Predicting Diabetes and Identifying Important Health Indicators

An Exploration of the Application of Machine Learning on the CDC's 2015 Telephone Survey

Stephanie Boissonneault
500616408
Supervisor: Tamer Abdou
Dec 9th, 2024

Ryerson University

Table of Contents

Abstract

Diabetes Mellitus is a multifaceted and widespread non-communicable disease affecting many Canadians and individuals worldwide. Early disease diagnosis is important for disease management and mitigating further health complications. A literature review examining 9 studies was conducted to explore the use and application of machine learning in predicting a diabetes diagnosis and identifying diabetes health indicators. While the use of machine learning in predicting diabetes diagnosis has been widely researched, many studies lack heterogeneity across the selected population, sample, machine learning techniques, and reported validation metrics. Further, much of the literature reveals a need for more transparency in sharing the features selected. These discrepancies pose challenges during model comparisons between studies when seeking to evaluate and choose the best models for clinical implementation. Transparency is crucial for clinicians to interpret and validate the model's diagnosis and disease detection. This project investigated associations between health indicator features and diabetes diagnosis features using results from the CDC American health survey to build predictive machine learning models for disease detection that address model interpretability for clinical application and that report on various metrics for model comparison between studies. It was found that multinomial logistic regression was the best model for feature selection, while the decision tree model and the multinomial logistic regression predictive models similarly performed best for predicting diabetes based on stability, validity, generalizability, interpretability, and efficiency. It is recommended that future studies investigate model performance on Canadian datasets with an emphasis on recall, F2 score, and Roc Auc for

clinical considerations, and opportunities for validation in clinical practice as part of their study continuity.

Introduction

Good health and well-being is one of the 17 Sustainable Development Goals in the Government of Canada's 2030 Agenda (Government of Canada, 2024b). To help achieve this goal, the Government of Canada works towards preventing non-communicable diseases, the leading cause of premature deaths globally (Government of Canada, 2024a). Diabetes Mellitus is a multifaceted widespread non-communicable disease affecting 1 in 10 adults worldwide (World Health Organization, n.d. as cited in Statistics Canada, 2023). There are different types of diabetes including prediabetes, gestational diabetes, type 1 diabetes, and the most common being type 2 diabetes (Diabetes Canada, n.d.b.). Several potential health complications can arise from diabetes mellitus including "kidney disease, foot and leg problems, eye disease (retinopathy) that can lead to blindness, heart attack & stroke, anxiety, nerve damage, amputation and erectile dysfunction" (Diabetes Canada, n.d.b.). Early diagnosis and disease management are therefore crucial for mediating further health complications and improving the health of Canadians. This research project investigates the application of machine learning techniques for predicting diabetes and identifying its risk factors for early disease detection and intervention using the publicly available Centers for Disease Control and Prevention (CDC) 2014 Behavioral Risk Factor Surveillance System survey data and health survey data.

Literature Review

Literature Search Strategy

A literature review was conducted in the third week of October 2024 examining peer-reviewed research articles published between January 1st, 2019, and October 13th, 2024. The literature review was conducted using PubMed, a free and public database with over 37 million works of biomedical and life sciences literature. The key search terms “diabetes”, and “machine learning” were used. Filters for Meta-Analysis, Randomized Control Trials, and Systematic Reviews were applied to help limit results and ensure articles selected for review were high up on the hierarchy of evidence. Studies were selected for review if they met the following inclusion criteria: Peer-review journal articles from open-access databases with full-text available online in English examining the use of machine learning for the identification of diabetes health indicators or for predicting diabetes in the general adult population. Studies were conducted on a global scale and special populations such as pregnant women and geriatric patients were included to gain a global view of the application of machine learning for predicting different types of diabetes diagnosis at different life stages. Studies focused on populations with specific underlying illnesses were excluded to focus the review on detecting diabetes diagnosis among healthy individuals. Studies investigating the use of machine learning for the subclassification of diabetes, disease management, or the prediction of diabetes health complications were also excluded to focus on predicting diabetes and identifying its risk factors. Deep learning techniques often use many features and apply decision-making processes that are not easily interpretable, making it difficult for clinicians to understand how the model arrived at a specific

diabetes diagnosis, which could complicate further discussions and decisions on treatment options; Studies focusing solemnly on deep learning techniques were therefore excluded to isolate the search to commonly used machine learning techniques with higher interpretability.

Literature Search Results

The initial results from the PubMed search revealed 111 articles. No duplicates were identified. Among these articles, 98 were excluded during the initial screening of the article's titles and abstracts as they did not meet the inclusion criteria. The remaining 13 articles were imported into Mendeley for a full-text review. During the full-text review, 4 articles were excluded based on the inclusion and exclusion criteria. A total of 9 articles were selected for inclusion in the literature review. This selection process is further summarized in the Search Strategy Diagram outlined in Figure 1.

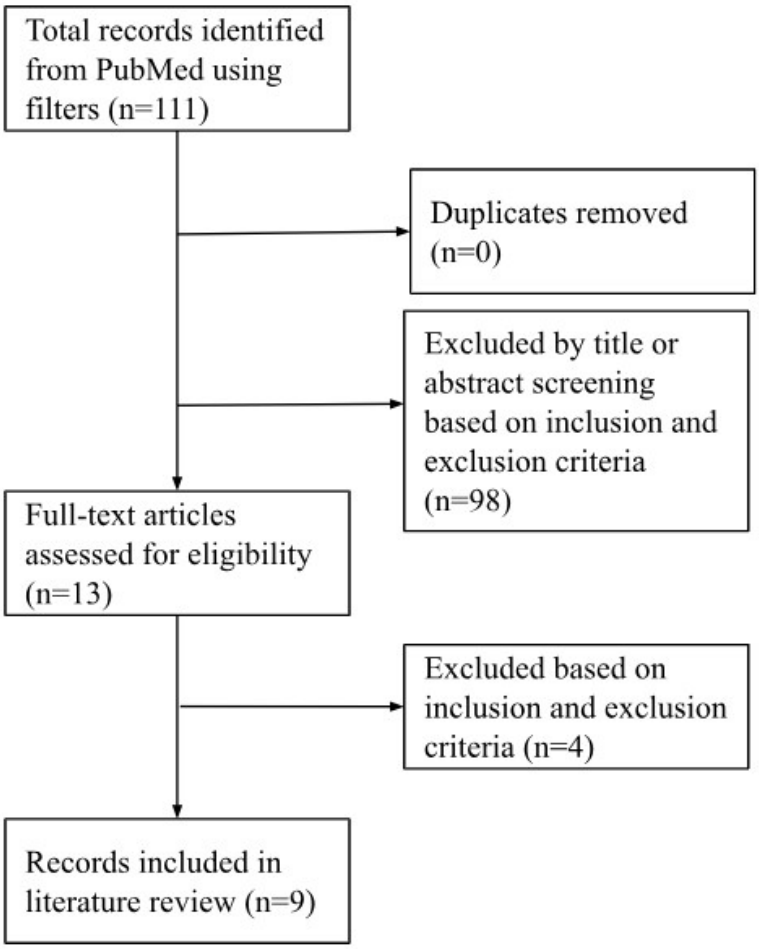


Figure 1. Search Strategy Diagram

Included articles consisted of 5 systematic reviews, 2 meta-analyses, 1 survey directed according to systematic review and meta-analysis guidelines, and 1 secondary analysis of a randomized control trial. Of the 9 articles, 3 investigated the use of machine learning models towards predicting various diseases, where one of the diseases investigated was diabetes, and are summarized in Table 1 (see Appendix A). The remaining 6 explored and compared the

application of different types of machine learning models towards detecting either the risk of diabetes, or identifying the disease, and are summarized in Table 2 (see Appendix A).

Diabetes as a Common Disease for the Application of Predictive Modeling

Research into the application of machine learning models towards predicting diabetes mellitus was found to be quite extensive in the literature. A systematic review conducted by Adbulazeem et al. (2023) selected and analyzed 106 articles examining the use of machine learning prediction models for health conditions. Out of the 42 health conditions covered in the review, diabetes mellitus consisted of 19.8% of the research and was found to be the most frequently targeted health condition by machine learning prediction models. Similarly, findings from a systematic review examining 70 articles investigating the application of machine learning towards aging-related concerns in healthy individuals aged 45 and up, further confirm the extensive application of machine learning toward predicting and identifying risk factors for diabetes mellitus (Das and Dhillon, 2023). Despite the extensive research into this topic, much of the current research lacks heterogeneity making comparisons between study models difficult. There is also an overall lack of consensus on the best machine-learning algorithms for building models for diabetes detection.

Commonly Applied Machine Learning Algorithms

The literature review revealed the application of many different types of machine learning algorithms for predicting diabetes and identifying its risk factors. Common machine learning algorithms identified during the search include random forest, logistic regression, X boost, support vector machine, extreme, light, and adaptive decision trees, gradient boosting trees, naive Bayes, k-nearest neighbors, LASSO, fuzzy logics, gradient boosting machines and

neural networks (Abdulazeem et al., 2023; Fregoso-Aparico et al., 2021; Kodama et al., 2022; Kumar et al., 2023; Olusanya et al., 2022; Zanelli et al., 2022). Disease risk factors were also commonly identified using techniques such as clustering algorithms like principle component analysis, and logistic regression classifier (Das & Dhillon, 2023).

Although there is no clear consensus on which algorithms build the best-performing model, multiple articles described decision tree models as the most common and best-performing models. A meta-analysis by Olusanya et al. (2022) reviewed 34 studies from different countries between 2010 and 2021 to investigate machine learning models' ability to predict type 2 diabetes mellitus and found that the most frequently used model for predicting type 2 diabetes mellitus consisted of decision tree models with a high pooled accuracy of 0.88. Non-linear dynamic machine learning models such as support vector machines or decision trees were also found to perform better by Abdulazeem et al. (2023). A systematic review by Fregoso-Aparicio et al. (2021) revealed a decision tree and random forest as the top-performing models based on performance metrics out of the 18 different types of models investigated.

Lastly, non-logistic regression models were found to perform better than logistic regression models in a meta-analysis by Zhang et al. (2022). The study examined 25 studies that developed machine learning prediction models for Gestational Diabetes Mellitus across the general population including women aged over 18 without a history of vital disease. A Prediction Model Risk Assessment Tool (PROBAST) was used to evaluate the risk of bias of each Machine learning model, while sensitivity analysis, a meta progression, and a subgroup analysis were also conducted to limit the influence of heterogeneity. They found that non-logistic

regression models had a pooled AUROC of 0.889, indicating a higher performance than the logistic regression models with a pooled AUROC of 0.8151 (Zhang et al., 2022).

