]. However, there is not consistency between studies on which attributes are strongly associated with prediabetes. Some studies find that factors like gender, age, or diet are not strongly associated with prediabetes, and others find that factors like marital status, income, and education level are significantly associated [25, 27, 29, 31, 32, 34, 37]. This inconsistency is likely due to unbalanced class sizes, low incidence of prediabetes patients, inconsistency in the definition of prediabetes, and overconfidence in the similarity of the pathogenesis of prediabetes and diabetes [8, 11, 20, 28].

Generally, the variance in risk factors for prediabetes is related to a greater difficulty identifying statistically significant risk factors for prediabetes and a reduced strength of prediction for risk factors that overlap with diabetes [27, 30, 32].

For example, a study by Okwechime et al. (2015) found that being overweight, obese, hypertensive, hypercholesterolemic, and arthritic were significantly associated with prediabetes. However, they also found that these predictors were more strongly associated with diabetes. Additionally, many predictors that were not strongly associated with prediabetes were strongly associated with diabetes, including age, income level, and level of physical activity [27]. This inconsistency demonstrates the need for more thorough analysis of prediabetes risk factors and how they are distinct from or related to diabetes risk factors.

2 Proposed Work

The goal of this project is to leverage logistic regression modeling techniques to measure the strength of potential risk factors in predicting prediabetes and diabetes as separate classes. First, we will fit a multinomial logistic regression model on all three of our classes of interest, diabetes, no diabetes, and prediabetes. Then, once we have created a good fit multinomial logistic regression model, we will create a logistic regression model to examine prediction of prediabetes vs. no prediabetes, leveraging insights gained from the first model. We leverage logistic regression modeling because it is a good fit for the data mining task and has been previously used as a successful method in this research domain [26, 27, 31, 35]. We will compare the risk factors for diabetes and prediabetes to see whether they are the same or different. Finally, we will compare against existing research and documented risk factors to confirm whether we were able to replicate results reported by previous studies.

We will first complete basic data cleaning and exploration. For the first multinomial logistic regression model, there is not extensive data cleaning or transformation required, as it is using an existing cleaned dataset. Null or missing values have been removed and numeric transformation of categorical attributes. However, for the second logistic regression model, we will extract and aggregate multiple years of BRFSS survey results to try and increase the sample size of prediabetes as a class, given the small set of responses in the first dataset. This will require recreating the initial data transformation and cleaning that was conducted on the first dataset, as well as the integration of additional features of interest as required.[41]

Once the data cleaning is completed, feature selection techniques will be used to identify the strongest potential predictors in each dataset. We will measure the variance and distribution of each predictor compared to the outcomes of interest. We will leverage tests like chi-square, one-way ANOVA and t-test to measure equality of variance and influence of the feature on our outcome of interest. We will compare the variance across all outcomes of interest. Variables with low or no variance will be dropped.

Additionally, we will leverage techniques like Variance Inflation Factor (VIF) measures, Principal Component Analysis (PCA), and correlation heatmaps to identify collinear features and drop them from our dataset.

In addition, for our first model, there are only a small number of responses for the prediabetes class, so we will leverage oversampling techniques like SMOTE to create a more balanced dataset.

Once we have identified the best potential predictors, we will use logistic regression modeling techniques to build our models. First, we will create a multinomial logistic regression model to predict between our three classes, no diabetes, diabetes and pre-diabetes. This model will leverage a similar approach to the work done by Okwechime et al [2015], but our study will be conducted on a more recent, larger and more representative sample. Using the information gained from the results of our first model, we will create a best fit model for prediabetes vs. not prediabetes on the aggregated dataset. This will mirror existing research on diabetes prediction but will instead focus on prediabetes as a unique class [27, 33]. The focus on prediabetes as a unique outcome and aggregation of historical survey results is the unique and novel exploration being conducted. Once we have the best fit models, we will compare the multinomial logistic regression model to the prediabetes logistic regression model and examine the difference in predictors and strength.

Once we have created and fit our models, we will evaluate their performance. We will use statistical measures like area under the receiver operating characteristic (ROC) curve (AUC) to evaluate the strength of our model. We will evaluate p-values for our coefficients to understand the strength of our feature coefficients, with $p <= 0.05$ as significant. Once we have identified the strength of our relative models, we will revise any poorly performing models, leveraging techniques like forward stepwise refinement to improve the performance of our models.

3 Data Sets

This project will leverage multiple data sets, comprised of results from the annual CDC Behavioral Risk Factor Surveillance System (BRFSS) survey. The annual survey has been conducted since 1984 and collects telephone survey responses from over 400,000 Americans on a range of health conditions, health behaviors, and use of health services [14]. The data sets used in this project are a subset of the publicly available records, comprised of features that are specifically related to diabetes, prediabetes, and related risk factors. The dataset for the first model is available via Kaggle. [15]. The datasets for the second model are also available via Kaggle. [42]

The first dataset is a cleaned dataset of 253,680 survey responses with 3 classes in the target variable, no diabetes (or diabetes only during pregnancy), prediabetes, and diabetes. There is a class imbalance in this dataset, with an uneven distribution amongst the three classes. There are 21 attributes of interest.

The remaining datasets are the raw BRFSS survey output from 2011, 2013 and 2015. There are missing questions from 2012 and 2014, so they were excluded from this study. Each dataset contains over 400,000 responses, and over 400 attributes. Each dataset will require extensive cleaning and transformation and will mirror the initial cleaned 2015 dataset. [41] The cleaned datasets will be combined into an aggregated dataset for the creation of the second logistic regression model.

