# Early-Onset vs. Late-Onset Diabetes Risk Prediction Using NHANES Data

## 1. Summary

Diabetes mellitus represents a significant and growing public health burden. Early-onset diabetes (diagnosed before age 40) is particularly concerning due to its association with more severe disease phenotypes and higher complication rates compared to late-onset diabetes. This study aimed to develop and compare predictive models for early-onset vs. late-onset diabetes using National Health and Nutrition Examination Survey data. Logistic regression and random forest classification models were implemented on two age groups (<40 vs. ≥40 years). Feature selection and cross-validation were conducted and class imbalance were addressed through techniques including ADASYN and SMOTE. Models demonstrated stronger predictive performance for early-onset diabetes (ROC AUC > 0.90) compared to late-onset diabetes (ROC AUC 0.62-0.79). Body mass index was the strongest predictor across all models, with significant differences in secondary predictors between age groups. Despite strong discriminatory ability, precision was limited by substantial class imbalance (2.15% diabetes prevalence in early-onset group). These findings suggest distinct risk patterns for early-onset diabetes and highlight the challenge of developing screening tools with acceptable false positive rates for younger populations.

## 2. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by hyperglycemia that results from impaired insulin secretion, diminished insulin action, or a combination of both pathophysiological mechanisms (Sacks et al., 2023). This rapidly growing epidemic affects 38.4 million people of the U.S population, which is a striking increase from the reported 34.2 million from just 3 years prior in 2018 (U.S. Centers for Disease Control and Prevention, 2024). This disease is associated with numerous complications, including hypertension, renal disease, and dyslipidemia, imposing substantial economic burden on healthcare systems (Cioana et al., 2021; Dabelea et al., 2017; Eppens et al., 2006; Pinhas-Hamiel & Zeitler, 2007). Notably, research on early-onset diabetes (diagnosed before age 40) remains limited despite its increasing prevalence and association with more severe disease phenotype and distinct risk factors (Chandrasekaran et al., 2024; Cveticanin & Arsenovic, 2025; Grisanti, 2018; James et al., 2023; Jiang & Lai, 2024; Strati et al., 2025). Therefore, enhancing preventive interventions, particularly refining personalized diabetes risk prediction models, is critical (Nijpels et al., 2019; Wareham, 2022).

Several recent studies have employed machine learning algorithms to develop diabetes risk prediction models specifically for youth populations, though this research area remains relatively emerging (McDonough et al., 2023; Mohsen et al., 2023; Vangeepuram et al., 2021). Vangeepuram et al. evaluated pediatric clinical screening guidelines based on American Diabetes Association diagnostic biomarkers using various machine learning algorithms to assess their predictive efficacy (Vangeepuram et al., 2021). They found significant limitations in these guidelines' ability to accurately predict diabetes risk in youth, underscoring the need for more refined prediction models (Vangeepuram et al., 2021). Similarly, McDonough et al. utilized an Ensemble Integration framework, a machine learning approach that integrates domain-specific models, to build a predictive model of youth prediabetes/diabetes status and identify the most significant variables associated with the condition (McDonough et al., 2023). Their work uncovered less common risk factors, highlighting the need for alternative data analysis methods to explore additional factors potentially contributing to diabetes development (McDonough et al., 2023).

Additionally, similar studies have been performed to compare the performance of multiple machine learning algorithms to predict diabetes in patients in their unique, respective datasets. One such research compared the performance of five machine learning models (decision tree, k-nearest neighbors [KNN] clustering, random forest [RF], support vector machine [SVM], and XGBoost) to test their performance in predicting type-2 diabetes mellitus (Jiang & Li, 2024). The conclusion from the study was that their tested linear model (logistic regression) had the most favorable performance, measured by the highest receiver operating characteristic (ROC) curve's area under the curve (AUC; Jiang & Li, 2024). In another study, data from the University of California Irvine machine learning repository was analyzed to test various machine learning models to predict diabetes given specific variables (Kibria, et al., 2021). The conclusion from that study as well was that the logistic regression model had the most favorable performance (Kibria, et al., 2021). Yet another study tested the performance of logistic regression, SVM, RF, and XGBoost to predict diabetes in medical records collected by the Hanaro Medical foundation in Seoul, South Korea (Deberneh & Kim, 2021). Although the study identified the logistic regression model as the best performing model, it was not by a large margin. Furthermore, Tasin et al. also found that the logistic regression model performed almost as equally as other models (RF, bagging, and SVM), in terms of the precision and recall scores (Tasin, et al., 2022).

Given these research gaps, our primary research question examines the key predictors of early-onset diabetes (diagnosed before age 40) compared to late-onset diabetes, and evaluates which regression models demonstrate superior performance in predicting the risk of both early-onset and late-onset diabetes.

Specifically, we intend to compare the performance of the logistic regression and random forest classification models between the age groups. The scikit-learn library will be primarily used to fit these models, as they are an accessible and open source tool for building predictive analytical models (Pedregosa, et al., 2011). Additionally, it provides useful functions for performance enhancements and analyzing results. In order to ensure that the model comparisons are as fair as possible, the same dataset will be used amongst all analysts/approaches, but each approach will optimize the results uniquely to the analyst and model. Our prior expectation for how the study will turn out, based on previous studies and general knowledge, is that the logistic regression model will outperform the random forest classification model in both age groups. This is because in the previous studies explored, although many models performed at comparable levels, the logistic regression model was chosen as the best predictive model by narrow margins in multiple study results. Furthermore, the logistic regression model is much more interpretable, and easier to implement. This might affect our study's results, as it takes less computational time to fit and optimize a linear model than a random forest model. A rebuttal to this predictive outcome could be that the random forest model performs better due to its ability to model complex relationships in wide datasets and its innate characteristic to reduce overfitting in data. However, due to analysis being limited by the hardware being used, these innate advantages of random forests might not be readily available at their fullest potential for this study.

## 3. Analysis

We studied the effectiveness of two classification models on two age groups: <40 and ≥40 years. We tested the performance of the best models identified in previous studies to see if one model consistently performed better across age groups. The project was divided into four tasks, with each team member responsible for building a model and evaluating its predictive ability on their assigned age group.

First, we merged the demographics, examination, laboratory, and questionnaire datasets from the latest National Health and Nutrition Examination Survey (NHANES; August 2021-August 2023) into one dataset. We conducted exploratory analysis to understand the data structure and summarized statistics, assessed assumptions, and created plots for visualization. Then, we splitted the dataset into <40 vs. ≥40 years for data analysis. The following analyses were completed to compare model performances and attain results.

Our approaches were building on established work in diabetes risk prediction using generalized linear models. We aimed to contribute to this field by building risk prediction models using various predicting variables that might improve the predictability of the model. While previous studies evaluated overall model performance on complete datasets, our approach uniquely examines prediction consistency across age-stratified groups. By analyzing distinct age cohorts separately, we could identify age-specific patterns previously overlooked, ultimately improving predictive precision.

*NHANES Dataset Overview:*

NHANES includes a broad set of datasets collected from August 2021-2023. These datasets include Demographic dataset, Dietary dataset, Examination dataset, Laboratory dataset, Questionnaire dataset. For the purpose of this research, Dietary dataset was excluded, and 4 other datasets were merged. Table DIQ_L.xpt which contains information about the status of patients with diabetes or no diabetes were used to establish the dependent variable. Furthermore, GHB_L.xpt and GLU_L.xpt tables which directly have indicators for diabetes measures were excluded. After merging the datasets, rows with more than 20% and columns with more than 5% missing values were removed. The resulting dataframe contained 88 fixed features. In the next step, missing values were imputed with their median feature value and collinearity was examined. VIF with threshold of 5 was used to drop multicollinear features. The resulting dataframe contained 64 features. In the next step, LassoCV was performed to show the importance and strength of selected features in predicting whether a participant in the survey had diabetes (Figure 1). The optimum lambda was achieved at 0.0023, MSE of 0.0899 and $R^2$ of 0.2121. The resulting dataframe by filtering the features with their absolute value of their coefficient greater than 0.0001 contained 25 features and 6582 samples. The description of the selected features are provided in Table 1.