Commonly Selected Features for Diabetes Prediction

The heterogeneity between datasets and features across studies results in many different types of models with differing features selected as the best diabetes predictors. Fregoso-Aparicio et al. (2021) conducted a systematic review of 90 articles examining the machine learning techniques used in type 2 diabetes prediction. Their study revealed that a combination of lifestyle, socioeconomic, and diagnostic data generally produced better predictive models. Kumar et al. (2023) surveyed systematic review and Meta-Analysis guidelines to examine the use of machine and deep learning classification for disease prediction. They found that common diabetes and blood glucose predictors include blood glucose, insulin, body mass index (BMI), stress, illness, medication amounts of sleep, and periodic heart rate. Features such as PPG and ECG were examined in a systematic review by Zanelli et al., (2022) and were also found to create promising machine-learning models for the prediction of diabetes. Kodama et al. (2022) conducted a systematic review of 12 studies between 1950-2020 comparing machine learning classification of diabetes with the actual incidence of the disease and found that the most frequently selected features were age, obesity, and blood glucose while physical activity and family history were rarely selected. Varga et al. (2021) conducted a secondary analysis of a randomized control trial on data from the Diabetes Prevention Program to evaluate the use of machine learning for predictive models of diabetes using features such as standard lipid measurements and NMR-measure lipoprotein size and concentration. These machine learning algorithms however did not perform better than logistic

regression suggesting a lack of sufficient interactions between the analytes assessed (Varga et al., 2021).

Commonly Reported Evaluation Metrics

While different studies reported on different evaluation metrics for evaluating model performance, some evaluation metrics were more commonly reported than others. During their literature review, Kumar et al, (2023) noted that common metrics for model evaluation applied throughout the literature include precision, recall, accuracy, and f1 score. Many of the studies analyzed by Fregoso-Aparicio et al. (2021) in their systematic review also reported on model performance based on metrics from a confusion matrix. Kadama et al. (2022) also reviewed and compared machine learning models based on results from a confusion matrix regarding the pooled sensitivity, specificity, positive likelihood ratio, and negative likelihood ratio. A lack of standardization of reported parameters increased the difficulty of model comparison between studies (Fregoso-Aparicio et al., 2021; Zanelli et al., 2022). There is an opportunity for future research to increase the number of reported parameters to include evaluation metrics such as accuracy, sensitivity, specificity, precision, and F1-score to increase opportunities for benchmarking and comparison between models (Fregoso-Aparicio et al., 2021).

Commonly Reported Challenges

A common challenge discussed across many of the articles is the high heterogeneity and lack of model validation. During their literature review, Kumar et al. (2023) noted that many of the studies examined used small sample sizes and had insufficient data. The Meta-analysis by Olusanya et al. (2022) also noted a high variance of estimates across studies, as well as heterogeneity between study populations, sample sizes, and burden of disease present among

the different groups. The differences in the selected population, sample, and features used in machine learning models increase the difficulty of model comparison between studies (Fregoso-Aparicio et al., 2021). Many studies also have a high risk of bias due to a lack of model validation (Abdulazeem et al., 2023). Further, a lack of transparency regarding the features selected and implemented in the machine-learning models is quite common, making the models difficult to interpret for use and validation by clinicians in healthcare settings (Fregoso-Aparicio et al., 2021). Increased transparency and validation are recommended for future studies for clinical implementation and to increase the model's generalizability (Das & Dhillon, 2023; Fregoso-Aparicio et al., 2021; Kumar et al., 2023; Zhang et al., 2022).

Other Recommendations and Future Opportunities

Diabetes is a multifaceted long-term chronic disease. It is therefore important that future studies apply machine learning algorithms on datasets that consider a combination of factors such as genetics, lifestyle, and environmental factors to provide insight into complex and dynamic disease prediction. (Abdulazeem et al., 2023). Machine learning models might also perform best on well-structured balanced datasets composed of many different feature types (Fregoso-Aparicio et al., 2021). Balancing the data and reducing dimensionality through feature selection to increase model accuracy is recommended (Fregoso-Aparicio et al., 2021; Kumar et al., 2023). Clinical need rather than accuracy can also be considered during feature selection to help build models that ensure the features used can be easily obtained during routine medicine (Zhang et al., 2022). Comparisons between machine learning models using risk models and the same database for the prediction of diabetes mellitus are also recommended (Fregoso-Aparicio et al., 2021; Kodama et al., 2022; Olusanya et al., 2022). Lastly, current policy changes should

be considered when analyzing historical data to avoid reinforcing outdated practices (Abdulazeem et al., 2023).

Literature Search Summary

The application of machine learning for predicting diabetes and identifying its risk factors has already been extensively researched in the literature. Numerous types of machine learning models have been built using many different datasets, often revealing high-performing models. Despite this, little consensus has been reached on the health indicators identified as the most important predictors of the disease. There is also little consensus on which types of models might be best used in clinical practice settings for helping with the early identification of the disease. This lack of consensus can be attributed to the common challenges with model comparisons between studies due to a lack of heterogeneity in the explored population, sample, features, machine learning algorithms, and reported evaluation metrics. A lack of model validation and transparency is also common across studies, limiting their application for disease identification in clinical healthcare settings. Opportunities identified include further investigation of important features for disease prediction using feature selection techniques on well-structured and balanced datasets. Further exploration into the application of different types of machine learning algorithms and extensive reporting on evaluation metrics is also recommended to allow benchmarking between studies.

Research Questions

The lack of consensus across studies for the models and features identified as the best predictors of diseases suggests a need to explore the use of machine learning techniques for identifying and understanding diabetes predictors. This project seeks to answer the question ‘’What health conditions and lifestyle factors commonly occur together in individuals with different diabetes diagnoses?’’ to identify diabetes screening questions that should be considered in conjunction in clinical practice settings. This project also investigates “What health indicators are considered most important for disease prediction?’’ and “What machine learning models best predict diabetes mellitus based on effectiveness, efficiency, stability, and interpretability?” to recommend predictive models that could improve diseases diagnosis of the for early intervention and disease management. These findings could also guide initial screenings in clinical settings, enhancing applicability. Lastly, the literature review revealed that age and obesity are common diabetes predictors (Kodama et al., 2022; Kumar et al., 2023). Diabetes Canada also lists that diabetes prevalence varies across age, sex, and weight in addition to other select demographics (Diabetes Canada, n.d.a.). This project will therefore also investigate “What patterns can be uncovered in subpopulations based on age, sex, and BMI?” to help provide further insights into variations of feature importance for tailoring programs that meet the subgroup’s needs. Overall, the proposed research questions for this project have the goal of helping identify diabetes and its risk factors through exploring opportunities that support Canada’s goal of health and well-being.

Research Methodology

To help answer the research questions, this project applies machine learning techniques to a consolidated version of the Centre of Disease Control and Prevention (CDC) 2014 Behavioral Risk Factor Surveillance System (BRFSS) Survey dataset titled “diabetes _ 012 _ health _ indicators _ BRFSS2015.csv” (Teboul, 2021; UC Irvine Machine Learning Repository, 2023). Tools such as Visual Studio Code, Python programming language, and Python libraries were used to clean the data, build machine learning models, and uncover patterns and trends in the data.

Data Description

The dataset consists of 22 variables and a sample of 253,680 observations from the CDC’s annual health-related telephone survey examining risk behaviors, chronic health conditions, and use of preventative services from 400,000 respondents across 50 states in the US (Teboul, 2021; Centers for Disease Control, 2015). The dataset contains 6 numeric type and 16 categorical type variables, one of which is Diabetes_012 classifying respondents as either (0) having no diabetes or only diabetes during pregnancy, (1) having pre-diabetes, or (2) being diagnosed with diabetes. All features included in the dataset and their datatypes are described in Table 3 (see Appendix B). An exploratory analysis report was generated using Python’s Panda library and is available in the GitHub repository’s project files.

Data Cleaning and Preprocessing

Data cleaning and preprocessing techniques were applied to the dataset to ensure data accuracy and reliability. The data was first checked for missing values and duplicate records. No missing values were identified in the dataset. The zeros recorded for MentHlth and PhysHlth are considered valid rather than missing data since they represent that there were no days where the individual felt that their mental health or physical health was compromised. It was determined that the duplicate values could represent valid reoccurrences of individual diabetes diagnoses and health indicators, which is meaningful when evaluating the itemset frequency when generating association rules. Duplicates are therefore not dropped during initial data cleaning but are however dropped at subsequent steps before applying feature selection techniques and when building predictive models to avoid biased results. The data was transformed to their appropriate datatypes; All data was transformed to the categorical datatype except for BMI which is a numeric continuous measure, and MentHlth and PhysHlth which are numeric discrete. This would allow descriptive statistics to be applied to various types of data. In addition to transforming the data to the appropriate type, categories were renamed to aid with the interpretation of the descriptive statistics. Subsequent preprocessing steps were applied to categorize the BMI, MentHlth, and PhysHlth values since select tests and algorithms could only be applied to categorical data. The BMI values were categorized according to the Centers for Disease Control and Prevention (2024) BMI categories for adults 20 and over. The MentHlth and PhysHlth were categorized based on 5-day increments. These categories are listed in Table 4 (see Appendix B) and were integrated into the data frame before applying chi-square during descriptive statistics, and before answering each research question to ensure

consistency in the data evaluated across different machine learning algorithms. Additional preprocessing techniques to ensure the data is appropriately formatted, such as one hot encoding, are outlined in the methods for the appropriate research question.

Descriptive Statistics

First, the frequency distribution of the target variable was visualized and revealed a large class imbalance (see Figure 2). The dataset's large class imbalance can impact the predictive accuracy of machine learning algorithms, which tend to be biased towards the majority class, and could therefore pose an issue. This target variable will therefore need to be balanced at a later stage before the application of machine learning techniques sensitive to these class imbalances. The variables CholCheck, Stroke, HeartDiseaseorAttack, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, were also determined as having a class imbalance based on the EDA report. The variables MentHlth and PhysHlth were also flagged as having a disproportionately high number of zeros by the EDA report.

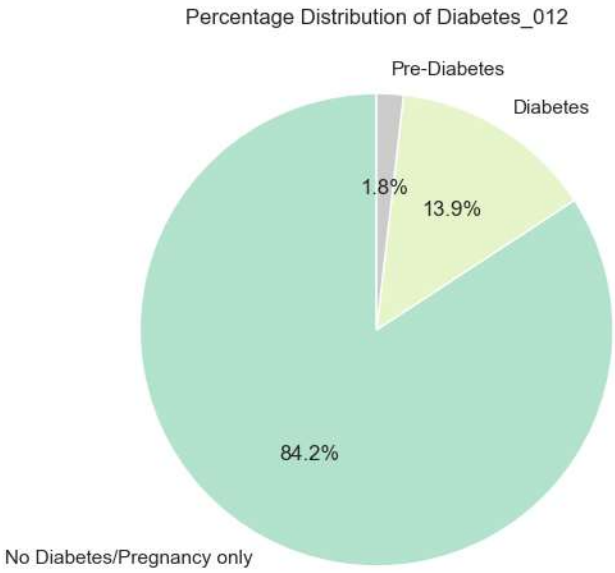


Figure 2. Frequency distribution of “Diabetes_012” target variable

Outliers in the dataset were also checked using boxplots for discrete and continuous data (see Figure 3). Boxplots show many outliers for BMI, MentHlth, and PhysHlth variables. While these outliers are outside of the normal distribution, they represent legitimate critical values of severe health conditions that may be relevant and provide valuable insights for disease prediction in clinical settings. BMI values range between 12 and 98. The values below 16.5 represent severe underweight and the values above 40 represent severe and extreme obesity, which are still possible. Although represented as outliers here, MentHlth and PhysHlth values are legitimate and remain between the 1 - 30 day range. These extreme values could provide insights into identifying and informing approaches to patients at high risk of severe health issues. While retaining these outliers helps visualize the full range of patient’s health, it is important to note that some machine learning models like linear regression and k-nearest neighbors are sensitive to outliers. Close monitoring of their performance metrics is therefore

recommended. If outliers seem to greatly impact the analysis of these models, the outliers could be capped at the 95th percentile or further transformed to reduce their impact.