4 Evaluation Methods

The primary evaluation method for our project will be to leverage statistical measures like ROC/AUC curve analysis and p-values of our coefficients to determine the strength of our models and of our predictors. Additionally, we plan to compare our identified risk factors against the existing literature. We expect to see similarities between our selected features and those identified by similar methods. Additionally, we will compare our results against identified practitioner best practices and screening standards to identify any similarities or novel differences.

5 Tools

The project will leverage industry best practices for statistical modeling. The analysis will be conducted in Python, and requires use of these key statistical, modeling, and visualization libraries: pandas, seaborn, matplotlib, scipy, statsmodels, sklearn, numpy, yellowbrick.

The project will be completed in an interactive Jupyter notebook file, to improve readability and understanding of the modeling process. Visualizations and written explanation will be displayed next to the model generation and statistical analysis code blocks.

Given the size of the dataset, the data cleaning and processing does not require any other specialized data mining tools outside of what will be completed in the Juypter notebook.

6 Milestones

Completed Milestones

6.1 Milestone 1. July 17th – Ingestion and initial visual exploration of features.

6.2 Milestone 2. July 24th – Initial logistic regression models.

Remaining Milestones

6.3 Milestone 3. July 31st – Analysis and re-work to strengthen model performance.

6.4 Milestone 4. August 7th – Comparison between results and existing research.

6.5 Milestone 5. August 14th – Finished write-up and project delivered.

7 Results

Feature Selection

The first dataset explored was the three class 2015 BRFSS dataset, with responses labeled as prediabetes vs. diabetes vs. no diabetes. Initial feature selection focused on identifying strong predictors and reducing collinearity between selected variables.

For the three continuous numerical predictors (BMI, mental health, physical health), one-way ANOVA tests were conducted to determine whether the means of the three groups were statistically significantly different. BMI ($F = 6768.36, p = 0.0$), mental health ($F = 717.12$, $p < 0.0$), and physical health ($F = 4078.70, p = 0.0$) all had statistically significant differences between means, suggesting that they are related to our outcome of interest.

However, all three numerical variables were also skewed, with many outliers. Because outliers can cause issues with convergence, the variables were transformed using domain knowledge into discrete categorical predictors.

BMI was transformed to capture CDC classification of BMI: below 18.5 is underweight, 18.5-24.9 is healthy, 25 – 29.9 is overweight, 30 – 34.9 is class 1 obesity, 35 – 39.9 is class 2 obesity, and 40 and over is considered class 3 or severe obesity. [38] Chi-square test results indicated that the categorical BMI attribute still had statistically significant differences between classes, X^2 (10, $N = 253{,}680$) = 14508.0, $p = 0.0$, so it was selected as an initial feature in our modeling.

Mental health was also transformed to better capture the distribution of the variable without introducing convergence issues. Because most respondents (69.25%) reported having no poor mental health days, it was important to capture no poor mental health days separately from any poor mental health days. This mirrors existing analysis on BRFSS responses on number of poor mental health days [39]. Responses were categorized into two values: no poor mental health days or some poor mental health days. Chi-square test results indicated that the categorical mental health attribute still had statistically significant differences between classes, X^2 (2, N = 253,680) = 256.13, $p < 0.0$, so it was selected as an initial feature in our modeling.

Physical health was transformed in a similar way to mental health. Most respondents (63.09%) reported having no poor physical health days, so responses were categorized into two values: no poor physical health days or some poor physical health days. This mirrors existing analysis on BRFSS responses on number of poor physical health days. [39] Chi-square test results indicated that the categorical physical health attribute still had statistically significant differences between classes, X^2 (2, N = 253,680) = 4671.23, $p = 0.0$, so it was selected as an initial feature in our modeling.

The remainder of the variables were categorical attributes, so chi-square tests were conducted on each predictor. All predictors were statistically significantly related to the difference in means between our classes, suggesting that they are strong predictors and should be selected as initial features.

Once the initial features were selected, it was necessary to investigate and reduce collinearity between features.

Multinomial logistic regression requires the assumption that features are independent, which means that any strongly collinear features must be dropped. A correlation heatmap using the Pearson correlation coefficient was generated to identify collinear features. The strongest correlations were between general health and difficulty walking (r = 0.46), education and income (r = 0.45), and general health and physical health (r = 0.41). However, a correlation coefficient of 0.8 or greater is usually used to identify collinear features. [40] Therefore, no additional predictors were dropped from the analysis.

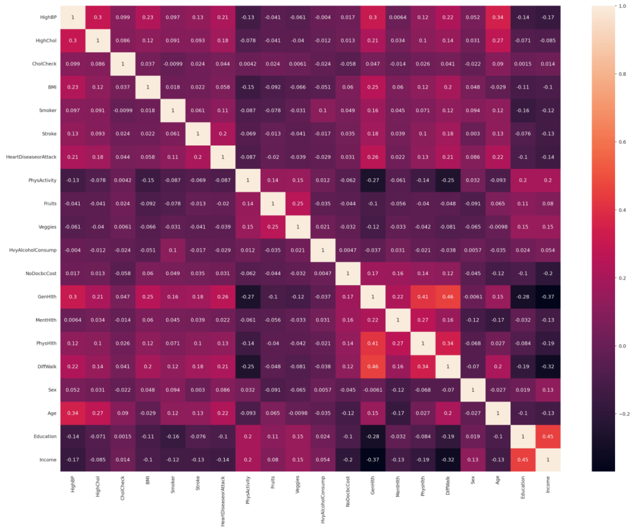
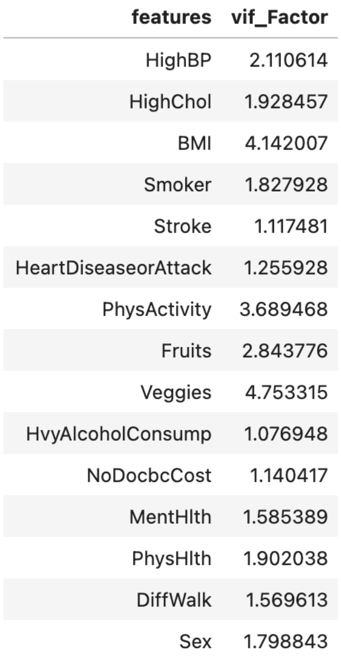