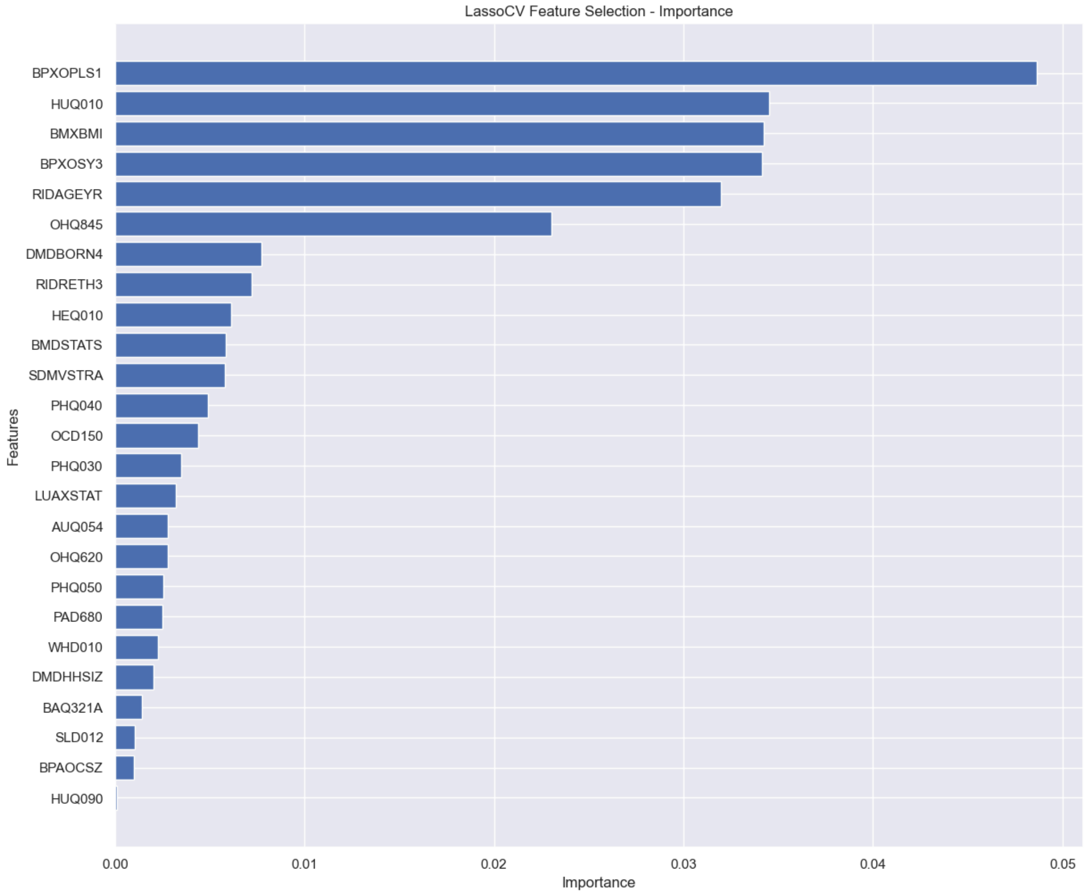


**Figure 1.** Features Selected by LassoCV vs Their Importance Predicting Diabetes for Patients of All Ages from NHANES Dataset.

**Table 1.** Characteristics of Selected Features.

| **Table Name** | **Data File** | **Description** |
|---|---|---|
| AUQ054 | Questionnaire Data | General condition of hearing |
| BAQ321A | Questionnaire Data | Past 12 months, problems with vertigo |
| HEQ010 | Questionnaire Data | Hepatitis B |
| HUQ010 | Questionnaire Data | General health condition |
| HUQ090 | Questionnaire Data | Seen mental health professional/past yr |
| OCD150 | Questionnaire Data | Type of work done last week |
| OHQ845 | Questionnaire Data | Rate the health of your teeth and gums |
| OHQ620 | Questionnaire Data | How often last yr. had aching in mouth? |
| PAD680 | Questionnaire Data | Minutes sedentary activity |
| SLD012 | Questionnaire Data | Sleep hours - weekdays or workdays |
| WHD010 | Questionnaire Data | Current self-reported height (inches) |
| PHQ030 | Laboratory Data | Alcohol |
| PHQ040 | Laboratory Data | Gum, mints, lozenges or cough drops |

| | | |
|---|---|---|
| PHQ050 | Laboratory Data | Antacids, laxatives, or anti-diarrheals |
| BMDSTATS | Examination Data | Body Measures Component Status Code |
| BMXBMI | Examination Data | Body Mass Index (kg/m**2) |
| BPAOCSZ | Examination Data | cuff size - oscillometric |
| BPXOSY3 | Examination Data | Systolic - 3rd oscillometric reading |
| BPXOPLS1 | Examination Data | Pulse - 1st oscillometric reading |
| LUAXSTAT | Examination Data | Elastography exam status |
| RIAGENDR | Demography | Gender |
| RIDRETH3 | Demography | Race/Hispanic origin w/ NH Asian |
| DMDBORN4 | Demography | Country of birth |
| DMDHHSIZ | Demography | Total number of people in the Household |
| SDMVSTRA | Demography | Masked variance pseudo-stratum |

### *Approach #1: Logistic Regression in <40 years (Sarah Park)*

*Methods:*

This study used a machine learning approach to predict diabetes risk in individuals under 40 years old using the master dataset described above. Participants that were pre-diabetic were excluded, and features with more than 5% missing values and observations with more than 20% missing values were excluded. The missing values were imputed using median. The diabetes outcome variable was derived from the DIQ010 field, which indicates diabetes status. Age stratification was performed to focus specifically on the under-40 population, addressing early-onset diabetes prediction. Prior to fitting models, exploratory analysis was conducted and assumptions were tested (the results are not shown here but the relevant code can be found in github).

A LassoCV approach with 10-fold cross-validation was implemented to select the most predictive features while minimizing multicollinearity. The feature selection pipeline incorporated median imputation for missing values and standardization of features. To address class imbalance during feature selection, balanced sample weights were applied. The optimal regularization parameter (alpha = 0.0431) was determined by minimizing mean squared error across cross-validation folds. From the initial variables, 10 features with non-zero coefficients were retained for the final model.

Following feature selection, a logistic regression model was developed using the selected features. The dataset was split into training (80%) and test (20%) sets using stratified sampling to maintain class distribution. Missing values were imputed using median imputation, and features were standardized using the training set parameters. To address the substantial class imbalance (2.15% diabetes prevalence), Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set, creating a balanced dataset with equal class proportions. The logistic regression model was configured with balanced class weights, a maximum of 1000 iterations, and L2 regularization (C = 1.0).

Model performance was assessed using 5-fold stratified cross-validation on the SMOTE-resampled training data. Evaluation metrics included accuracy, precision, recall, and F1-score. The model was then validated on the original imbalanced test set to assess real-world performance. To optimize classification performance, probability thresholds ranging from 0.1 to

0.9 were systematically evaluated against precision, recall, and F1-score metrics. The threshold maximizing F1-score (0.9) was selected as optimal for final model evaluation. Comprehensive performance assessment included confusion matrix analysis, ROC curve with AUC calculation, and precision-recall curve with average precision score. Feature importance was quantified using the absolute values of the logistic regression coefficients, providing insight into the relative contribution of each predictor variable.

*Results:*

The dataset showed significant class imbalance, with approximately 2.15% of samples being positive for diabetes in the under-40 age group (not shown). Age distribution analysis revealed that approximately one-third of the studied population was under 40 years old (not shown). Correlation analysis identified numerous highly correlated features (r > 0.7), particularly among laboratory measurements, necessitating dimensionality reduction.

The LassoCV procedure with 10-fold cross-validation identified an optimal alpha value of 0.0431 (Figure 2), resulting in the selection of 10 features from the original variables. Body mass index (BMXBMI) was the most important predictor, followed by blood pressure (BPXODI3) and prescription medication use (RXQ033; Figure 3).

Prior to fitting a logistic regression, class imbalance was addressed through SMOTE. Five-fold cross-validation on the training set demonstrated robust performance with an accuracy of 0.866, precision of 0.861, recall of 0.874, and F1-score of 0.867. However, initial performance on the imbalanced test set showed substantial discrepancy, with high accuracy (0.847) but low precision (0.090) and F1-score (0.158), though maintaining reasonable recall (0.667) and excellent ROC AUC (0.90, Figure 4).

Threshold optimization analysis across probability thresholds from 0.1 to 0.9 identified 0.9 as the optimal threshold, maximizing F1-score at 0.214. With this threshold, final test set performance achieved accuracy of 0.947, precision of 0.158, and recall of 0.333. The precision-recall curve (Figure 5) demonstrated average precision of 0.18, substantially above the no-skill baseline (0.0215), while the confusion matrix (Figure 6) revealed 3 true positives, 16 false positives, 393 true negatives, and 6 false negatives at the optimized threshold. Body mass index (BMXBMI) was the most important positive predictor in both magnitude of coefficient (0.794) and relative importance for the logistic regression model (Figure 7). Other significant positive predictors included diastolic blood pressure (BPXODI3, 0.768), general health condition (HUQ010, 0.517), and pulse rate (BPXOPLS1, 0.506). Prescription medication use (RXQ033) was the strongest negative predictor with a coefficient of -0.851.