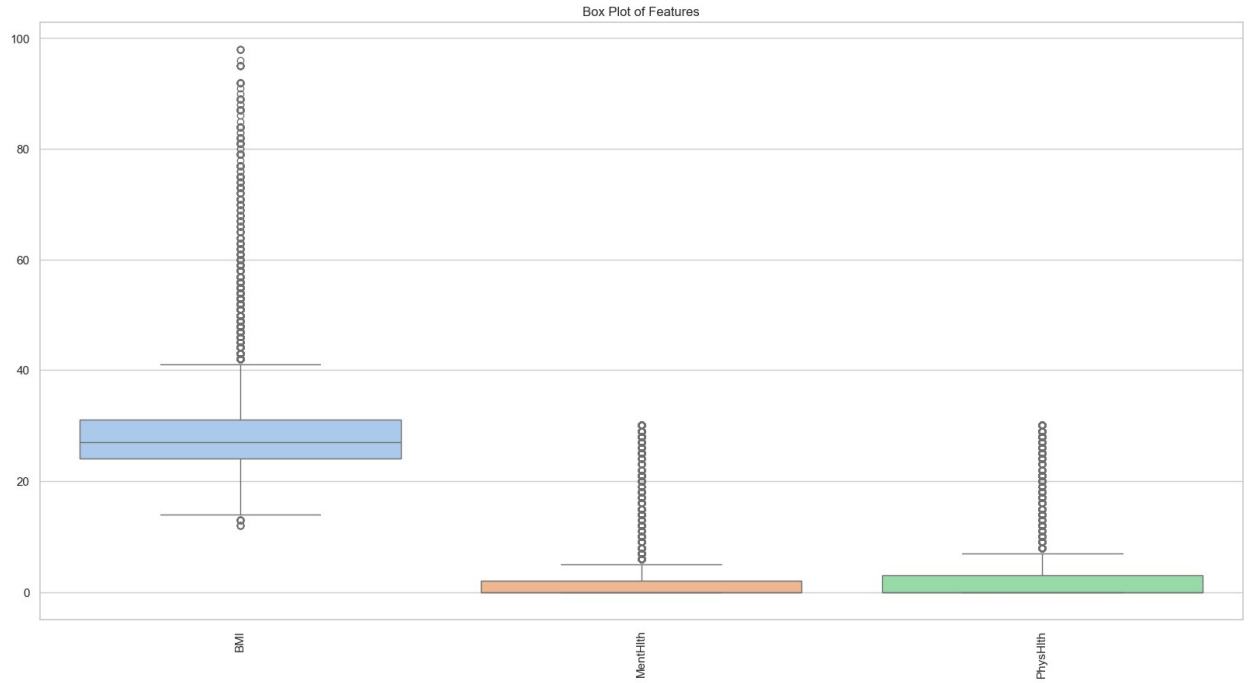


Figure 3. Boxplot of discrete and continuous data

Variables described throughout the literature review as having been identified as good predictors of diabetes including age, obesity/BMI, stress, and illness for which data was also made available in this dataset were further visualized through descriptive analysis. Descriptive statistics were generated for diabetes diagnosis across age ranges in Figure 4. While no diabetes/pregnancy diabetes remains the highest diagnosis for all age categories, the diagnosis for pre-diabetes seems to increase between ages 45 and 74 and decreases between ages 74 and 80. Diabetes diagnosis is the highest for individuals ages 50 to 80 and over. The median of

individual BMIs is lowest for individuals with no diabetes/pregnancy only and is the highest for individuals diagnosed with diabetes (see Figure 5).

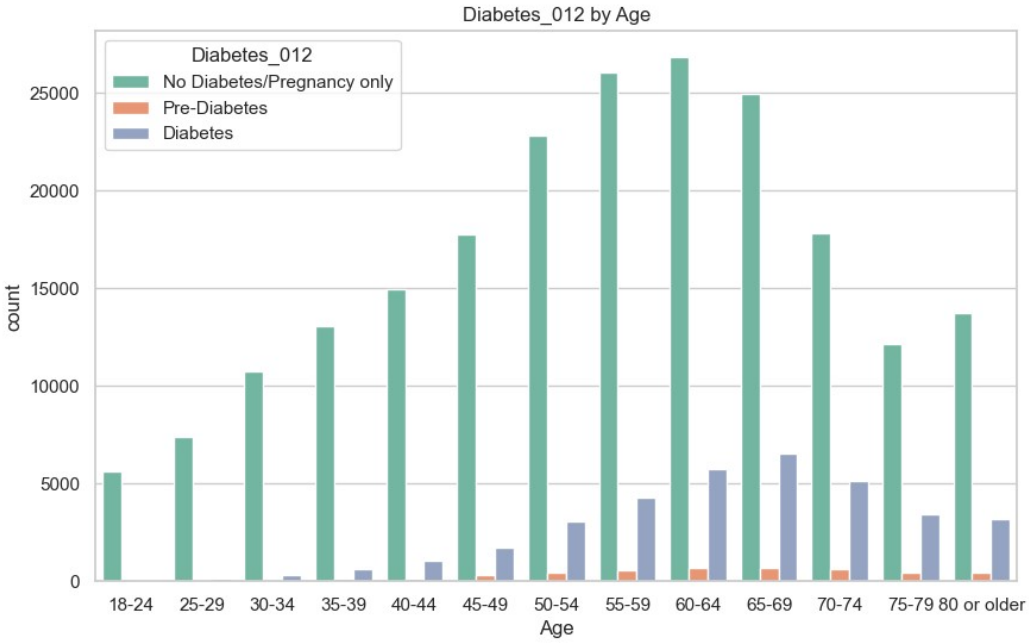


Figure 4. Frequency of diabetes diagnosis across age ranges

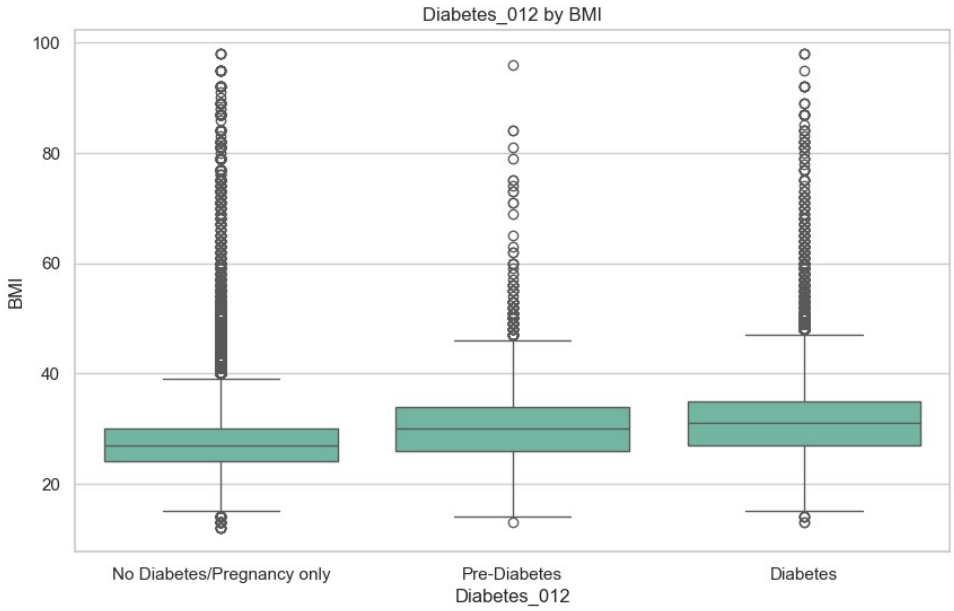


Figure 5. BMI across diabetes diagnosis

While individuals with no diabetes or pregnancy-only diabetes reported a higher proportion of excellent to very good general health ratings, individuals with diabetes reported a higher proportion of fair and poor general health (see Figure 6). The distribution of reported poor mental health days is quite similar across different diabetes diagnoses (see Figure 7).

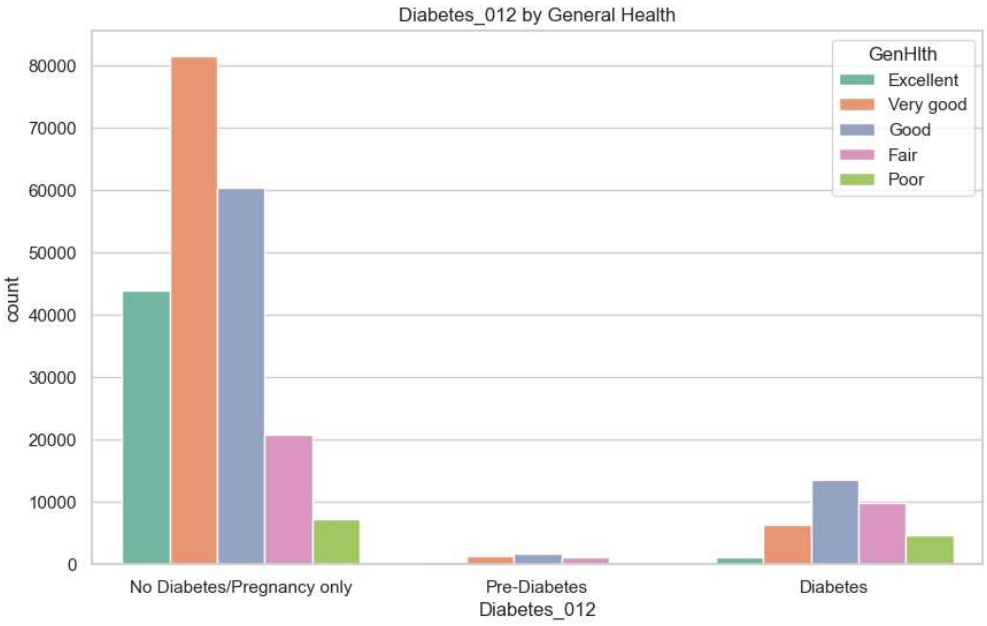


Figure 6. Diabetes by General Health Rating

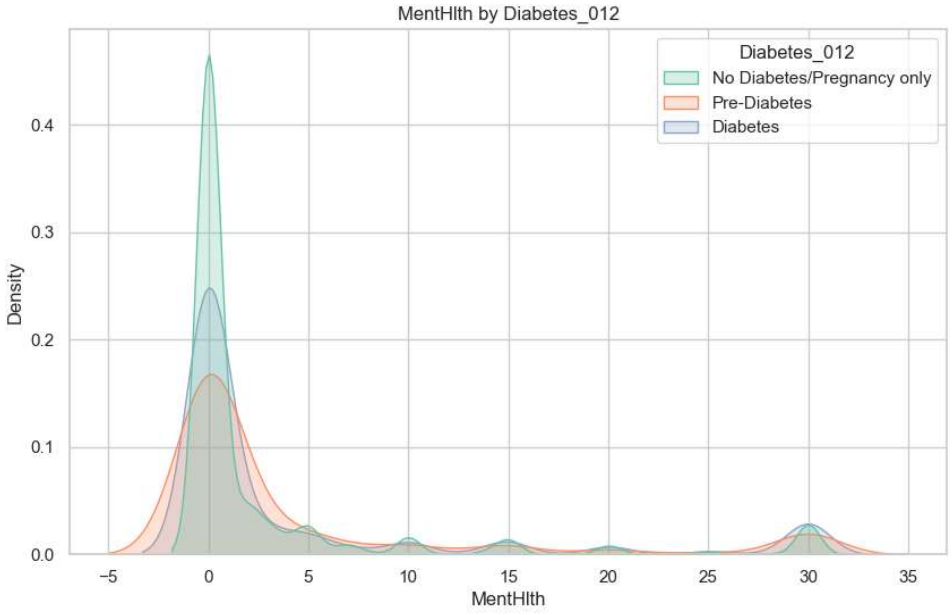


Figure 7. Days of Reported Poor Mental Health by Diabetes Diagnosis

A heatmap of the significant p-values (p=<0.05) for chi-square tests between each feature of the dataset was generated (see Figure 8). The p-values of less than 0.05 for chi-square tests between all pairs of features except for Stroke and Sex indicate that we can reject the null hypothesis that the features are independent, and conclude that there is a statistically significant association between the features. It can be concluded that all health indicators in the dataset have a statistically significant association with the diabetes target variable.