Figure 1: Pearson Correlation Coefficient Heatmap for Three Class Dataset

To further identify collinear features, a Variance Inflation Factor (VIF) calculation was used. A VIF score of greater than 5 indicates moderate to strong collinearity between variables. [40] Variable selection using VIF score was done recursively, dropping the variable with the highest VIF score and recalculating until all variables had a VIF score of less than 5. Education (*VIF* = 28.77), cholesterol check (*VIF* = 21.00), any healthcare (*VIF* = 16.99), general health (*VIF* = 8.96), income (*VIF* = 8.45), and age (*VIF* = 7.12) had high VIF scores and were dropped from analysis. The remaining attributes had VIF scores less than 5 and were retained in the analysis.



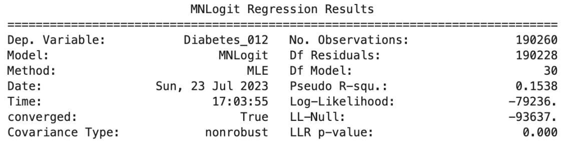
features	vif_Factor
HighBP	2.110614
HighChol	1.928457
BMI	4.142007
Smoker	1.827928
Stroke	1.117481
HeartDiseaseorAttack	1.255928
PhysActivity	3.689468
Fruits	2.843776
Veggies	4.753315
HvyAlcoholConsump	1.076948
NoDocbcCost	1.140417
MentHlth	1.585389
PhysHlth	1.902038
DiffWalk	1.569613
Sex	1.798843

Figure 2: Final VIF Score Output for Initial Selected Features for Three Class Dataset

The feature selection process identified 15 attributes out of the original 21 features for initial modeling and analysis.

Initial Multinomial Logistic Regression Model

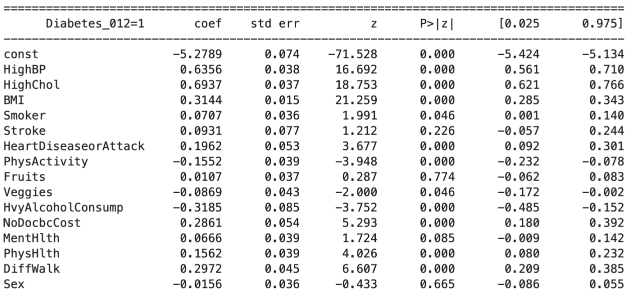
To create and fit a multinomial logistic regression model, a simple logistic regression model was used. The initial model, with all selected features, demonstrated a reasonable fit to the data (pseudo-R^2 = 0.1538), but had room for improvement, given that there were variables with statistically insignificant coefficient values.



```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:            Diabetes_012   No. Observations:               190260
Model:                         MNLogit   Df Residuals:                   190228
Method:                            MLE   Df Model:                           30
Date:                 Sun, 23 Jul 2023   Pseudo R-squ.:                  0.1538
Time:                         17:03:55   Log-Likelihood:                -79236.
converged:                        True   LL-Null:                       -93637.
Covariance Type:             nonrobust   LLR p-value:                     0.000
```

Figure 3: Simple Multinomial Logistic Regression Model for Three Class Dataset – Descriptive Statistics

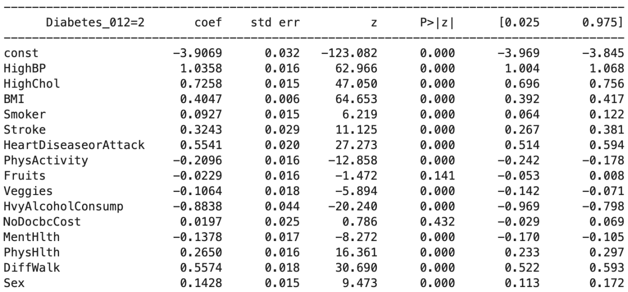
Variables with coefficient p-values of greater than 0.05 were dropped and the model refitted. For prediction of the prediabetes class, stroke (p = 0.226), consumption of fruits (p = 0.774), mental health (p = 0.085) and sex (p = 0.665) had p-values greater than 0.05 and were deemed not statistically significant predictors.



```
==============================================================================
      Diabetes_012=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                   -5.2789      0.074    -71.528      0.000      -5.424      -5.134
HighBP                   0.6356      0.038     16.692      0.000       0.561       0.710
HighChol                 0.6937      0.037     18.753      0.000       0.621       0.766
BMI                      0.3144      0.015     21.259      0.000       0.285       0.343
Smoker                   0.0707      0.036      1.991      0.046       0.001       0.140
Stroke                   0.0931      0.077      1.212      0.226      -0.057       0.244
HeartDiseaseorAttack     0.1962      0.053      3.677      0.000       0.092       0.301
PhysActivity            -0.1552      0.039     -3.948      0.000      -0.232      -0.078
Fruits                   0.0107      0.037      0.287      0.774      -0.062       0.083
Veggies                 -0.0869      0.043     -2.000      0.046      -0.172      -0.002
HvyAlcoholConsump       -0.3185      0.085     -3.752      0.000      -0.485      -0.152
NoDocbcCost              0.2861      0.054      5.293      0.000       0.180       0.392
MentHlth                 0.0666      0.039      1.724      0.085      -0.009       0.142
PhysHlth                 0.1562      0.039      4.026      0.000       0.080       0.232
DiffWalk                 0.2972      0.045      6.607      0.000       0.209       0.385
Sex                     -0.0156      0.036     -0.433      0.665      -0.086       0.055
```

Figure 4: Correlation Coefficients for Prediction of Prediabetes Class

For prediction of the diabetes class, consumption of fruits (p = 0.141) and no doctor as a result of cost (p = 0.432) had p-values greater than 0.05 and were deemed not statistically significant predictors.