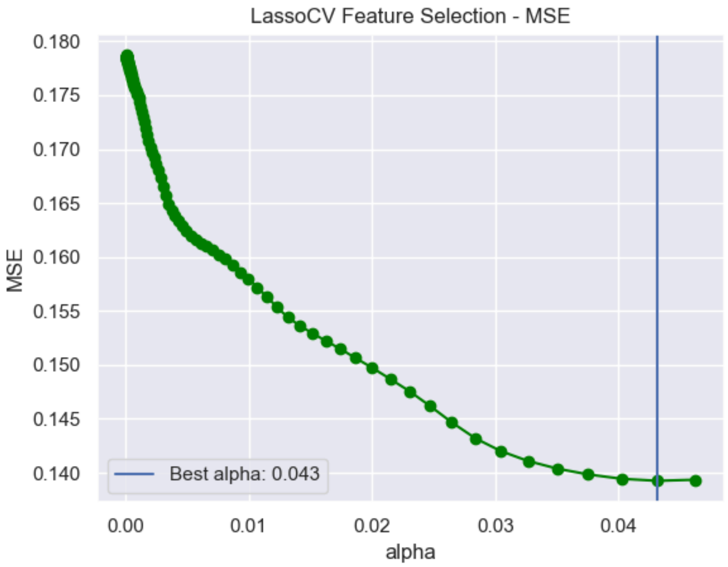


**Figure 2.** Mean Squared Error by Regularization Parameter (Alpha) in LassoCV Feature Selection.

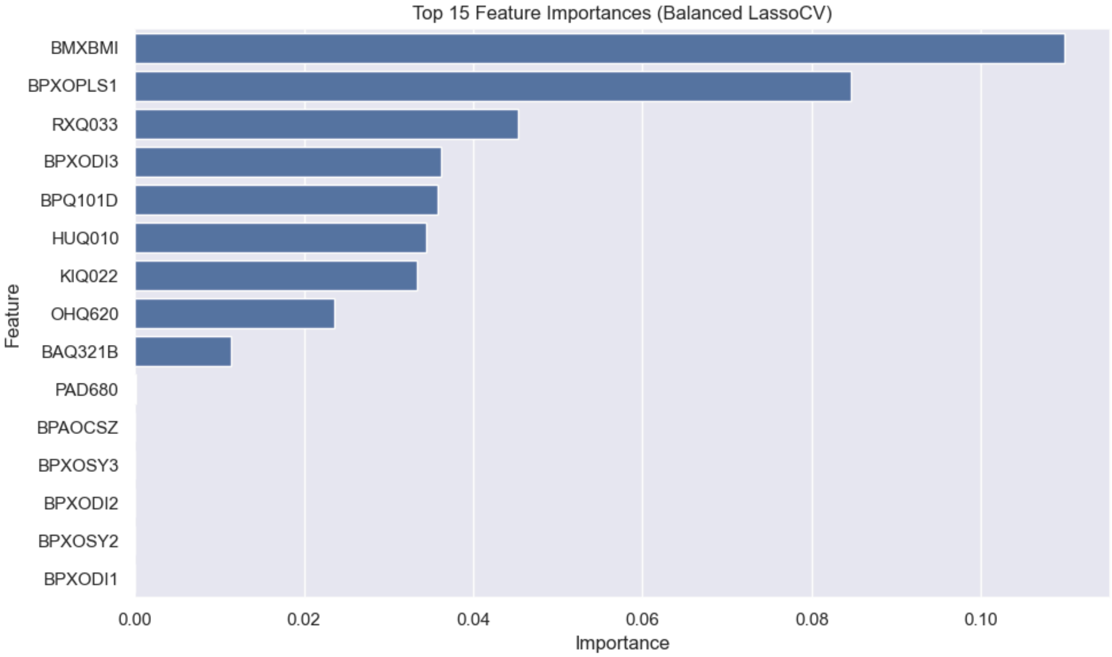


**Figure 3.** Top 15 Feature Importances from Balanced LassoCV Selection for Diabetes Prediction.

BAQ321B = Symptoms of dizziness, light-headedness, or balance problems; BMXBMI = Body mass index; BPQ101D = Taken medication to lower blood cholesterol; BPXODI3 = Diastolic - 3rd oscillometric reading; BPXQPLS1 = Blood pressure; HUQ010 = General health condition; KIQ022 = Weak/failing kidneys; OHQ620 = Frequency of painful aching in the mouth during the last year; RXQ033 = Taken prescription medicine.

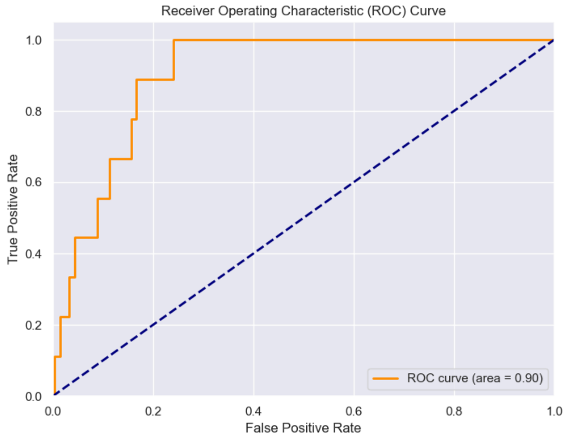


**Figure 4.** Receiver Operating Characteristic (ROC) Curve for Early-Onset Diabetes Prediction Model (AUC = 0.90).

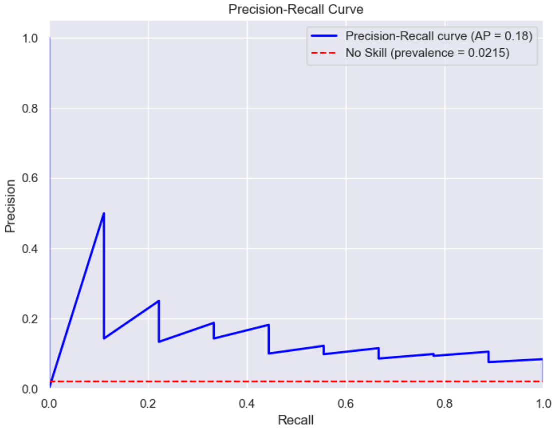


**Figure 5.** Precision-Recall Curve for Early-Onset Diabetes Classification Model (Average Precision = 0.18).

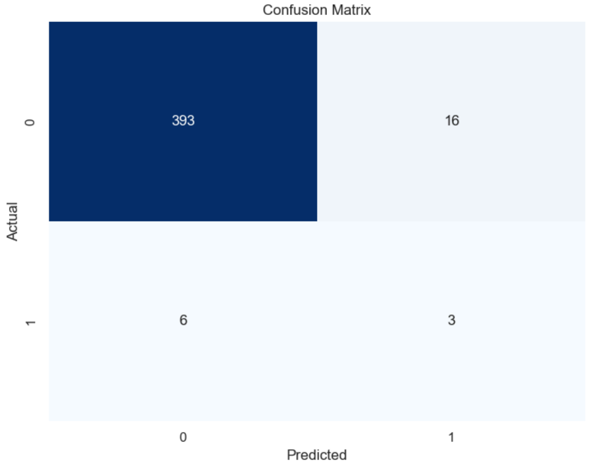


**Figure 6.** Confusion Matrix of Optimized Logistic Regression Model at 0.90 Probability Threshold.

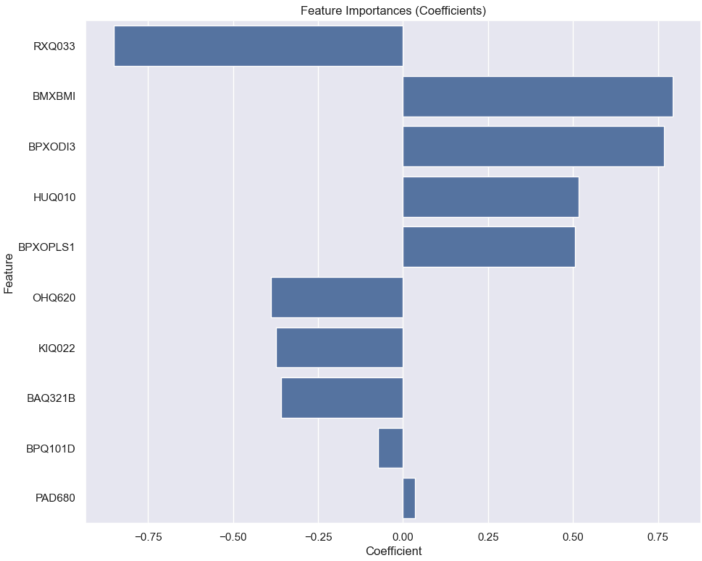


**Figure 7.** Feature Importance Coefficients from Logistic Regression Model for Early-Onset Diabetes Prediction.

BAQ321B = Symptoms of dizziness, light-headedness, or balance problems; BMXBMI = Body mass index; BPQ101D = Taken medication to lower blood cholesterol; BPXODI3 = Diastolic - 3rd oscillometric reading; BPXQPLS1 = Blood pressure; HUQ010 = General health condition; KIQ022 = Weak/failing kidneys; OHQ620 = Frequency of painful aching in the mouth during the last year; PAD680 = Minutes sedentary activity; RXQ033= Taken prescription medicine.