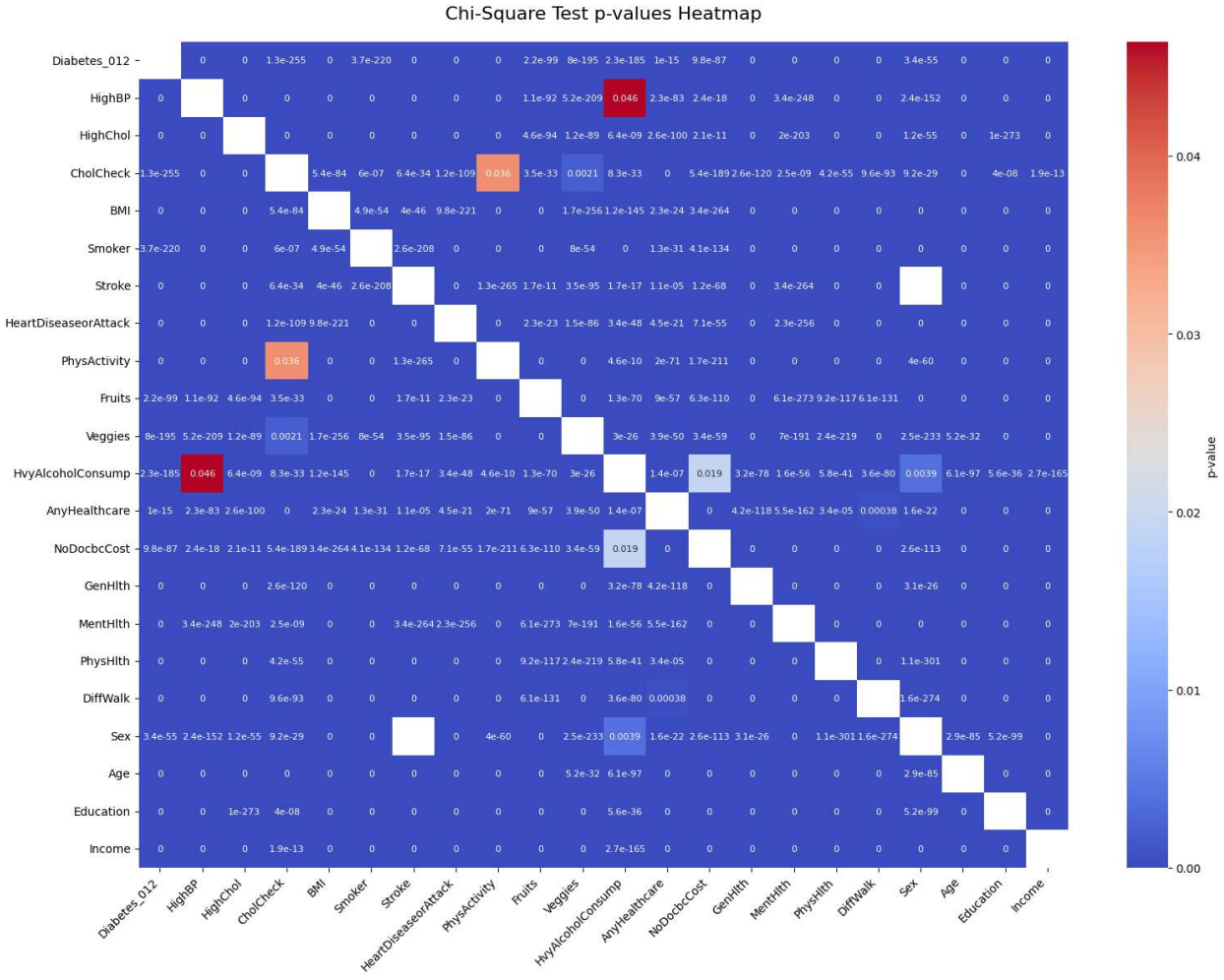


Figure 8. Days of reported poor mental health by diabetes diagnosis

Data Analysis Procedures for Answering Research Questions

Question 1

The apriori algorithm was applied to the dataset to generate itemsets and association rules as outlined in Figure 9. to help identify “What health conditions and lifestyle factors commonly occur together in individuals with different diabetes diagnoses?”. The dataset was transformed into one-hot encoded data using “get_dummies” from the Pandas library, ensuring that boolean values are converted to integers and that all data are subdivided into appropriate categories. The one hot-encoded dataset was saved as a new CSV file for future use. Features not considered to be health conditions or lifestyle factors ((sex, age, education, income, AnyHealthcare, and NoDocbcCost) were dropped from the data frame to focus the analysis on features of interest as identified in the research questions and for memory efficiency considerations. Random sampling of 50% of the health conditions and lifestyle factors dataset was conducted to further reduce memory usage, resulting in a 126,840 sample size.

The apriori algorithm was applied to the sampled dataset using “apriori” from Python’s “mlxtend.frequent_pattern” package. A minimum support[1] value of 0.10 was selected, indicating that the itemsets must appear in at least 10% of the transactions. This minimum support value was selected to help reduce computational time compared to a lower support value and to strike a balance between selecting health conditions and lifestyle factor itemsets that are both rarer and reasonably frequent. The top ten frequent itemsets were visualized in a bar chart for

[1] Support = count of transaction (LHS ∪ RHS) / total number of transactions

analysis. Association rules were generated for when the consequent of the rule (the RHS) is a diabetes diagnosis using “association_rules” from Python’s “mlxtend.frequent_pattern” package.

A minimum confidence[2] threshold of 0.8 was originally selected for generating the association rules, indicating that the RHS (diabetes diagnosis) occurs at least 80% of the time given the antecedent (LHS). This initial high confidence level was selected considering that the rules generated should have a stronger predictive power for application in predicting diabetes diagnosis in clinical settings. Rules for each RHS diabetes diagnosis were further filtered to only include association rules with a lift[3] higher than 1 to filter for stronger associations compared to random chance where an occurrence of the antecedent increases the likelihood of the consequent. Visualizations of the resulting association rules for each RHS diabetes type were generated to extract meaningful insights regarding the strength of the frequent itemsets’ ability to predict the diabetes diagnosis and make recommendations for health indicators that could be a diabetes predictor and that could be assessed together in practice.

[2] Confidence = support of transaction (LHS ∪ RHS) / support(LHS)

[3] Lift = confidence (LHS→RHS)/support(RHS)

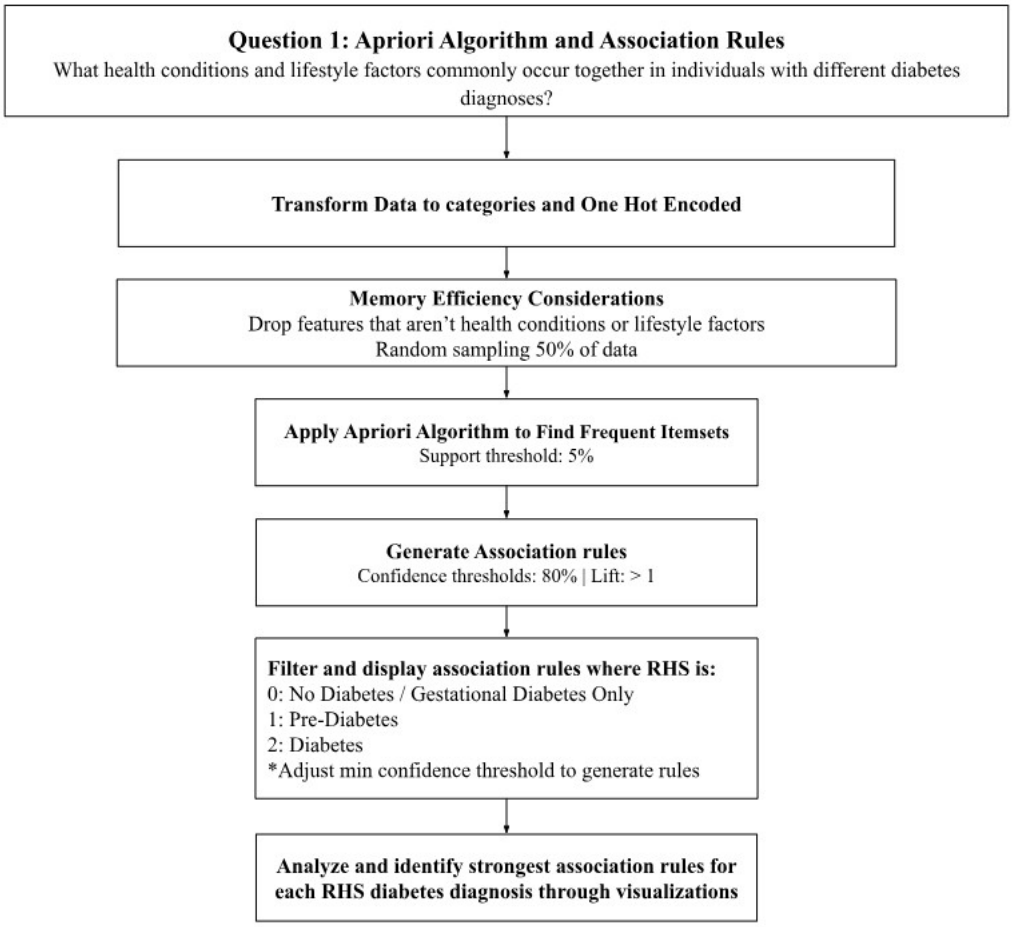


Figure 9: Questions 1 Apriori Algorithm Approach

Question 2

Different feature selection models were analyzed and compared to help answer the question “What health indicators are considered most important for disease prediction?”. Preprocessing techniques were applied to ensure all datatypes were converted to the

appropriate categorical type. Duplicate values were identified and dropped from the dataset to limit biased or overfitted models that could occur when select features are over-represented. The data was split into train and test sets using an 80/20 split. This split ratio was selected because the dataset is a large size (202,619 rows and 22 columns) and a 20% test set would allow for sufficient test data (40,524) to evaluate the model's performance and generalizability. The train and test sets were divided into X containing all features except the target variable, and Y, the diabetes target variable. Oversampling using the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to the training set to balance the data since feature selection techniques can be biased towards the majority class and underrepresent meaningful patterns in the minority class. SMOTE was selected as an oversampling technique because it generates synthetic samples that are diverse, uniform, and balanced across the features of the minority class, increasing variability and generalizability by introducing new information, rather than repeating identical instances. SMOTE is also less sensitive to class imbalances than methods like Adaptive Synthetic Sampling (ADASYN) and has better computational efficiency since it does not apply density calculations.