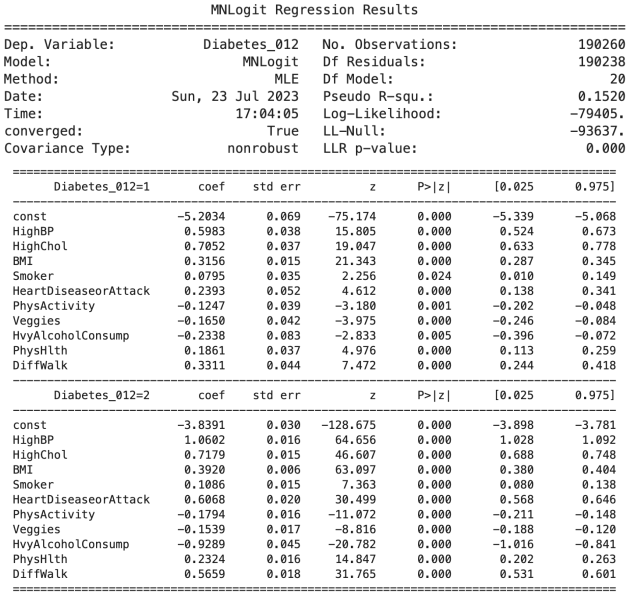


```
------------------------------------------------------------------------------
      Diabetes_012=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                   -3.9069      0.032   -123.082      0.000      -3.969      -3.845
HighBP                   1.0358      0.016     62.966      0.000       1.004       1.068
HighChol                 0.7258      0.015     47.050      0.000       0.696       0.756
BMI                      0.4047      0.006     64.653      0.000       0.392       0.417
Smoker                   0.0927      0.015      6.219      0.000       0.064       0.122
Stroke                   0.3243      0.029     11.125      0.000       0.267       0.381
HeartDiseaseorAttack     0.5541      0.020     27.273      0.000       0.514       0.594
PhysActivity            -0.2096      0.016    -12.858      0.000      -0.242      -0.178
Fruits                  -0.0229      0.016     -1.472      0.141      -0.053       0.008
Veggies                 -0.1064      0.018     -5.894      0.000      -0.142      -0.071
HvyAlcoholConsump       -0.8838      0.044    -20.240      0.000      -0.969      -0.798
NoDocbcCost              0.0197      0.025      0.786      0.432      -0.029       0.069
MentHlth                -0.1378      0.017     -8.272      0.000      -0.170      -0.105
PhysHlth                 0.2650      0.016     16.361      0.000       0.233       0.297
DiffWalk                 0.5574      0.018     30.690      0.000       0.522       0.593
Sex                      0.1428      0.015      9.473      0.000       0.113       0.172
```

Figure 5: Correlation Coefficients for Prediction of Diabetes Class

What is interesting even at this early stage of modeling was that there were differences in the features that were statistically significant predictors of the outcome of prediabetes compared to the outcome of diabetes. The features that were not statistically significant predictors of diabetes were consumption of fruit (p = 0.141) and no doctor because of cost (p = 0.432). However, no doctor was a statistically significant predictor of the pre-diabetes class ($p < 0.00$). Consumption of fruit was not a strong predictor of pre-diabetes (p=0.774), but history of stroke (p=0.226), sex (p=0.665), and mental health (p=0.085) were also not statistically significant predictors of prediabetes, even though they were of diabetes.