*Discussion:*

The findings from this study provide valuable insights into early-onset diabetes prediction in individuals under 40 years of age using machine learning approaches. Body mass index was found to be the strongest positive predictor, consistent with established literature identifying obesity as a primary risk factor for early-onset diabetes (Jiang & Lai, 2024; Chandrasekaran et al., 2024). The significant association between diastolic blood pressure and diabetes risk aligns with previous research demonstrating that hypertension often precedes diabetes diagnosis in younger populations (James et al., 2023). Our model identified prescription medication use (RXQ033) as a strong negative predictor, suggesting potential protective effects of certain medications or reflecting healthcare-seeking behaviors that might mitigate diabetes risk. The identification of pulse rate as a significant predictor supports emerging evidence linking autonomic nervous system function to diabetes pathophysiology (Grisanti, 2018).

The substantial class imbalance observed (2.15% diabetes prevalence) reflects the epidemiological reality of early-onset diabetes but presented significant methodological challenges. Despite robust cross-validation performance after SMOTE application, the marked discrepancy in test set performance highlights the limitations of synthetic sampling techniques in highly imbalanced real-world datasets. The high ROC AUC (0.90) demonstrates good discriminative ability, but the low precision and F1-scores even after threshold optimization underscore the difficulty in developing clinically applicable early-onset diabetes prediction models. Our findings regarding optimal threshold selection (0.90) suggest that in screening contexts where false positives carry substantial resource implications, highly conservative probability thresholds may be necessary. However, this comes at the cost of reduced sensitivity

(0.333), potentially missing two-thirds of early-onset cases. This precision-recall tradeoff represents a significant clinical challenge requiring careful consideration in implementation.

Several limitations warrant consideration. First, the cross-sectional nature of NHANES data precludes assessment of temporal relationships between identified predictors and diabetes development. Second, despite comprehensive feature selection, potentially important genetic and environmental factors were unavailable for analysis. Third, while NHANES provides nationally representative data, sample sizes for early-onset diabetes were limited, affecting model stability. Future research should explore ensemble approaches that may better handle extreme class imbalance and incorporate longitudinal data to establish temporal relationships. Integration of genetic risk scores and environmental exposures could enhance predictive performance. Additionally, validation in diverse populations would strengthen generalizability, particularly given the increasing incidence of early-onset diabetes in specific ethnic groups.

In conclusion, this study identified key predictors of early-onset diabetes and demonstrated both the promise and limitations of machine learning approaches in this context. While achieving perfect predictive performance remains challenging, even incremental improvements in early identification could substantially impact public health through targeted prevention efforts.

### *Approach #2: Random Forest in <40 years (Ye Jun Kim)*

*Methods:*

The master dataset was used to perform analysis on the predictive capabilities of a random forest  model on patients younger than 40 years. A random forest model could have a more optimal feature selection method than a logistic regression, so multiple runs were tested to ensure the highest predictive capacity. Firstly, the master dataset was filtered by records of patients under 40 years of age. Secondly, as NHANES stores the diabetes status as multiclass, the "borderline" value was removed from the dataset, so that there was a clear binary response of "0" being classified as "patient has no diabetes" and "1" being classified as "patient has diabetes". Thirdly, since a Random Forest model cannot run with the existence of NaN values in the dataset, median imputation was utilized in columns with NaN values. Finally, the resulting clean table was then split into training and testing predictor and response datasets in a 80/20 ratio. It was immediately apparent that there was an immense disproportion when it cam   e to the response values, with about 97.9% of the response values being 0, classified as "non-diabetic". Surely with such a disproportion on a supervised machine learning model, the resulting predictive model would heavily favor predictions showing no diabetes in patients, meaning there would be inevitable bias in the results of the model. In order to address this, the training and testing datasets in each cycle of analysis were split into a 50/50 ratio of response values 0/1. Then a Random Forest model was built using the training sets, and the resulting model's predictive capabilities were measured using the testing datasets.

The Random Forest Classifier results attained through the scikit-learn library has a built-in feature selection function to quantify the importance of the features used in the model. It utilizes various internal calculations, including the Gini importance and mean decrease in accuracy metrics, to measure impurity decrease within each decision tree (Pedregosa, et al., 2011). Through this function unique to random forest models, the importance of features were quantified and visualized. The features that were reduced to zero importance during the random foresting were then removed from the predictor dataset, and a secondary model was created with the remaining predictor variables. This approach was taken again for a third cycle, until no features were reduced to zero during random foresting.

*Results:*

According to the predictive performance metrics used, the initial model had an accuracy of approximately 69.4%, with an Out-of-Bag (OOB) score of 98.2%. The OOB score is a common method used to evaluate the performance of ensemble models such as random forests. It is calculated by using data points that were not included in the training of the model, hence “out-of-bag” samples. Although the resulting model boasted a high OOB score, the accuracy and the R-squared results certainly had room for improvement. The calculated pseudo R-squared value was -0.22. As mentioned above, feature selection was measured by the built-in function for Random Forests through scikit-learn. The features that were reduced to zero during random foresting were removed from the datasets, and modeled again. This secondary random forest results had slightly higher accuracy of 73.2%, and a slightly lower OOB score of 98.0%. The pseudo R-squared value was still extremely low at a value of -0.07. Finally, the same process was repeated after dropping the features that was reduced to zero during the second model. This would be the final time it was needed, as no more features were reduced to zero after the third model was built. The third model had an accuracy of 71.6% and an OOB score of 98.3%. The pseudo R-squared value was -0.14, which still left a lot to be desired.

The figures below outline the results of the Random Forest model in the “<40 years” age group. When feature importance was visualized on the third model, no features were identified to have reduced to zero. The example decision trees shown for the third model showed less complexity than the first, due to the removal of predictors that were deemed unimportant by the models. The most important result of interest however is the confusion matrix and the ROC curve of the resulting final model. The confusion matrix was built through prediction values of the subsampled testing dataset. Due to the abundance of “non-diabetic” records in the dataset, the model is extremely good at identifying those without diabetes. However, when predicting patients with diabetes, where the response value is 1, the performance considerably decreases. However, according to the probability of predictions visualized through the ROC curve, the model is extremely good at discriminating between classes. This discrepancy between the accuracy of the predictions themselves and the probability of predictions are likely due to the fact that the ROC calculations heavily favor imbalance datasets, and orders results using a different metric than an accuracy calculation. The high ROC AUC in this case represents the fact that the classifier is extremely good at identifying when the response value is 0 at the cost of high error when response value = 1. This is also reflected by the confusion matrix as well.

Further consideration and research must be explored in order to clearly outline the full predictive capabilities of random forest models in predicting diabetes in NHANES data. Due to the existence of largely disproportionate response values, the predictive capabilities of the model were also impacted, since data had to be manually filtered and altered for the model to make predictions. If the training and testing datasets were not split with a 50/50 even split of the binary response values, there would be an incredibly disproportional dataset where the response value is zero. Additionally, due to limited computational capacity, the random forest was run after pre-selection of some features and data preparation, instead of starting with all features and reducing the variables through their importance.