Feature selection was conducted on the balanced dataset as advised by findings from the literature review and as outlined in Figure 10. The balanced dataset did not need to be standardized because all columns had been converted to categorical. K-fold cross-validation was applied to the training set using k=5 to help reduce bias that might otherwise occur when feature selection is dependent and applied on a single split. Using multiple folds helps simulate repeated training on different subsets of the data, which helps identify features that are consistently important across folds. Only 5 folds were selected because the dataset is already quite large, however using multiple folds can still help increase generalizability. The feature

selection techniques chi-square (filter method), decision tree (wrapper method), and multinomial logistic regression (embedded method using penalty L2 and regularization strength C) were trained for each k-fold. These feature selection techniques were selected because they are interpretable, increasing transparency for clinical use and validation by practitioners in healthcare settings. A confusion matrix was generated for each k-fold and evaluation metrics (accuracy, precision, recall, and f1 score) were aggregated for each feature selection technique for analysis. The top 10 features by importance were also displayed in a table for each fold for each feature selection model for comparison in determining the stability of the model, where consistency in the features selected would demonstrate a more stable model. Feature importance was determined by the ChiSquare value where a higher value indicates a stronger association of the feature with the diabetes diagnosis target variable. Feature importance for the decision tree model was calculated based on a reduction in gini impurity. The value for feature importance for the decision tree ranges between 0 and 1 where a higher value is indicative of the feature's higher importance, having a greater impact on the model's prediction of the target variable. With logistic regression, feature importance is determined by the absolute coefficient values where a larger absolute coefficient value for a feature indicates that it has a stronger association and more power over predicting the target variable. Positive coefficients indicate a positive relationship where an increase in the feature increases the likelihood of the target variable outcome, while a negative coefficient indicates that a decrease in the feature will increase the target variable outcome.

Hyperparameters tuning was conducted for the decision tree and the multiclass logistic regression model to identify parameters that would result in the best generalization performance for each model. This step was completed by applying GridSearchCV and cross-validation to get

the best parameters. The selected parameters were then applied to the final model which was trained on the train set where the top 10 features were selected based on importance and compared to the features selected during each k-fold. Note that chi-square did not undergo this step because there are no parameters to tune. Lastly, the final models were applied to the test to test the generalizability of the model. Performance metrics including accuracy[4], sensitivity[5], specificity[6], precision[7], F1-Score[8], and F2-Score[9] were recorded as a result of the model's application to the test and compared to the performance metrics generated by the train test to analyze the generalizability of the models. The Receiver Operating Characteristic (ROC), which calculates the true positive rate and the false positive rate at various thresholds, and the Area under the Curve (AUC) which summarises the model's performance based on the area under the ROC curve, were also generated for the test set to help identify the overall model's capability at distinguishing between positive and negative classes and determine the best-performing model across all classification thresholds (Google for Developers, n.d.). The wide range of evaluation metrics was selected to enable benchmarking between different models from different studies. The features with the best model interpretability and performance were selected to train multiple predictive machine-learning models in question 3.

[4] Accuracy indicates the proportion of correct predictions of the model. Accuracy = (TP + TN) / Total
[5] Sensitivity, also known as recall, is the rate of true positives. Recall = TP / (TP + FN)
[6] Specificity is the true negative rate. Specificity = TN / (TN + FP)
[7] Precision measure the proportion of correctly identified positive cases out of all cases predicted as positive. Precision = TP / (TP + FP)
[8] F1 Score balances precision and recall. F1 = 2 ((Precision)(Rrecall) / (Precision + Recall))
[9] F2 Score is the harmonic mean of precision and recall that emphasizes importance on recall. F2 = (1+2^2) * ((Precision)(Recall)/((2^2 * Precision) + Recall))

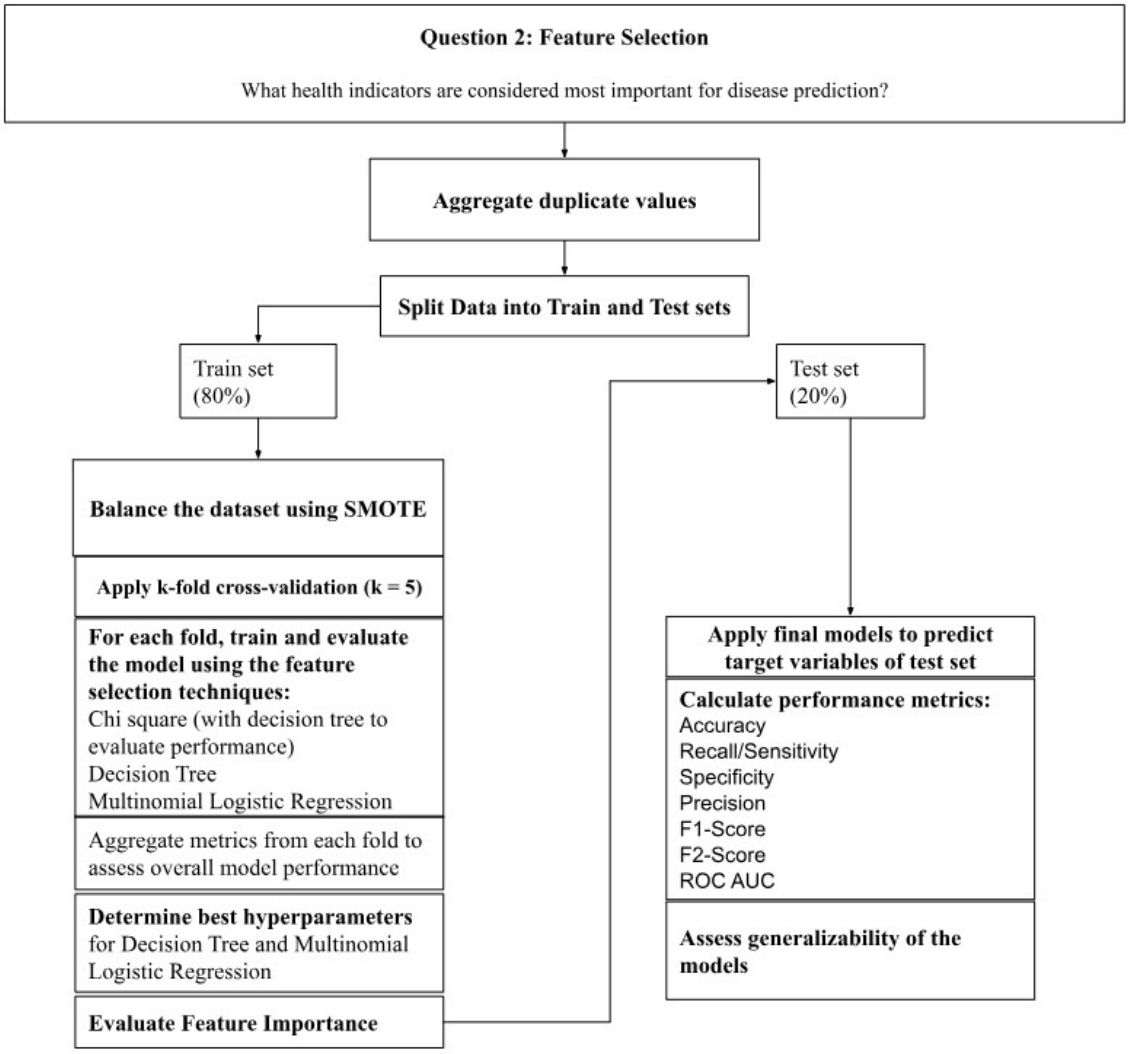


Figure 10: Questions 2 Feature Selection Approach

Question 3

This study trained and tested four different types of machine learning models to answer the question “What machine learning models best predict diabetes mellitus based on

effectiveness, efficiency, stability, and interpretability?" as outlined in Figure 4. The machine learning algorithms decision tree, multinomial logistic regression, Naive Bayes, and k-nearest neighbors were selected because they are machine learning models that have higher interpretability compared to other models that may be more accurate. Interpretability is of concern because, in clinical practice, health professionals require clear and transparent models they can interpret when making a diagnosis or recommendations to their patients. Some models were commonly selected and previously investigated as uncovered during the literature review. Machine learning models with high memory requirements such as random forest were excluded due to hardware limitations in memory capacity. Data preparation techniques applied for question 2 were repeated in question 3; data was converted to the appropriate categorical types and duplicate values were dropped. The data was split into train and test sets using an 80/20 split. The train set was balanced using SMOTE since the severe class imbalance of the target variable could lead to biased or misleading machine learning during training of the model as well as misleading performance metrics and affect feature selection. The best feature selection model as identified in question 2 was applied to the train set. The dimensions of the training set and the test set were reduced to only keep 10 columns representing the top 10 selected features. This step was performed to help with memory efficiency while training and testing the machine-learning models. A k-fold cross-validation was initiated to train each machine learning model on 5 folds. Combining K-folds and train-test split increases the reliability and generalizability of the models, ensuring that bias is minimized. Results for each model on each k-fold were compared to identify the stability of the model. Metrics for the accuracies, precision, recall, f1 score, f2 score, ROC AUC, and confusion matrix were stored for each fold and then aggregated. Hyperparameter tuning for each model was conducted using GridSearchCV

because this technique tests all possible parameter combinations to select the best hyperparameters that would create a final mode. The final model was applied to the train set. The final models were then applied to the test set to evaluate their predictive ability. Several evaluation metrics were reported including accuracy, sensitivity, specificity, precision, F1-score, and F2-score, and Roc Auc to enable benchmarking between models from different studies. Performance metrics from the test set were compared to those of the train set to assess the generalizability of the models. Lastly, the efficiency of the final models on the test set was also evaluated based on computation time, memory usage, and model size.