The features that were statistically insignificant for predicting prediabetes or diabetes were dropped from analysis and the model was refitted. High blood pressure (p = 0.000), high cholesterol (p = 0.000), BMI (p = 0.000), smoker (p = 0.024), heart disease or attack (p = 0.000), physical activity (p = 0.001), vegetable consumption (p = 0.000), heavy alcohol consumption (p = 0.005), physical health (p = 0.000), and difficulty walking (p = 0.000) all had statistically significant coefficient values. However, the performance of the model was not noticeably improved (pseudo-R^2 = 0.1520).



```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           Diabetes_012   No. Observations:               190260
Model:                        MNLogit   Df Residuals:                   190238
Method:                           MLE   Df Model:                           20
Date:                Sun, 23 Jul 2023   Pseudo R-squ.:                  0.1520
Time:                        17:04:05   Log-Likelihood:                -79405.
converged:                       True   LL-Null:                       -93637.
Covariance Type:            nonrobust   LLR p-value:                     0.000
  ============================================================================
      Diabetes_012=1       coef    std err          z      P>|z|      [0.025      0.975]
  ----------------------------------------------------------------------------
const                   -5.2034      0.069    -75.174      0.000      -5.339      -5.068
HighBP                   0.5983      0.038     15.805      0.000       0.524       0.673
HighChol                 0.7052      0.037     19.047      0.000       0.633       0.778
BMI                      0.3156      0.015     21.343      0.000       0.287       0.345
Smoker                   0.0795      0.035      2.256      0.024       0.010       0.149
HeartDiseaseorAttack     0.2393      0.052      4.612      0.000       0.138       0.341
PhysActivity            -0.1247      0.039     -3.180      0.001      -0.202      -0.048
Veggies                 -0.1650      0.042     -3.975      0.000      -0.246      -0.084
HvyAlcoholConsump       -0.2338      0.083     -2.833      0.005      -0.396      -0.072
PhysHlth                 0.1861      0.037      4.976      0.000       0.113       0.259
DiffWalk                 0.3311      0.044      7.472      0.000       0.244       0.418
  ----------------------------------------------------------------------------
      Diabetes_012=2       coef    std err          z      P>|z|      [0.025      0.975]
  ----------------------------------------------------------------------------
const                   -3.8391      0.030   -128.675      0.000      -3.898      -3.781
HighBP                   1.0602      0.016     64.656      0.000       1.028       1.092
HighChol                 0.7179      0.015     46.607      0.000       0.688       0.748
BMI                      0.3920      0.006     63.097      0.000       0.380       0.404
Smoker                   0.1086      0.015      7.363      0.000       0.080       0.138
HeartDiseaseorAttack     0.6068      0.020     30.499      0.000       0.568       0.646
PhysActivity            -0.1794      0.016    -11.072      0.000      -0.211      -0.148
Veggies                 -0.1539      0.017     -8.816      0.000      -0.188      -0.120
HvyAlcoholConsump       -0.9289      0.045    -20.782      0.000      -1.016      -0.841
PhysHlth                 0.2324      0.016     14.847      0.000       0.202       0.263
DiffWalk                 0.5659      0.018     31.765      0.000       0.531       0.601
  ============================================================================
```

Figure 6: Second Iteration of Simple Multinomial Logistic Regression Model, Three Class Dataset – Descriptive Statistics

To improve the performance of the model, the sklearn logistic regression model function was used to generate a multinomial logistic regression model with our remaining predictors. This function is more flexible, and often increases accuracy and strength of a prediction model because it can incorporate more complex relationships between features. The model had a training accuracy of 0.85 and test accuracy of 0.85. However, as can be seen in the confusion matrix, the model was not predicting any prediabetes cases.

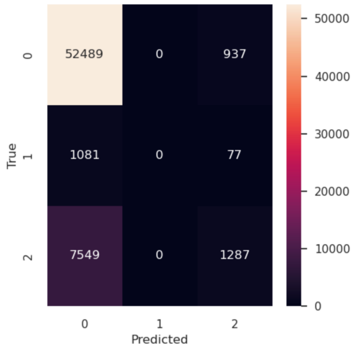


Figure 7: Confusion Matrix for Complex Multinomial Logistic Regression Model, Three Class Dataset

While the model performed well overall (micro-average *AUC* = 0.95, macro-average *AUC* = 0.75), it performed worst on the prediabetes class (*AUC* = 0.67) and about the same for no diabetes (*AUC* = 0.79) and diabetes (*AUC* = 0.79). This means that the model achieves decent performance on diabetes and no diabetes prediction but could be improved to better differentiate prediabetes cases.

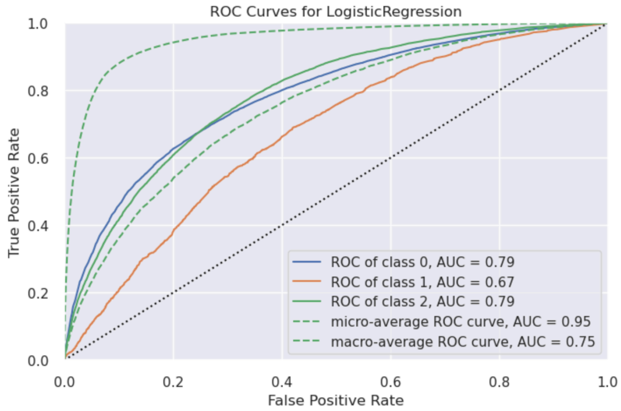


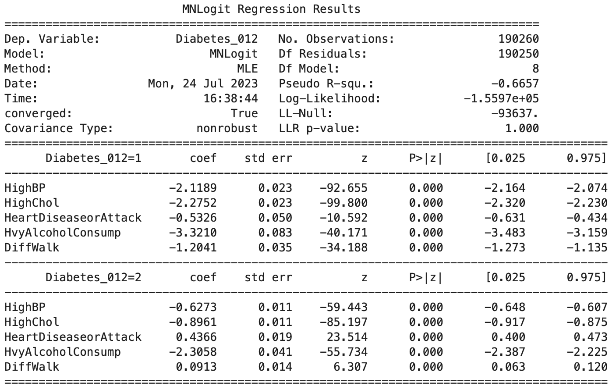
Figure 8: ROC Curves for Complex Multinomial Logistic Regression, Three Class Dataset

One of the reasons why this occurred is because the classes are highly imbalanced, with prediabetes representing only 1.83% of the total response population. To try and improve the performance of the model, and to accommodate for this class size imbalance, the SMOTE technique was used on the training data to oversample the prediabetes class. The prediabetes class size was oversampled to increase the number of samples from 3,473 to 10,000 responses.

However, this did not improve the performance of the model. The model achieved a training accuracy of 0.82, and a testing accuracy of 0.85, and saw little improvement in the overall AUC score (micro-average *AUC* = 0.95, macro-average *AUC* = 0.75). It still experienced greater difficulty predicting the prediabetes class (*AUC* = 0.67) relative to predicting the diabetes (*AUC* = 0.79) and the no diabetes (*AUC* = 0.78) classes.