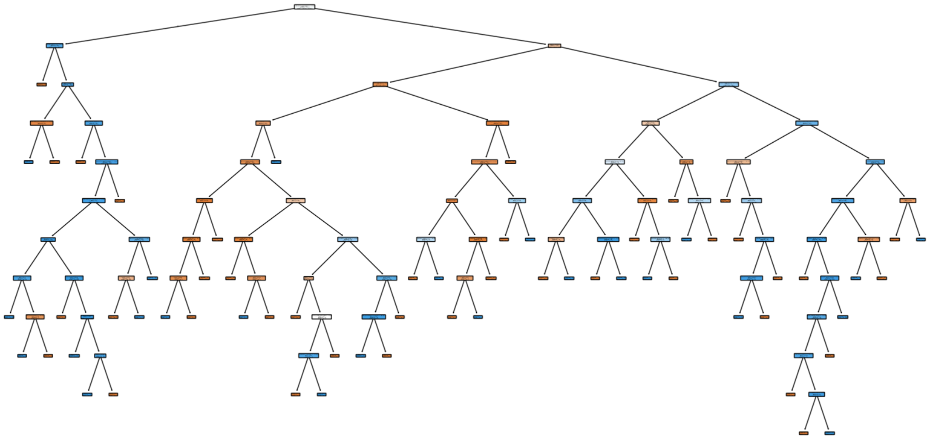


**Figure 8.** Random Forest < 40 years, Example Decision Tree.

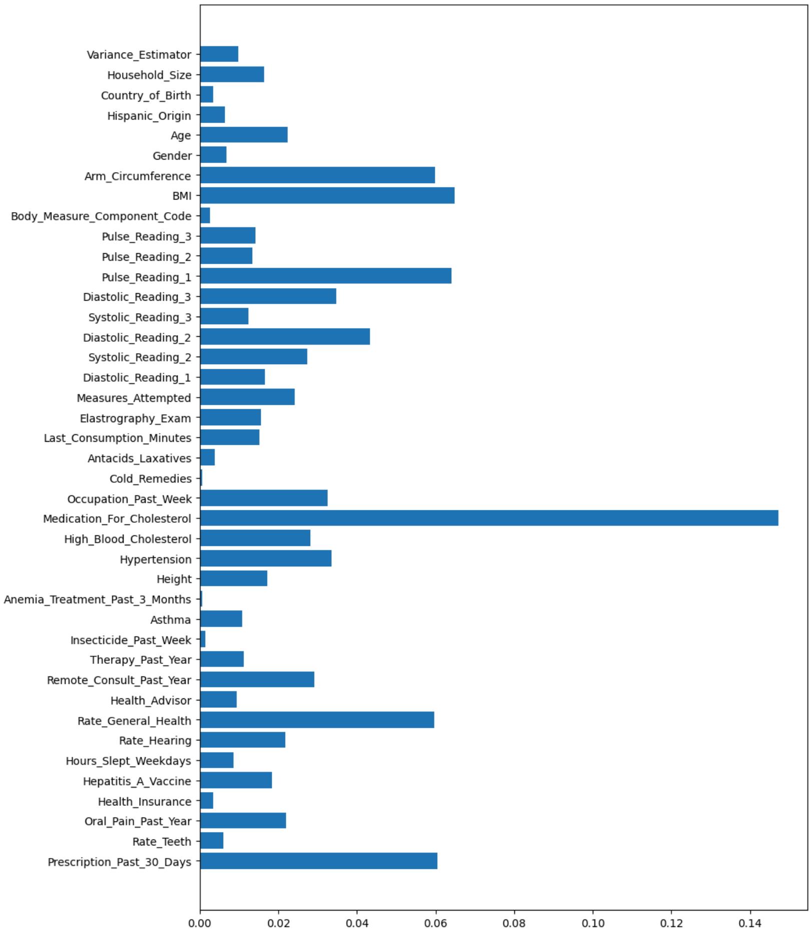


**Figure 9.** Feature Importance Visualization from Random Forest Model <40 years.

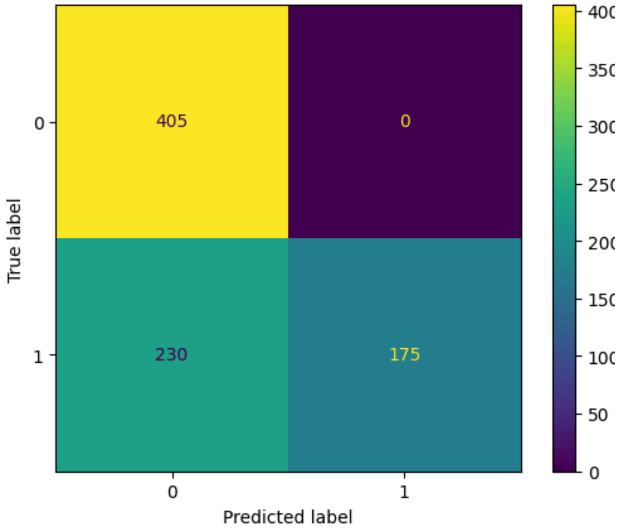


**Figure 10.** Random Forest <40 years Confusion Matrix Result.

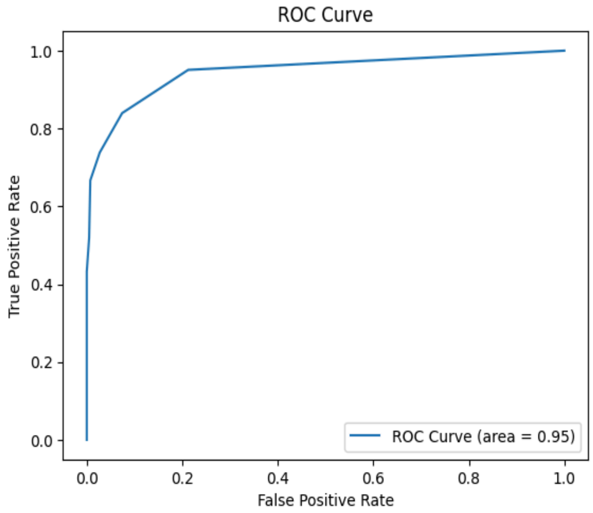


**Figure 11.** ROC Curve and AUC of Random Forest <40 years

*Approach #3: Logistic Regression in ≥40 years (Sihan Wang):*

*Method:*

Master dataset was created by joining multiple xpt files by using SEQN identifier from 4 main NHANES directories: Questionnaire Data, Laboratory Data, Examination Data, and Demographic Data. Value 0 represents not diagnosed with diabetes and 1 is with diabetes. Dataset was cleaned by removing duplicated rows and columns, also rows and columns were removed with a number of missing values above threshold (5% for columns and 20% for row). A small amount of data indicating borderline/prediabetes were removed from the dataset due to a convergence issue and our target model is binary classification. Subdataset filtered to include only patients aged 40 years and older was created for logistic regression analysis.

Variance Inflation Factor was implemented to evaluate multicollinearity. Before doing so, missing values were imputed using the median strategy, and features were standardized. At threshold 5, there are 18 features with VIF that are above threshold. Those features were removed from the dataset to mitigate multicollinearity.

Dataset was then split into training and testing subsets using an 80/20 stratified split. LogisticRegressionCV was used with L1 penalty and negative log loss as scoring for feature selection. Missing value, data scaling were handled inside the pipeline. Due to imbalanced data, a balanced class weight parameter was used to provide higher weight to the minority class, in this case, it was diabetes diagnosed. I also considered SMOTE and ADASYN to compare results with only using LogisticRegressionCV build-in feature for imbalanced data, model performance was very similar.