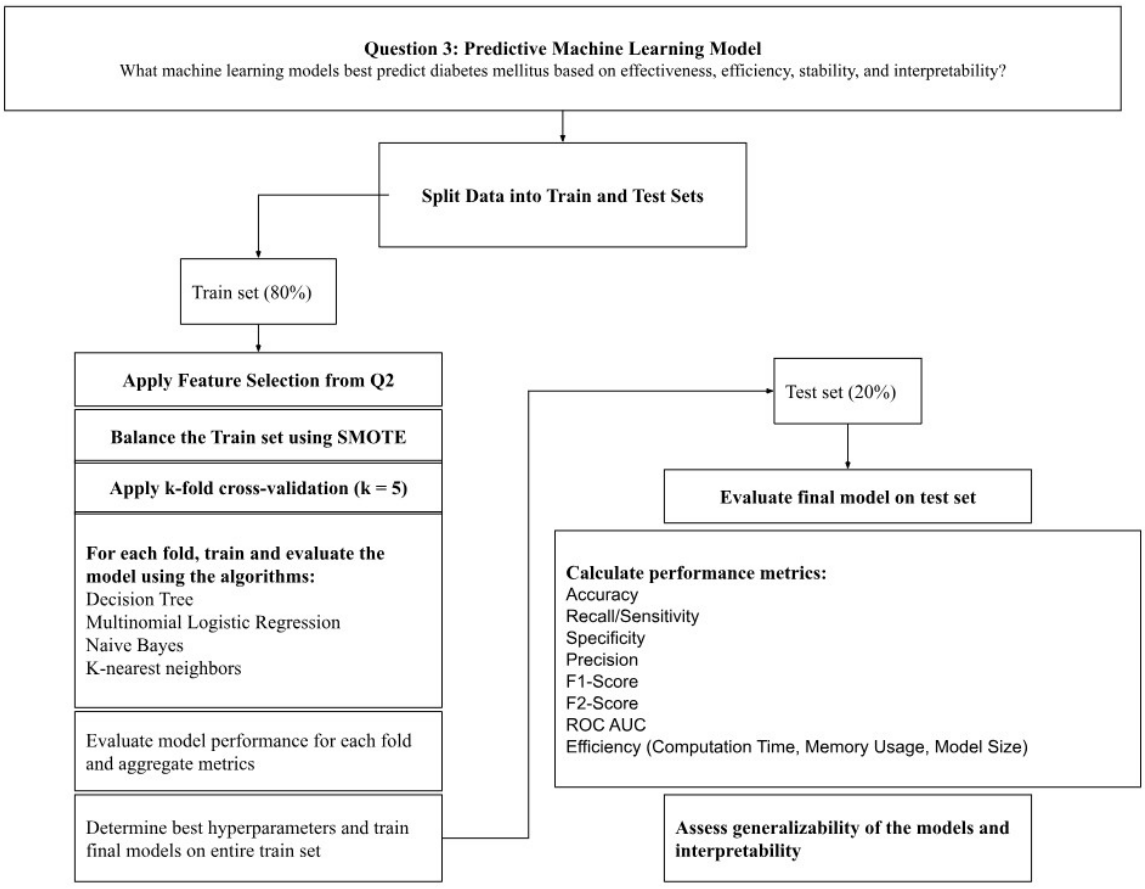


Figure 11. Question 3 Predictive Model Approach

Question 4

An examination of the explained variance from the results of principle component analysis (PCA) on subsets of the population data was used to learn "What patterns can be uncovered in subpopulations based on demographics including age, sex, and BMI" as outlined in Figure 12. Data preparation included one-hot encoding of the dataset. Age categories were subdivided into four bins to represent ages 18-29, 30-49, 50-64, and 65+ to help reduce the amount of subcategories for comparison. The diabetes target variable was dropped from the data frame and each age category was further subdivided by sex, and BMI class, forming a total of 48 data frame subsets, each representing a different combination of age, gender, and BMI subpopulation. A description of each subset and its index are listed in Table 5 of Appendix B. PCA was applied to each subpopulation for analysis. An elbow plot was generated for each subpopulation to use the elbow method to help determine the number of principal components to retain for analysis based on how many principal components generally explained 90% of the variance across subpopulations. The average number of principal components needed to reach a 90% explained variance threshold was also calculated. To identify the features that contribute the most to the explained variance for each subset, I extracted the loadings for the selected number of principal components needed to reach 90% explained variance. The loadings[10] for each subset were then sorted based on the absolute values to select the top 3 features. These

[10] Loadings weighs how much each feature contributes to the principle component. Loadings range between -1 to 1 where 0 indictes minimal feature contribution to the PC and where the absolute value 1 would indicate that the feature strongly contributes to the principle component. The sign indicates the direction of the contribute (positive or negative).

top 3 features were then compared between subsets using visualizations to assess patterns between subsets.

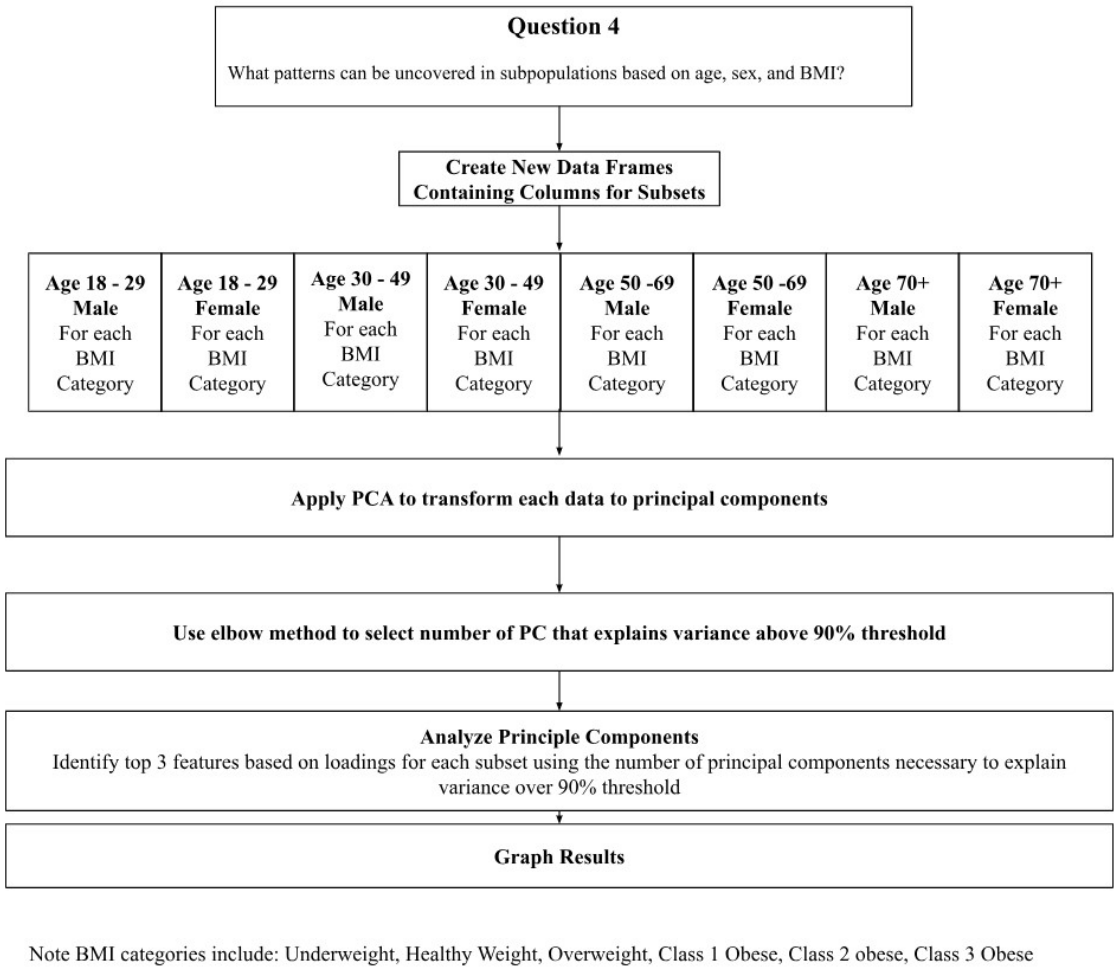


Figure 12. Question 4 PCA Approach

Results

Question 1

The top ten frequent itemsets with the highest support value (the highest reoccurrence in the dataset) are listed in a bar chart (see Figure 13). It was identified that CholCheck and diabetes_0 features occur together in 81% of the dataset, Veggies and CholCheck occur together 78% of the time, and Veggies and Diabetes occur together in 69% of the data.

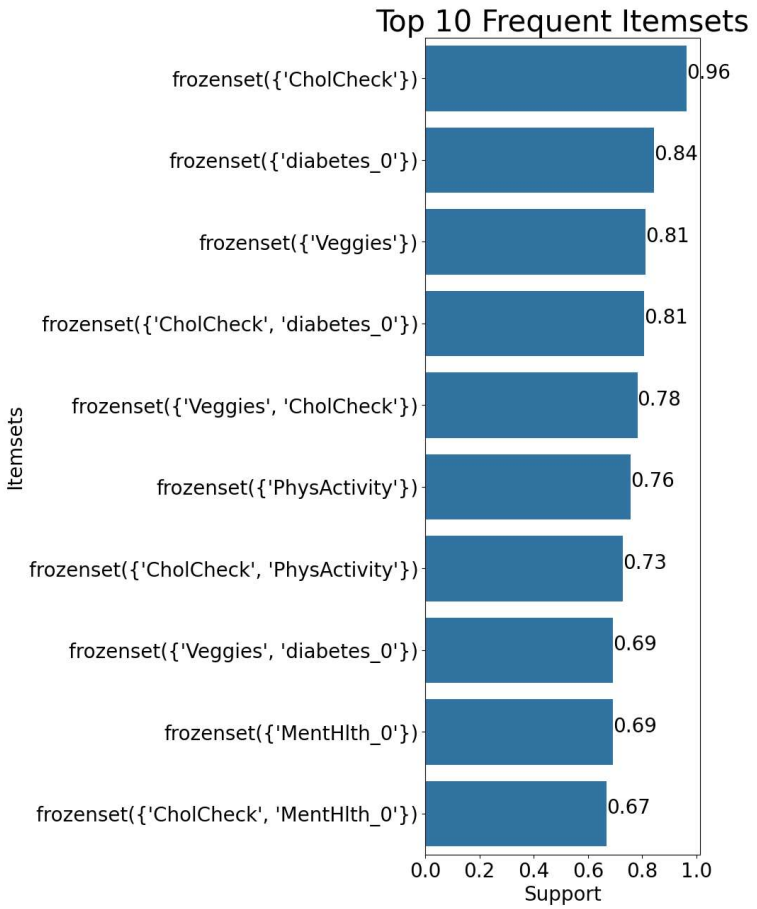


Figure 13. Top 10 Generated Frequent Itemsets with Highest Support Values

A total of 697 association rules were generated with a minimum confidence threshold of 0.80 and lift > 1 for consequents (RHS) that contain diabetes_0. Network graphs for association rules where the support is > 50% (Figure 14) and where confidence is above 97.2% (Figure 15) were generated to identify the most commonly occurring rules and rules with antecedents most likely to result in the diabetes diagnosis. No association rules were generated for consequents (RHS) containing diabetes_1 or diabetes_2 when the minimum confidence threshold was set to (80%). Therefore, the minimum confidence threshold was lowered for each RHS diabetes diagnosis until association rules were generated for analysis. The confidence level was adjusted and reduced to generate association rules for analysis. No association rules could be generated when the RHS contains diabetes_1 (no rules were generated when the confidence was as low as 0.01). Three association rules were generated for RHS containing diabetes_2 when the confidence threshold was reduced to a minimum threshold of 20% (see Table 1). The support, confidence, and lift for RHS diabetes_0 and RHS diabetes_2 association rules were visualized in a scatter plot for comparison (see Figure 16).