An additional reason why there might be difficulty fitting this model is the large number of features.

Recursive feature elimination was used to identify the strongest predictors for the multinomial logistic regression model. The process identified five features as the best predictors: high blood pressure, high cholesterol, history of heart disease or heart attack, heavy alcohol consumption, and difficulty walking. When fitting a simple multinomial logistic regression model with statsmodels, the performance of the model was significantly improved (pseudo-R^2 = -0.6657). However, the LLR p-value (p = 1.0) indicated that this model was not a statistically significantly improvement relative to the null model (log-likelihood = -155,970, log-likelihood null = -93,637).



```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:           Diabetes_012   No. Observations:               190260
Model:                        MNLogit   Df Residuals:                   190250
Method:                           MLE   Df Model:                            8
Date:                Mon, 24 Jul 2023   Pseudo R-squ.:                 -0.6657
Time:                        16:38:44   Log-Likelihood:            -1.5597e+05
converged:                       True   LL-Null:                       -93637.
Covariance Type:            nonrobust   LLR p-value:                     1.000
=======================================================================================
      Diabetes_012=1       coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
HighBP                  -2.1189      0.023    -92.655      0.000      -2.164      -2.074
HighChol                -2.2752      0.023    -99.800      0.000      -2.320      -2.230
HeartDiseaseorAttack    -0.5326      0.050    -10.592      0.000      -0.631      -0.434
HvyAlcoholConsump       -3.3210      0.083    -40.171      0.000      -3.483      -3.159
DiffWalk                -1.2041      0.035    -34.188      0.000      -1.273      -1.135
---------------------------------------------------------------------------------------
      Diabetes_012=2       coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
HighBP                  -0.6273      0.011    -59.443      0.000      -0.648      -0.607
HighChol                -0.8961      0.011    -85.197      0.000      -0.917      -0.875
HeartDiseaseorAttack     0.4366      0.019     23.514      0.000       0.400       0.473
HvyAlcoholConsump       -2.3058      0.041    -55.734      0.000      -2.387      -2.225
DiffWalk                 0.0913      0.014      6.307      0.000       0.063       0.120
=======================================================================================
```

Figure 9: Simple Multinomial Logistic Regression Model Using Features Selected Using RFE

When usingthe sklearn library to generate a complex multinomial logistic regression model fitted with the RFE selected features, we still achieved similar performance to previous models. The model had a training accuracy of 0.84 and test accuracy of 0.84. Additionally, there was a slight decrease in AUC scores overall, with prediabetes class still the worst (*AUC* = 0.65), with diabetes (*AUC* = 0.77) and no diabetes (*AUC* = 0.76) still performing similarly.

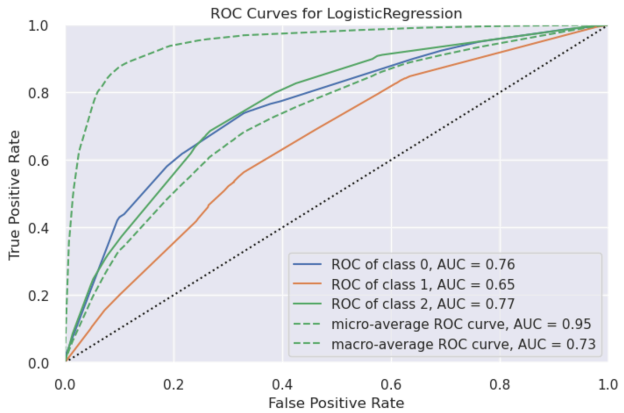


Figure 10: ROC Curves for Complex Multinomial Logistic Regression Model Using Features Selected using RFE

To understand why there was not improved performance as a result of selecting the top features identified during RFE, we used principal component analysis to understand how much variance is explained by each attribute. Unfortunately, each variable explains only a small proportion of the variance in our outcome classes, suggesting that a good fit model will likely need to contain many dimensions.

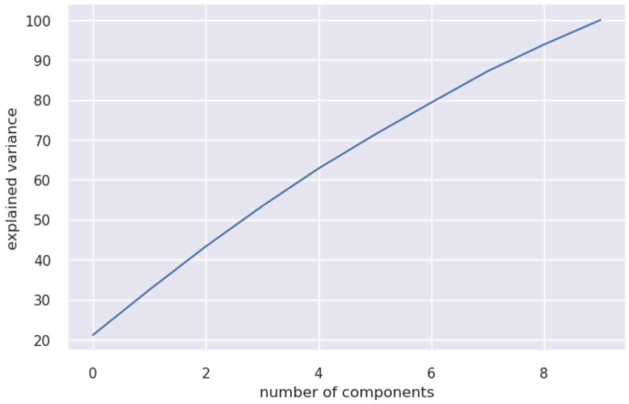


Figure 11: Principal Component Analysis of Available Features in Three Class Dataset

To best understand how to differentiate between prediction of prediabetes and diabetes, we need to create a logistic regression model to predict prediabetes as a separate class and identify the strongest features so that we can better inform a multinomial logistic regression model and improve understanding of whether there's any significant difference between prediabetes and diabetes predictive features.