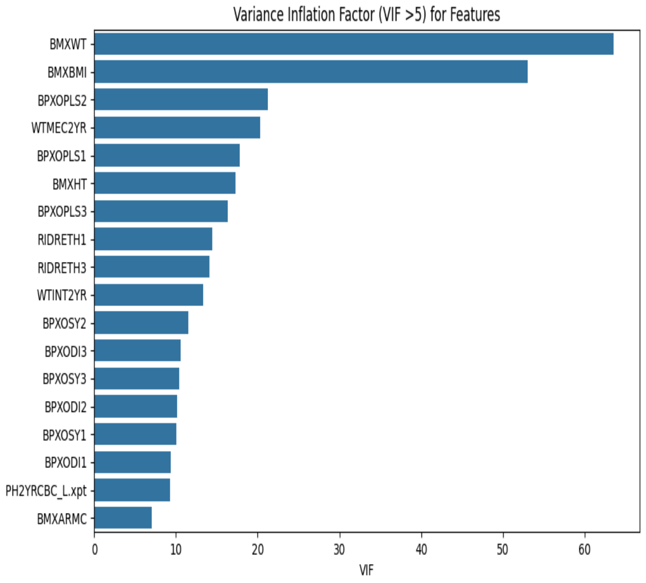


**Figure 12.** Features with VIF Above Threshold.

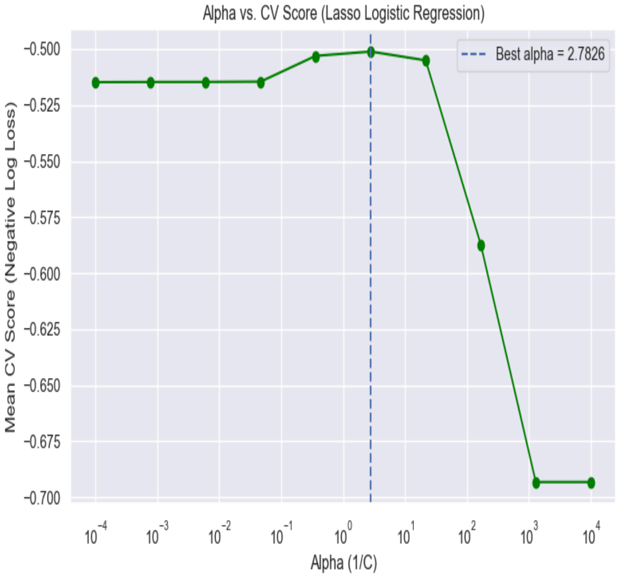
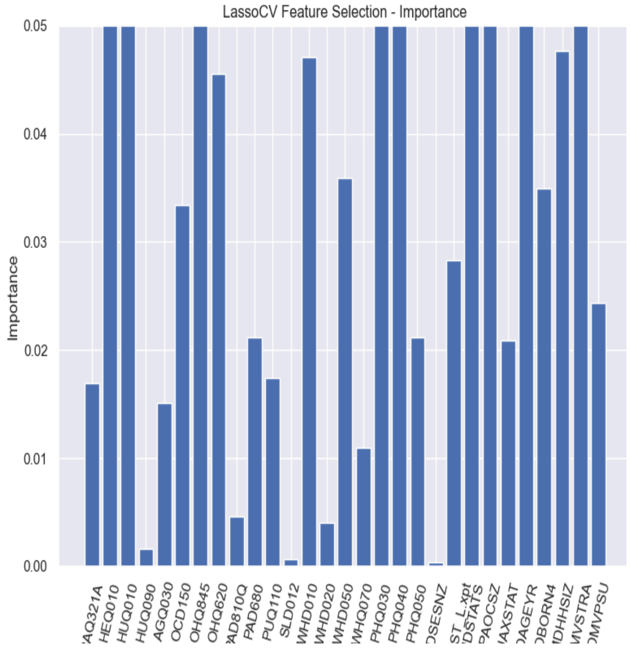


**Figure 13.** Alpha vs. Mean Cross-Validation Score for Logistic Regression Lasso.

*Model performance:*

Upon fitting the model, the optimal regularization parameter C was selected based on the mean cross-validation score measured by negative log loss. The inverse of C was interpreted as the model's alpha, indicating the level of penalization applied to the coefficients. Performance on the test set was evaluated. The classification report revealed precision, recall, and F1 scores for each class. However, precision and f1 score were not as high as expected. The model achieved a reasonable ROC AUC score as 0.79, indicating the model is able to distinguish between classes to a certain degree. A confusion matrix was plotted to visually inspect the classification outcomes, and the ROC curve further illustrated the model's ability for classification. Feature coefficients were scaled back for interpretability.



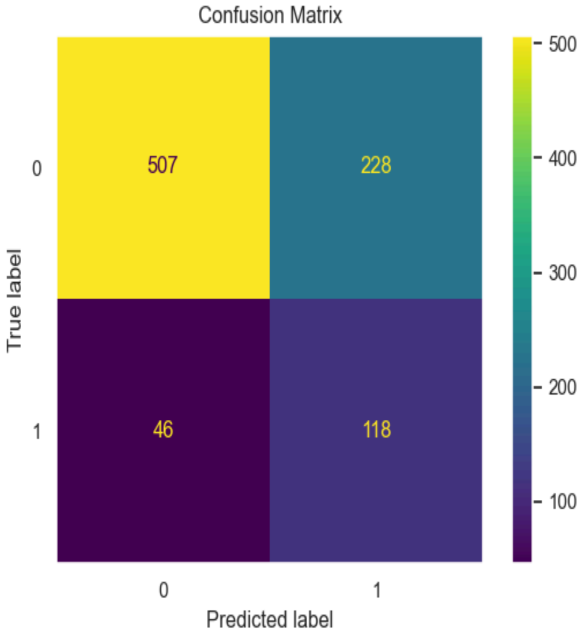


**Figure 14.** Logistic Regression Age≥40 Feature Importance Visualization.

**Figure 15**. Logistic Regression Age≥40 Confusion Matrix.

**Table 2.** Classification Report.

| | **Precision** | **recall** | **f1** |
|---|---|---|---|
| **0** | 0.92 | 0.69 | 0.79 |
| **1** | 0.34 | 0.72 | 0.46 |

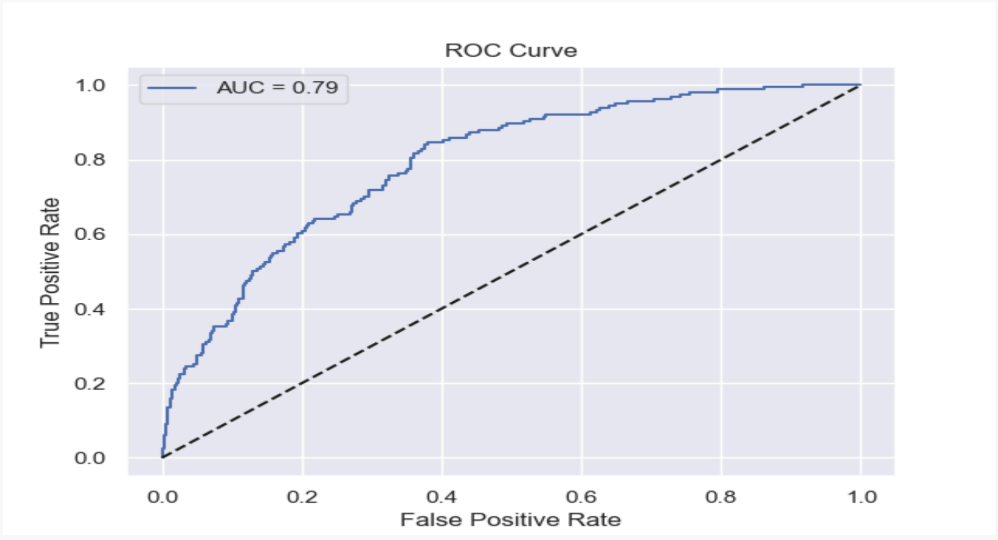


**Figure 16.** ROC Curve for Logistic Regression, Age≥40.

*Future work:*

In future research, we could explore the application of more sophisticated models, such as XGBoost, or deep learning approaches like neural networks that fit better for nonlinear relationships.

As for data processing and feature engineering, it is critical to explore advanced techniques such as SMOTE variance, particularly for datasets that contain both categorical and numerical variables. Furthermore, an evaluation of outliers should be conducted to mitigate noise. Applying interaction terms may also improve model performance by capturing underlying dependencies between features.

Instead of sticking to the conventional classification threshold of 0.5, it may be beneficial to explore alternative thresholds and assess their impact on model results.

*Approach #4: Decision tree/random forest Regression in >40 years (Yahya Hosseini):*

The sub-dataset containing information for the age group +40 year was obtained by splitting the dataset using the RIDAGEYR column. LassoCV with 10-fold cross-validation, and imputation of missing values with the column median was used for feature selection. General feature selection technique is the same as explained in NHANES dataset section. with optimum lambda achieving at 0.005 with MSE of 0.1188 and $R^2$ of 0.2037 (Figure 17, left image). Comparing these values with the LassoCV performance over the whole dataset, it seems that the prediction power of the LassoCV (although LassoCV is just used for feature selection), is slightly less on patients >40 years compared to patients of all ages. Figure 17, right image also shows the selected features vs their coefficients using LassoCV. Comparing the selected features with the features from the whole dataset, it can be observed that for this analysis, predictor Age has lost its significance from #5 to #7. Ignoring the predictor, Age, the top 5 significant predictors look the same for >40 year dataset and the entire dataset.

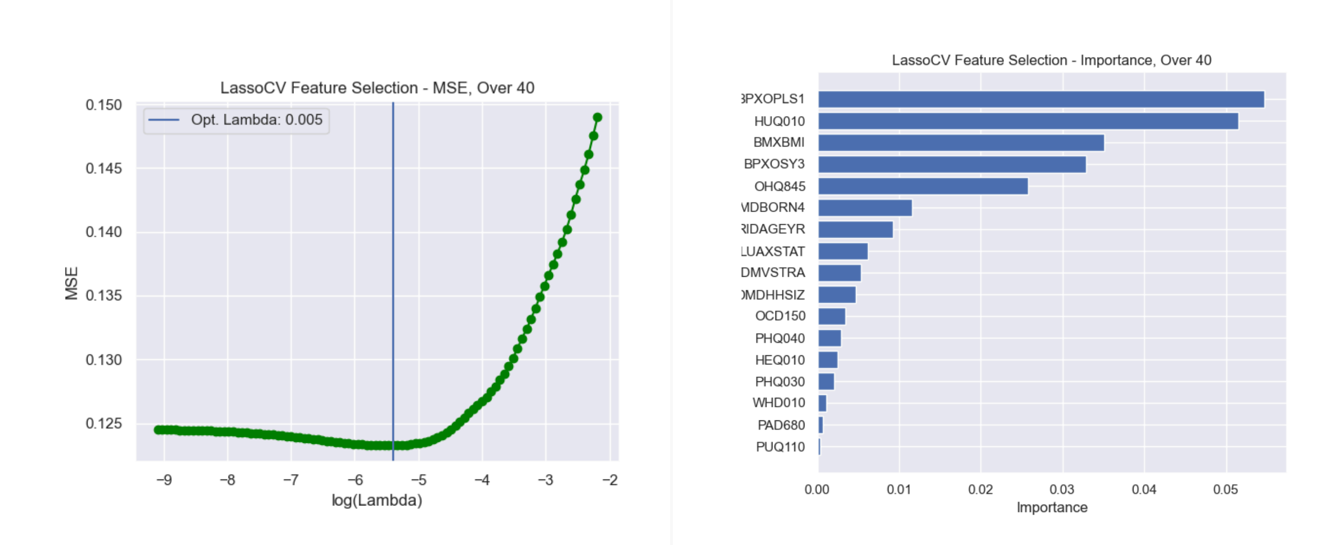


**Figure 17.** (Left) LassoCV MSE vs Lambda Using NHANES Dataset Patients Aged 40 Years Plus. (Right) Features Selected by LassoCV for NHANES Dataset Predicting the Strength of Features in Predicting Patients with Diabetes.