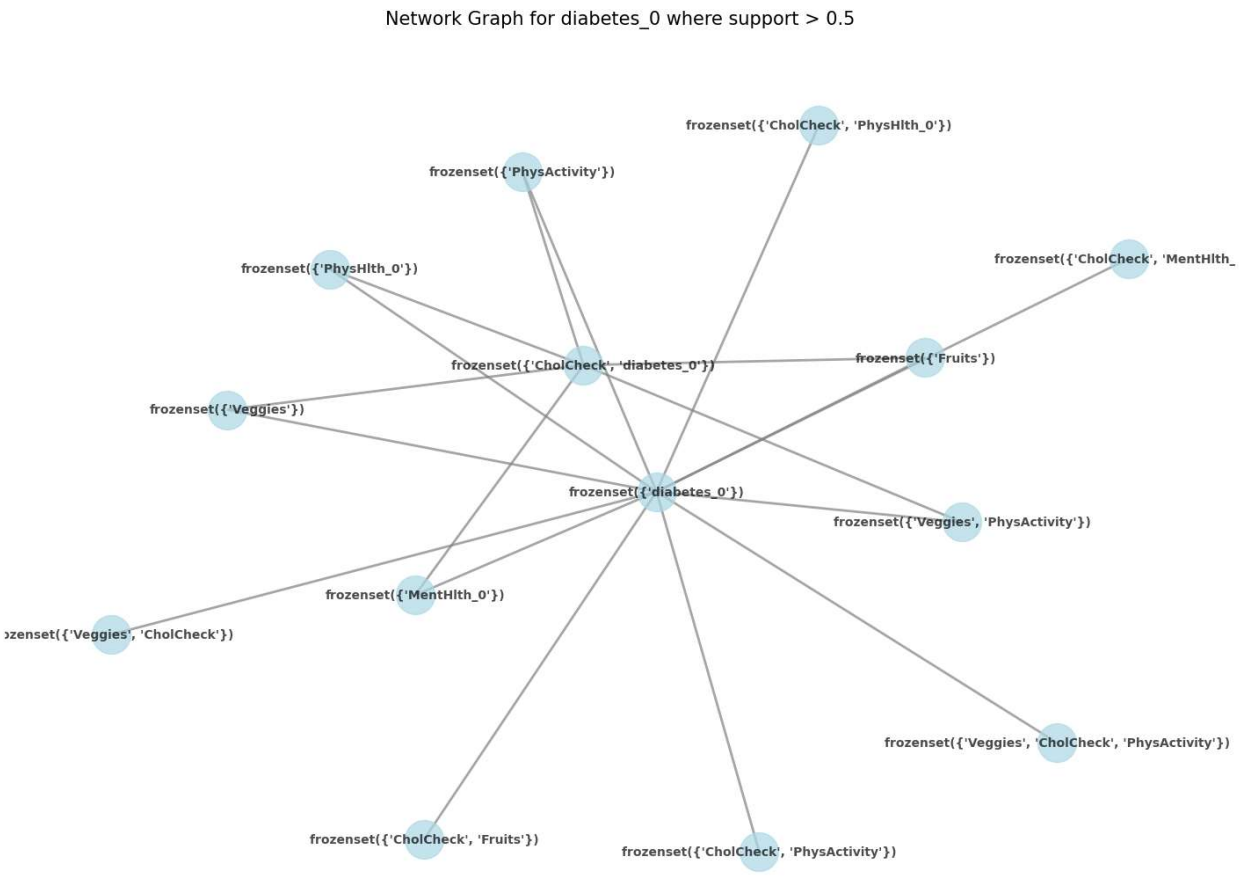


Figure 14. Network graph of association rules for RHS of diabetes_0 with support > 50%

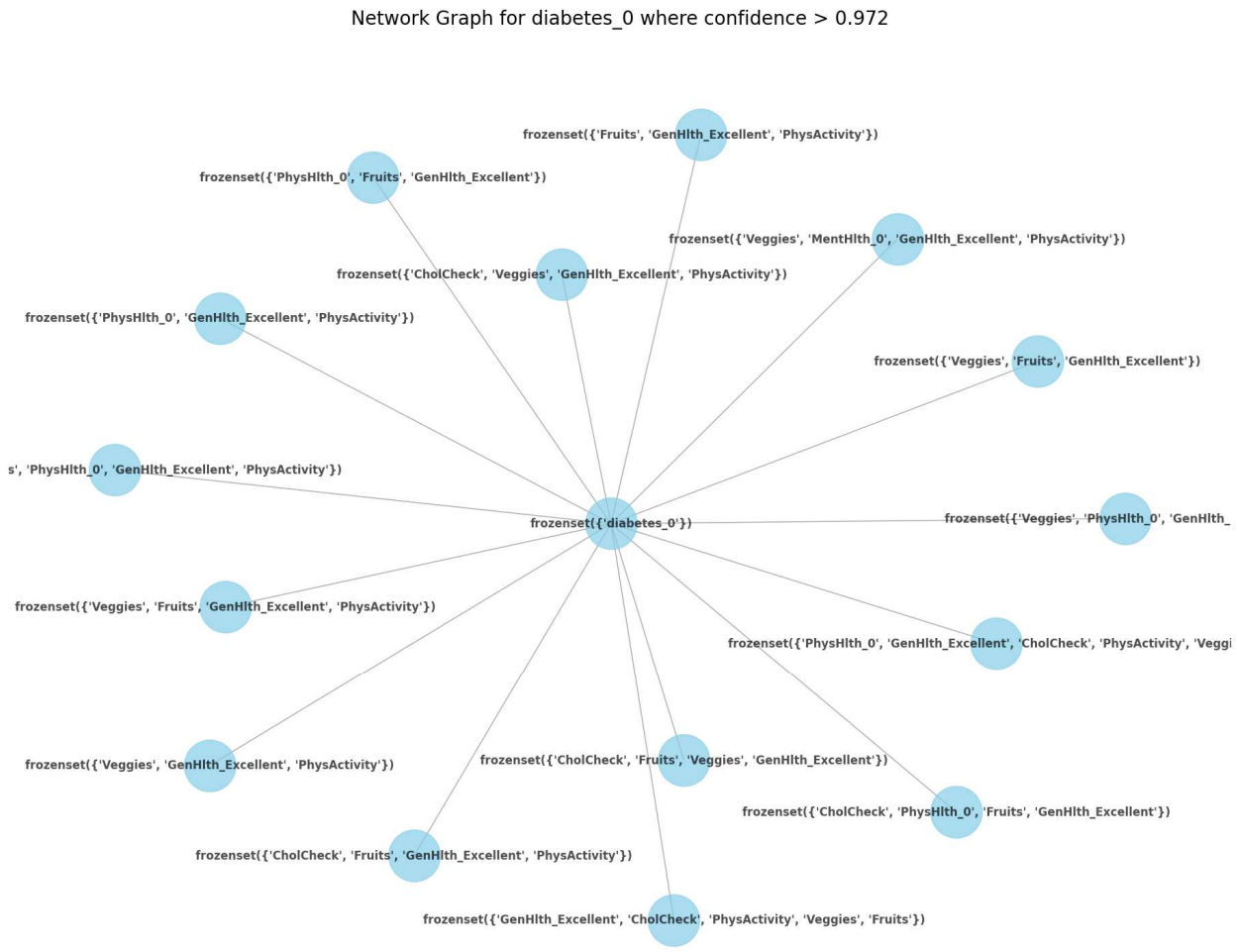


Figure 15. Network graph of association rules for RHS of diabetes_0 with confidence > 97.2%

The top three itemsets that were identified as the highest predictors for diabetes_0 (no diabetes/only during pregnancy) were “Veggies, Fruits, GenHlth_Excellent, PhysActivity” (support =0.106875, confidence = 0.976798, lift = 1.158945), ‘’GenHlth_Excellent, CholCheck, PhysActivity, Veggies, Fruit’’ (support = 0.101671, confidence = 0.975639, lift = 1.157571), and ‘’Fruits, GenHlth_Excellent, PhysActivity’’ (support = 0.114120, confidence = 0.974944, lift = 1.156746) followed by other itemset combining of the same features. The three itemsets that

were generated for diabetes_2 (diabetes diagnosis) involved HighBP as an antecedent and CholCheck. While association rules generated for diabetes_2 had a lower confidence level than those generated for diabetes_0 (a confidence of 0.24 as opposed to a confidence as high as 0.97), indicating that although the presence of the HighBP (alone or in combination with CholCheck) is less likely to result in diabetes_2.

Table 1. Association Rules Generated for RHS diabetes_2

	antecedent	consequents	support	confidence	lift
16	(HighBP)	(diabetes_2)	0.104959	0.245243	1.761116
303	(HighBP, CholCheck)	(diabetes_2)	0.104447	0.247774	1.779296
306	(HighBP)	(CholCheck, diabetes_2)	0.104447	0.244045	1.763198

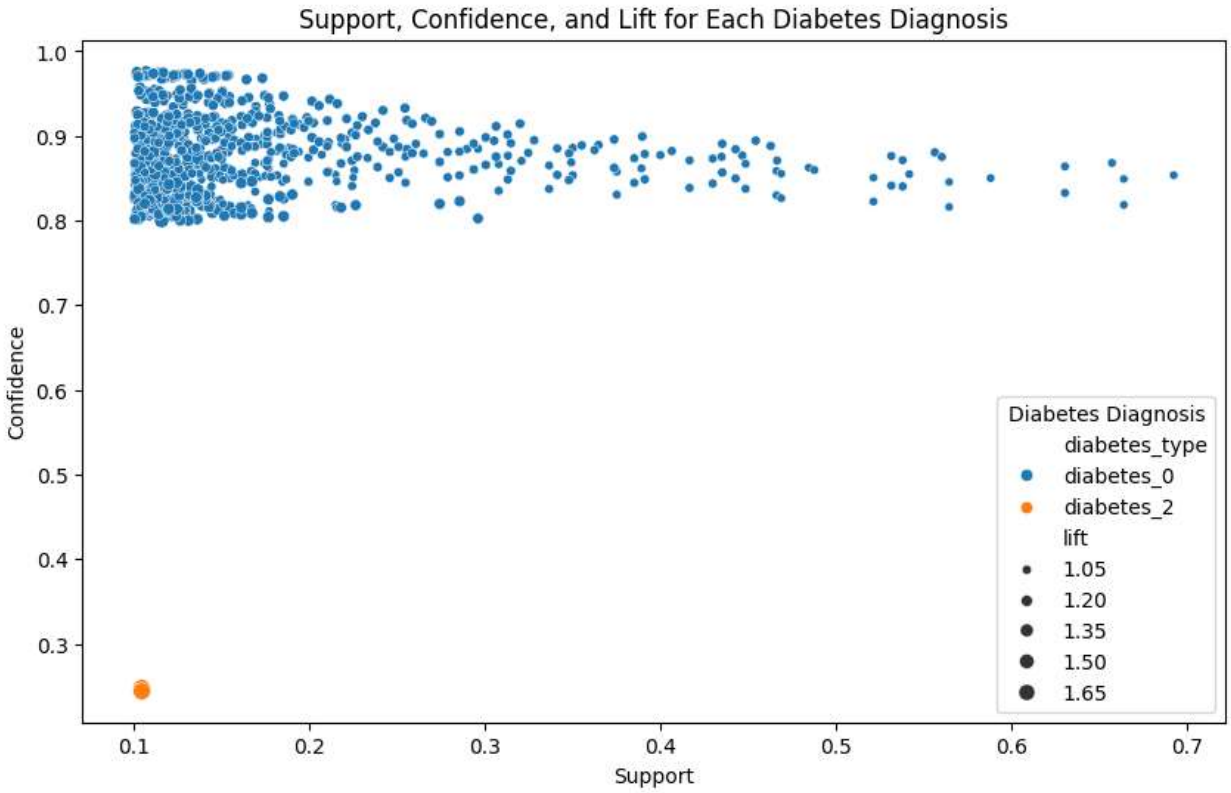


Figure 16. Scatter plot comparing association rules for RHS of diabetes_0 and RHS diabetes_2

The diabetes_2 association rules did have a higher lift, ranging from 1.76-1.78 as opposed to 1.01-1.02 (see Figure 17). Although a high lift is usually indicative that the association rule is stronger than if it were to occur by random chance, the combination of low support and low confidence values amplifies the lift value based on the lift calculation. The high lift value is therefore less reliable or actionable in this instance.