REFERENCES

[1] CDC. 2023. Type 2 Diabetes. (April 2023). Retrieved July 10, 2023 from https://www.cdc.gov/diabetes/basics/type2.html

[2] Sarah J Hallberg, Victoria M Gershuni, Tamara L Hazbun, and Shaminie J Athinarayanan. 2019. Reversing Type 2 Diabetes: A Narrative Review of the Evidence. *Nutrients* 11, 4, (April 2019), 766. DOI: https://doi.org/10.3390/nu11040766

[3] Adam G. Tabák, Christian Herder, Wolfgang Rathmann, Eric J. Brunner, and Mika Kivimäki. Prediabetes: a high-risk state for diabetes development. *Lancet* 379, 9833, (June 2012), 2279-2290. DOI: https://doi.org/10.1016/S0140-6736(12)60283-9

perspectives. *Clinical Diabetes and Endocrinology* 5, 5, (May 2019). DOI: https://doi.org/10.1186/s40842-019-0080-0

[5] US Preventive Services Task Force. Screening for Prediabetes and Type 2 Diabetes: US Preventive Services Task Force Recommendation Statement. *JAMA*. 326, 8, (August 2021), 736–743. DOI: https://10.1001/jama.2021.12531

[6] Kenneth Lam and Sei J. Lee. Prediabetes – A Risk Factor Twice Removed. *JAMA Internal Medicine* 181, 4, (August 2021), 520-521. DOI: https://doi.org/10.1001/jamainternmed.2020.8773

[7] Aditya K. Khetan and Sanjay Rajagopalan. Prediabetes. *Canadian Journal of Cardiology* 34, 5, (May 2018), 615-623. DOI: https://doi.org/10.1016/j.cjca.2017.12.030

[8] Ashkann Zand, Karim Ibrahim and Bhargavi Patham. Prediabetes: Why Should We Care? Methodist Debakey Cardiovascular Journal,14, 4, (October 2018), 289-297. DOI: https://doi.org/10.14797/mdcj-14-4-289

[9] Justin B. Echouffo-Tcheugui and Elizabeth Selvin. Prediabetes and What It Means: The Epidemiological Evidence. *Annual Review of Public Health* 42, 59, (April 2021), 59-77. DOI: https://doi.org/10.1146/annurev-publhealth-090419-102644

[10] Alicia Diaz-Redondo, Carolina Giráldez-García, Lourdes Carrillo et al. Modifiable risk factors associated with prediabetes in men and women: a cross-sectional analysis of the cohort study in primary health care on the evolution of patients with prediabetes (PREDAPS-Study). *BMC Family Practice*, 16, Article 5, (January 15). DOI: https://doi.org/10.1186/s12875-014-0216-3

[11] J.F. Elgart, R. Torrieri, M. Ré et al. Prediabetes is more than a pre-disease: additional evidences supporting the importance of its early diagnosis and appropriate treatment. *Endocrine* 79, (January 2023), 80-85. DOI: https://doi.org/10.1007/s12020-022-03249-8

[12] Jan Brož, Jana Malinovska, Marisa A. Nunes et al. Prevalence of diabetes and prediabetes and its risk factors in adults aged 25-64 in the Czech Republic: A cross-sectional study. *Diabetes Research and Clinical Practice*, 170, Article 108470, (September 2020). DOI: https://doi.org/10.1016/j.diabres.2020.108470

[13] Khaled K. Aldossari, Abdulrahman Alidiab, Jamaan M. Al-Zahrani et al. Prevalence of Prediabetes, Diabetes, and Its Associated Risk Factors among Males in Saudi Arabia: A Population-Based Survey. *Journal of Diabetes Research* 2018, Article 2194604 (April 2018), 12 pages. DOI: https://doi.org/10.1155/2018/2194604

[14] CDC. Behavioral Risk Factor Surveillance System. (May 2023). Retrieved July 10, 2023 from https://www.cdc.gov/brfss/index.html

[15] Alex Teboul. Diabetes Health Indicators Dataset. (2021). Retrieved July 10, 2023 from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[16] Sania Siddiqui, Hadzliana Zainal, Sabariah Noor Harun, Siti Maisharah Sheikh Ghadzi, and Saadia Ghafoor. Gender differences in the modifiable risk factors associated with the presence of prediabetes: A systematic review. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14, 5, (July 2020), 1243-1252. DOI: https://doi.org/10.1016/j.dsx.2020.06.069

[17] CDC. Diabetes Risk Factors. (April 2022). Retrieved July 10, 2023 from https://www.cdc.gov/diabetes/basics/risk-factors.html

[18] Barbara Fletcher, Meg Gulanick, and Cindy Lamendola. Risk Factors for Type 2 Diabetes Mellitus. *The Journal of Cardiovascular Nursing*, 16, 2, (January 2002), 17-23.

[19] Phillipa J Talmud, Aroon D. Hingorani, Jackie A Cooper et al. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *BMJ* 340, Article 4838, (January 2010). DOI: https://doi.org/10.1136/bmj.b4838

[20] Crystal Man Ying Lee, Stephen Colagiuri, Mark Woodward et al. Comparing different definitions of prediabetes with subsequent risk of diabetes: an individual participant data meta-analysis involving 76,513 individuals and 8,208 cases of incident diabetes. *BMJ Open Diabetes Research and Care,* 7, 1, (November 2019). DOI: https://doi.org/10.1136/ bmjdrc-2019-000794

[21] Bennasar-Veny M, Fresneda S, López-González A, Busquets-Cortés C, Aguiló A, Yañez AM. Lifestyle and Progression to Type 2 Diabetes in a Cohort of Workers with Prediabetes. *Nutrients,* 12, 5, Article 1538 (May 2020,. DOI: https://doi.org/10.3390/nu12051538

[22] Rooney MR, Rawlings AM, Pankow JS, et al. Risk of Progression to Diabetes Among Older Adults With Prediabetes. *JAMA Intern Med*., 181, 4, (February 2021), 511-519. DOI: https://doi.org/10.1001/jamainternmed.2020.8774