The Adaptive Synthetic Sampling technique (He, et. al, 2008) was used to overcome the imbalance issue between subjects with diabetes and no diabetes. In order to train the decision tree and random forest classifiers, the dataset was split 80/20 for training and testing. For both classifiers, GridSearchCV was used to run 5-fold cross-validation on the training dataset.
For the decision tree and random forest classifiers, the hyperparameters outlined in Table 3 were explored to achieve the best accuracy. For each classifier, the highlighted value indicates the optimum parameter for each category with which the best cross–validation accuracy was achieved.

**Table 3.** Hyperparameters for Decision Tree and Random Forest Classifiers.

| **Parameters** | **Decision Tree** | **Random Forest** |
|---|---|---|
| criterion | 'gini', **'entropy'**, 'log_loss' | 'gini', **'entropy'**, 'log_loss' |
| n_estimators | - | 50, **100**, 200 |
| max_depth | 10, **20**, 25 | 10, 20, **25** |
| min_sample_split | **3**, 5, 10 | **3**, 5, 10 |
| min_samples_leaf | 3, 5, **10** | **3**, 5, 10 |

Figure 18 shows the importance of features selected by the decision tree (left) and random forest (right) classifiers. Decision tree classifier has put significant weight on the HUQ10 (General health condition) parameter for prediction. This makes sense as the decision tree is a greedy classifier. It means it takes the best feature for splitting at root and puts significant weight for splitting. Since the decision tree consists of one single tree, HUQ10 has received much higher importance compared to other features. Random forest consists of many different random trees for classification. Since features are split randomly, all the features get the chance to be evaluated and averaged. Like the decision tree classifier, in the random forest HUQ10 (General health condition) parameter is the most important feature. However, it can be observed that its weight is less than decision tree classifier and other parameters have also received higher significance. One major difference between LassoCV and decision tree/random forest based classifiers is that BPXOPLS1 (Pulse - 1st oscillometric reading) is the feature with most significance in LassoCV but in the ensemble based classifiers, it is placed in rank 5-6.

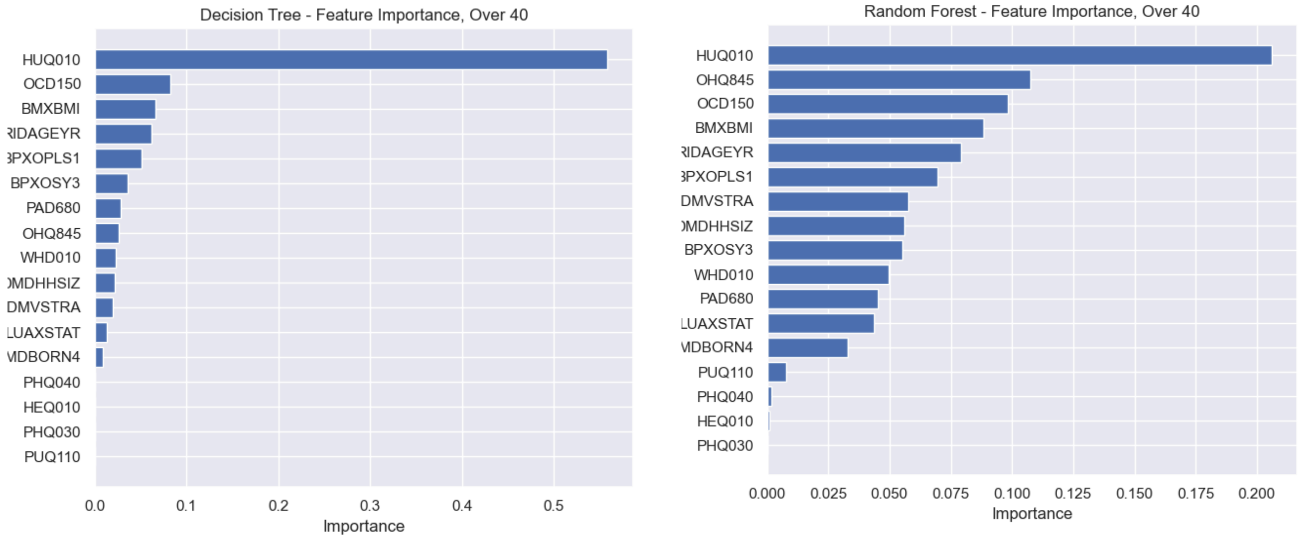


**Figure 18.** Features Selected by LassoCV for NHANES Dataset Predicting the Strength of Features in Predicting 40+ Age Patients with Diabetes, Using (Left) Decision Tree Classifier, (Right) Random Forest Classifier.

Table 4 shows the unweighted performance comparison between decision tree classifier vs random forest classifier. In all 5 metrics, the random forest classifier showed slightly better performance compared to the decision tree classifier. However, it seems both classifiers are underperforming and a classifier with non-linear kernel might perform better.

**Table 4.** Performance Comparison of Decision Tree and Random Forest Classifier.

| | **Accuracy** | **Precision** | **Recall** | **f1-score** | **ROC AUC** |
|---|---|---|---|---|---|
| **Decision Tree** | ***0.75*** | ***0.60*** | ***0.62*** | ***0.61*** | ***0.62*** |
| **Random Forest** | ***0.82*** | ***0.70*** | ***0.63*** | ***0.65*** | ***0.63*** |

## 4. Explanation of Changes

Dietary dataset was not included in this study because of the size and complexity of the dataset. This dataset included many duplicate rows for the same user with a few columns with a wide range of categorical data. Expanding them to dummy variables wouldn't have made sense. We used other 4 datasets with which we still complied to using at least 3 minimum datasets. The NHANES dataset section shows that even after feature selection, we are using 4 datasets. SMOTE, ADASYN techniques and balanced class weight setting were used to handle the imbalanced dataset.

## 5. Conclusions

Our study demonstrates that predictive modeling approaches can effectively differentiate between early-onset and late-onset diabetes risk factors using NHANES data. Both age cohorts showed distinct predictive patterns, with early-onset diabetes models achieving notably higher performance (ROC AUC > 0.90) compared to late-onset models (ROC AUC 0.62-0.79). This performance discrepancy suggests potentially different pathophysiological mechanisms between age groups and highlights the value of age-stratified approaches to diabetes risk prediction. Body mass index was consistently the strongest predictor across all models, while important age-specific differences in secondary predictors were identified that could inform targeted screening strategies. These findings contribute to our understanding of early-onset vs. late-onset diabetes risk factors and provide a foundation for developing more precise risk assessment tools for different age groups.

A key strength of this study was the successful handling of messy real-world data through systematic preprocessing, including addressing missing values, removing duplicated columns and rows, and mitigating multicollinearity through variance inflation factor analysis and feature selection techniques. We effectively addressed the challenge of class imbalance using multiple approaches including SMOTE, ADASYN techniques, and balanced class weight settings, which substantially improved minority class prediction despite the extremely low prevalence. Both modeling approaches provided complementary insights, with logistic regression providing better interpretability through coefficient analysis while random forest captured complex non-linear relationships. For early-onset diabetes, the random forest model resulted in higher ROC AUC compared to the logistic regression model (0.95 vs. 0.90), while it was the opposite for late-onset diabetes (0.63 vs. 0.79). These models achieved reasonable performance across metrics, providing a solid baseline for future advanced modeling techniques.

Several limitations affect the interpretation of the findings. Despite our efforts, model precision remained suboptimal, particularly in the early-onset group, reflecting the inherent difficulty of achieving high precision with extremely imbalanced datasets. The feature set was limited to available NHANES variables, excluding potentially important genetic, environmental, and dietary data that could improve prediction accuracy. While we applied several techniques to address class imbalance, our models remained sensitive to this challenge, suggesting that more sophisticated machine learning approaches might yield better results. Our use of median imputation, while practical, may have oversimplified the missing value patterns in the data. Additionally, the synthetic oversampling techniques used (SMOTE/ADASYN) introduce potential overfitting risks by creating artificial data points that may add noise to the models. Future work should explore alternative approaches such as undersampling the majority class or applying SMOTE/ADASYN inside cross-validation loops to mitigate these concerns. Longitudinal data would also help establish temporal relationships between identified predictors and diabetes development, strengthening causal inference.

## 6. References

August 2021-August 2023 Demographics Data - Continuous NHANES. (2021). Cdc.gov. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&Cycle=2021-2023

August 2021-August 2023 Laboratory Data - Continuous NHANES. (2021). Cdc.gov. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory&Cycle=2021-2023

August 2021-August 2023 Questionnaire Data - Continuous NHANES. (2021). Cdc.gov. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2021-2023

CDC. (2024, May 31). *Diabetes Basics*. Diabetes. https://www.cdc.gov/diabetes/about/index.html#cdc_disease_basics_res-resources

Centers for Disease Control and Prevention. (2024, May 15). *National diabetes statistics report*. CDC. https://www.cdc.gov/diabetes/php/data-research/index.html

Chandrasekaran, P., & Weiskirchen, R. (2024). The Role of Obesity in Type 2 Diabetes Mellitus—An Overview. International Journal of Molecular Sciences, 25(3), 1882. https://doi.org/10.3390/ijms25031882

Cioana, M., Deng, J., Hou, M., Nadarajah, A., Qiu, Y., Chen, S. S. J., Rivas, A., Banfield, L., Chanchlani, R., Dart, A., Wicklow, B., Alfaraidi, H., Alotaibi, A., Thabane, L., & Samaan, M. C. (2021). Prevalence of Hypertension and Albuminuria in Pediatric Type 2 Diabetes: A Systematic Review and Meta-analysis. *JAMA Network Open*, *4*(4). https://doi.org/10.1001/JAMANETWORKOPEN.2021.6069

Cveticanin, L., & Arsenovic, M. (2025). Prediction Models for Diabetes in Children and Adolescents: A Review. *Applied Sciences 2025, Vol. 15, Page 2906*, *15*(6), 2906. https://doi.org/10.3390/APP15062906

Dabelea, D., Stafford, J. M., Mayer-Davis, E. J., D'Agostino, R., Dolan, L., Imperatore, G., Linder, B., Lawrence, J. M., Marcovina, S. M., Mottl, A. K., Black, M. H., Pop-Busui, R., Saydah, S., Hamman, R. F., Pihoker, C., Koebnick, C., Reynolds, K., Holmquist, K., Li, X., … Pierce, J. (2017). Association of Type 1 Diabetes vs Type 2 Diabetes Diagnosed During Childhood and Adolescence With Complications During Teenage Years and Young Adulthood. *JAMA*, *317*(8), 825–835. https://doi.org/10.1001/JAMA.2017.0686

Deberneh, H. M., & Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *International Journal of Environmental Research and Public Health*, *18*(6), 3317. https://doi.org/10.3390/ijerph18063317

Eppens, M. C., Craig, M. E., Cusumano, J., Hing, S., Chan, A. K. F., Howard, N. J., Silink, M.,

& Donaghue, K. C. (2006). Prevalence of diabetes complications in adolescents with type 2 compared with type 1 diabetes. *Diabetes Care*, *29*(6), 1300–1306. https://doi.org/10.2337/DC05-2470

Grisanti L. A. (2018). Diabetes and Arrhythmias: Pathophysiology, Mechanisms and Therapeutic Outcomes. Frontiers in physiology, 9, 1669. https://doi.org/10.3389/fphys.2018.01669

He, H., Bai, Y., Garica, E. A., Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969

James, S., Perry, L., Lowe, J., Harris, M., Colman, P. G., Craig, M. E., & Australasian Diabetes Data Network Study Group (2023). Blood pressure in adolescents and young adults with type 1 diabetes: data from the Australasian Diabetes Data Network registry. Acta diabetologica, 60(6), 797–803. https://doi.org/10.1007/s00592-023-02057-4

Jiang, Y., & Lai, X. (2024). Clinical features of early-onset type 2 diabetes and its association with triglyceride glucose-body mass index: a cross-sectional study. *Frontiers in Endocrinology*, *15*, 1356942. https://doi.org/10.3389/FENDO.2024.1356942/BIBTEX

Jiang, W., & Li, Z. (2024). Comparison of Machine Learning Algorithms and Nomogram Construction for Diabetic Retinopathy Prediction in Type 2 Diabetes Mellitus Patients. *Ophthalmic research*, *67*(1), 537–548. https://doi.org/10.1159/000541294

Kibria, H. B., Matin, A., Jahan, N., & Islam, S. (2021). A comparative study with different machine learning algorithms for diabetes disease prediction. *2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 1–8. https://doi.org/10.1109/cce53527.2021.9633043

McDonough, C., Li, Y. C., Vangeepuram, N., Liu, B., & Pandey, G. (2023). Facilitating youth diabetes studies with the most comprehensive epidemiological dataset available through a public web portal. *MedRxiv*, 2023.08.02.23293517. https://doi.org/10.1101/2023.08.02.23293517

Mohsen, F., Al-Absi, H. R. H., Yousri, N. A., El Hajj, N., & Shah, Z. (2023). A scoping review of artificial intelligence-based methods for diabetes risk prediction. *Npj Digital Medicine*, *6*(1), 197. https://doi.org/10.1038/s41746-023-00933-5

Nijpels, G., Beulens, J. W. J., van der Heijden, A. AWA, Elders, P. J. (2019). Innovations in personalised diabetes care and risk management. *European Journal of Preventive Cardiology*, 26(2_suppl), 125–132. https://doi.org/10.1177/2047487319880043

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Pinhas-Hamiel, O., & Zeitler, P. (2007). Acute and chronic complications of type 2 diabetes mellitus in children and adolescents. *Lancet (London, England)*, *369*(9575), 1823–1831. https://doi.org/10.1016/S0140-6736(07)60821-6

Sacks, D. B., Arnold, M., Bakris, G. L., Bruns, D. E., Horvath, A. R., Lernmark, Å., Metzger, B. E., Nathan, D. M., & Kirkman, M. S. (2023). Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. *Diabetes Care*, *46*(10), e151–e199. https://doi.org/10.2337/DCI23-0036

Strati, M., Moustaki, M., Psaltopoulou, T., Vryonidou, A., & Paschou, S. A. (2024). Early onset type 2 diabetes mellitus: an update. *Endocrine*, *85*(3), 965–978. https://doi.org/10.1007/s12020-024-03772-w

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*. https://doi.org/10.1049/htl2.12039

Vangeepuram, N., Liu, B., Chiu, P. hsiang, Wang, L., & Pandey, G. (2021). Predicting youth diabetes risk using NHANES data and machine learning. *Scientific Reports 2021 11:1*, *11*(1), 1–9. https://doi.org/10.1038/s41598-021-90406-0

Wareham N. J. (2022). Personalised prevention of type 2 diabetes. Diabetologia, 65(11), 1796–1803. https://doi.org/10.1007/s00125-022-05774-7

## 7. Appendix

*Sarah Park*
https://github.gatech.edu/ISyE6414-SP2025/phantom_phoenixes/tree/main/spark978/
I contributed to brainstorming the topic, writing and editing the analysis plan and final report (Summary, Introduction, Methodology, Analysis - Approach #1, Explanation of Changes, Conclusions, References), and writing the code and carrying out the analysis for the Analysis - Approach #1.

*Ye Jun Kim*
https://github.gatech.edu/ISyE6414-SP2025/phantom_phoenixes/tree/main/ykim3060
Contributed to creating and initiating project documentation, and handled the analysis for building and measuring performance of a random forest model on age groups less than 40 years old from the master dataset. Also assisted with key aspects of documentation, such as writing, organization, and formatting.

*Sihan Wang*
https://github.gatech.edu/ISyE6414-SP2025/phantom_phoenixes/tree/main/swang943/

I implemented the code and analysis for approach #3 logistic regression with age group ≥40 years in the Analysis section. I contributed to writing and editing the analysis plan and final report (Plan, Introduction, Conclusion, Reference, Summary)

*Yahya Hosseini*
https://github.gatech.edu/ISyE6414-SP2025/phantom_phoenixes/tree/main/shosseini37
I handled the coding to parse, clean, and organize the dataset. This included importing the fixed and dependent variables into a structured DataFrame for the rest of the team to work with.
Wrote the NHANES dataset overview section in Analysis section
Wrote the code and analysis for the approach #4 in the Analysis section.
Contributed to Explanation of Changes section