[23] N. Anthony, V. Lenclume, A. Fianu, N.Le Moullec, X. Debussche, P. Gérardin, C. Marimoutou, E. Nobécourt, Association between prediabetes definition and progression to diabetes: The REDIA follow-up study. *Diabetes Epidemiology and Management,* 3, Article 100024 (November 2021). DOI: https://doi.org/10.1001/jamainternmed.2020.8774

[24] Hai Wang, Xin Zheng, Zheng-Hai Bai, Jun-Hua Lv, Jiang-Li Sun, Yu Shi, and Hong-Hong Pei. A retrospective population study to develop a predictive model of prediabetes and incident type 2 diabetes mellitus from a hospital database in Japan between 2004 and 2015. *Medical Science Monitor*, 26, (April 2020). DOI: https://doi.org/10.12659/MSM.920880

[25] Jiahua Wu, Jiaqiang Zhou, Xueyao Yin, Yixin Chen, Xihua Lin, Zhiye Xu, and Hong Li. A Prediction Model for Prediabetes Risk in Middle-Aged and Elderly Populations: A Prospective Cohort Study in China. *International Journal of Endocrinology*, 2021, Article 2520806 (November 2021). DOI: https://doi.org/10.1155/2021/2520806.

[26] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors, *The Kaohsiung Journal of Medical Sciences*, 29, 2 (October 2012), 93-99. DOI: https://doi.org/10.1016/j.kjms.2012.08.016

[27] Ifechukwude Obiamaka Okwechime , Shamarial Roberson, and Agricola Odoi. Prevalence and Predictors of Pre-Diabetes and Diabetes among Adults 18 Years or Older in Florida: A Multinomial Logistic Modeling Approach. *PLOS ONE,* 10, 12, Article e0145781 (2015). DOI: https://doi.org/10.1371/journal.pone.0145781

[28] Gary S. Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med*icine 9, 1, (December 2011), 1-14.

[29] Mayo Clinic. 2022. Prediabetes. (November 2022). Retrieved July 10, 2023 from https://www.mayoclinic.org/diseases-conditions/prediabetes/symptoms-causes/syc-20355278

[30] Vijay Viswanathan, Satyavani Kumpatla, Vigneswari Aravindalochanan et al. Prevalence of Diabetes and Pre-Diabetes and Associated Risk Factors among Tuberculosis Patients in India. *PLOS ONE,* 7, 7, (2012), Article ID e41367. DOI: https://doi.org/10.1371/journal.pone.0041367

[31] Zhao, Ming, Hongbo Lin, Yanyan Yuan, Fuyan Wang, Yang Xi, Li Ming Wen, Peng Shen, and Shizhong Bu. Prevalence of Pre-Diabetes and Its Associated Risk Factors in Rural Areas of Ningbo, China*. International Journal of Environmental Research and Public Health* 13, 8, Article 808 (August 2016). DOI: https://doi.org/10.3390/ijerph13080808

[32] Hemavathi Dasappa, Farah Naaz Fathima, Rugmani Prabhakar, Sanjay Sarin. Prevalence of diabetes and pre-diabetes and assessments of their risk factors in urban slums of Bangalore. *Journal of Family Medicine and Primary Care*, 4, 3 (July 2015), 399-404. DOI: https://doi.org/10.4103/2249-4863.161336

[33] Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Prev Chronic Dis* 16, Article 130, (September 2019). DOI: https://doi.org/10.5888/pcd16.190109

[34] F Hadaegh, A Derakhshan, N Zafari, D Khalili, M Mirbolouk,N Saadat, and F Azizi. Pre-diabetes tsunami: incidence rates and risk factors of pre-diabetes and its different phenotypes over 9 years of follow-up. *Diabet Med*. 34, 1, (January 2017), 69-78. DOI: https://doi.org/10.1111/dme.13034

[35] Rui Wang, Peng Zhang, Zhijun Li et al. The prevalence of pre-diabetes and diabetes and their associated factors in Northeast China: a cross-sectional study. *Scientific Reports*, 9, Article 2513, (February 2019). https://doi.org/10.1038/s41598-019-39221-2

[36] Parisa Amiri, Sara Jalali-Farahani, Mehrdad Karimi et al. Factors associated with pre-diabetes in Tehranian men and women: A structural equations modeling. *PLOS ONE*, 12, 12, Article e0188898, (December 2017). https://doi.org/10.1371/journal.pone.0188898

[37] Anna Zamora-Kapoor, Amber Fyfe-Johnson, Adam Omidpanah, Dedra Buchwald, and Ka'imi Sinclair. Risk factors for pre-diabetes and diabetes in adolescence and their variability by race and ethnicity. *Preventive Medicine*, 115, (August 2018), 47-52. DOI: https://doi.org/10.1016/j.ypmed.2018.08.015

[38] CDC. Defining Adult Overweight & Obesity. (2022). Retrieved July 23, 2023 from https://www.cdc.gov/obesity/basics/adult-defining.html

[39] Lindsey Lanigan and Colin Planalp. Measuring State-level Disparities in Unhealthy Days. (2022). Retrieved July 23, 2023 from https://www.shadac.org/news/measuring-unhealthy-days-SHC

[40] Noora Shrestha. Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and* Statistics, 8, 2, (June 2020), 39-42. DOI: https://doi.org/10.12691/ajams-8-2-1

[41] Alex Teboul. Diabetes Health Indicators Dataset Notebook. (2022). Retrieved July 10, 2023 from https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook/notebook

[42] CDC. Behavioral Risk Factor Surveillance System. (2017). Retrieved July 10, 2023 from https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system?select=2011.csv