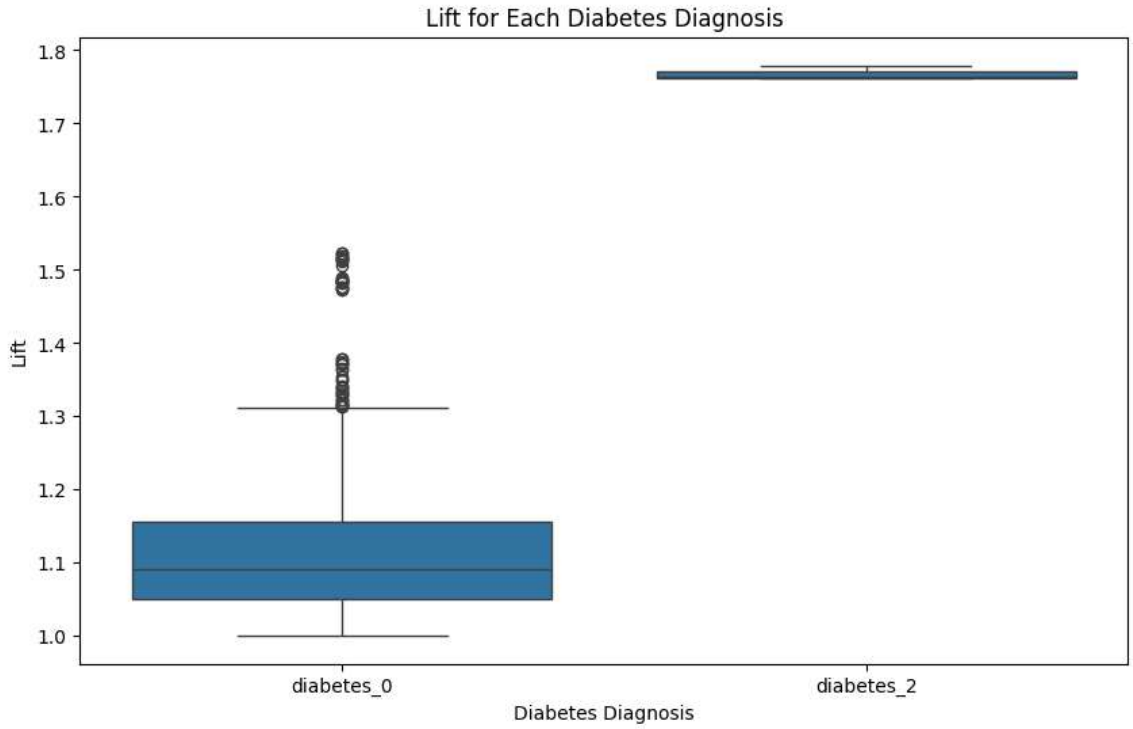


Figure 17. Box plot comparing lift for RHS of diabetes_0 and RHS diabetes_2 association rules

Question 2

The aggregated mean model performance metrics for each fold of the feature selection models applied to the initial train set are outlined in Figure 18. Further, the top 10 selected features based on importance for each k-fold for each feature selection technique are listed in Table 2, Table 3, and Table 4. Based on these initial results, we can conclude that the decision tree model has an overall higher accuracy, precision, recall, and F1 Score average across all classes as well as for the diabetes diagnosis class compared to the Chi-Square and the Logistic Regression. It can be concluded that based on the train set, the decision tree model is a better Feature Selection model. The standard deviation for all performance metrics is low indicating

stability across k-folds. While the order of selected features based on importance for the chi-square and the decision tree models remained the same across all five folds (see Table 2 and Table 3), the order of features selected for the multiclass logistic regression model showed inconsistencies across folds, indicating the logistic regression model is less stable.

Figure 18. Heatmap of Mean Model Performance Training Each Feature Selection Model

Table 2. Top 10 Features in Order of Importance for Each Chi-Square Fold

	Feature	Chi 2 Value	P Value
Fold 1	PhysHlth	16181.924106	0.0
	Age	12727.896405	0.0
	HvyAlcoholConsump	11022.642888	0.0
	HighBP	8519.057137	0.0
	GenHlth	7912.632458	0.0
	Income	6921.592790	0.0
	DiffWalk	6459.286504	0.0
	HeartDiseaseorAttack	5231.256222	0.0
	BMI	4545.632283	0.0
	HighChol	4479.467302	0.0
Fold 2	PhysHlth	16181.924106	0.0
	Age	12727.896405	0.0
	HvyAlcoholConsump	11022.642888	0.0
	HighBP	8519.057137	0.0
	GenHlth	7912.632458	0.0
	Income	6921.592790	0.0
	DiffWalk	6459.286504	0.0
	HeartDiseaseorAttack	5231.256222	0.0
	BMI	4545.632283	0.0
	HighChol	4479.467302	0.0
Fold 3	PhysHlth	15480.230170	0.0
	Age	12550.839112	0.0
	HvyAlcoholConsump	10625.038563	0.0
	HighBP	8463.896034	0.0
	GenHlth	7845.677928	0.0
	Income	6838.458088	0.0
	DiffWalk	6427.259736	0.0
	HeartDiseaseorAttack	5201.372218	0.0
	BMI	4515.520906	0.0
	HighChol	4502.390180	0.0
Fold 4	PhysHlth	15480.230170	0.0
	Age	12550.839112	0.0
	HvyAlcoholConsump	10625.038563	0.0
	HighBP	8463.896034	0.0
	GenHlth	7845.677928	0.0
	Income	6838.458088	0.0
	DiffWalk	6427.259736	0.0
	HeartDiseaseorAttack	5201.372218	0.0
	BMI	4515.520906	0.0
	HighChol	4502.390180	0.0

Fold 5	PhysHlth Age HvyAlcoholConsump HighBP GenHlth Income DiffWalk HeartDiseaseorAttack BMI HighChol	16363.438500 12458.396869 10764.000655 8497.754807 7905.941829 6958.928292 6602.150825 5308.423819 4625.569820 4401.917028	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

Table 3. Top 10 Features in Order of Importance for Each Decision Tree Fold

	Features	Importance
Fold 1	Age Income Education BMI PhysHlth MentHlth GenHlth PhysActivity Veggies Fruit	0.151525 0.140620 0.096402 0.087240 0.073678 0.060924 0.053252 0.040412 0.038634 0.038595
Fold 2	Age Income Education BMI PhysHlth MentHlth GenHlth PhysActivity Veggies Fruit	0.151000 0.142345 0.094398 0.086339 0.073675 0.060325 0.052071 0.043969 0.040084 0.035429
Fold 3	Age Income Education BMI	0.150403 0.141865 0.098419 0.086250

	PhysHlth MentHlth GenHlth PhysActivity Veggies Fruit	0.071083 0.061393 0.052346 0.042154 0.039962 0.036707
Fold 4	Age Income Education BMI PhysHlth MentHlth GenHlth PhysActivity Veggies Fruit	0.153526 0.143338 0.092841 0.085211 0.075023 0.058271 0.051449 0.041770 0.041106 0.038441
Fold 5	Age Income Education BMI PhysHlth MentHlth GenHlth PhysActivity Veggies Fruit	0.151315 0.144687 0.091946 0.082855 0.073579 0.059194 0.053430 0.043833 0.040921 0.039276

Table 4. Top 10 Features in Order of Importance for Each Multinomial Logistic Regression Fold

	Features	Coefficient
Fold 1	HvyAlcoholConsump Stroke CholCheck NoDocbcCost HeartDiseaseorAttack HighBP DiffWalk Smoker	0.929754 0.679349 0.455479 0.348613 0.306585 0.240425 0.221168 0.193460

	GenHlth Sex	0.188137 0.177527
Fold 2	HvyAlcoholConsump Stroke CholCheck NoDocbcCost HeartDiseaseorAttack HighBP DiffWalk Smoker GenHlth Sex	1.032062 0.746004 0.411101 0.354456 0.321150 0.240384 0.215149 0.186465 0.183876 0.171986
Fold 3	HvyAlcoholConsump Stroke NoDocbcCost CholCheck HeartDiseaseorAttack HighBP AnyHealthcare Smoker HighChol Sex	1.022899 0.615709 0.419631 0.357358 0.321219 0.259313 0.227317 0.221521 0.181926 0.180095
Fold 4	HvyAlcoholConsump Stroke NoDocbcCost CholCheck HeartDiseaseorAttack HighBP DiffWalk Smoker Sex GenHlth	1.120971 0.720465 0.341835 0.340903 0.270417 0.247929 0.241490 0.186071 0.182211 0.174303
Fold 5	HvyAlcoholConsump Stroke NoDocbcCost CholCheck HeartDiseaseorAttack HighBP DiffWalk Smoker AnyHealthcase	1.106387 0.685670 0.393492 0.334034 0.304121 0.259412 0.223754 0.206742 0.189355

	Sex	0.182441

Hyperparameter tuning for the decision tree in a final decision tree model with the following parameters: no max dept (max_dept=None), a minimum sample leaf of 1 (min_samples_leaf=1), and a minimum sample split of 2 (min_sampls_split=2). The parameters selected for the multiclass logistic regression model based on the GridSearchCV includes a C value of 0.01 (C=0.001) indicating strong regularization to help prevent overfitting and a penalty of L2 (penalty = L2) which penalizes large coefficients also useful in avoiding overfitting. The final decision tree models trained on the train set resulted in the same order of top 10 features selection as those from each k-fold except fruits and veggies which swapped places in the order of importance and included (starting with the most important): Age, Income, Education, BMI, PhysHlth, MentHlth, GenHlth, PhysActivity, Fruits, Veggies. For the final logistic regression model, the resulting top 10 features selected based on the absolute coefficients are (starting with the most important): HvyAlcoholConsump, Stroke, NoDocbcCost, CholCheck, AnyHelthcare, Smoker, HeartDiseaseorAttack, DiffWalk, HighBP, HighChol. The top 10 features selected for the final models trained on the train set are summarized in Figure 19.

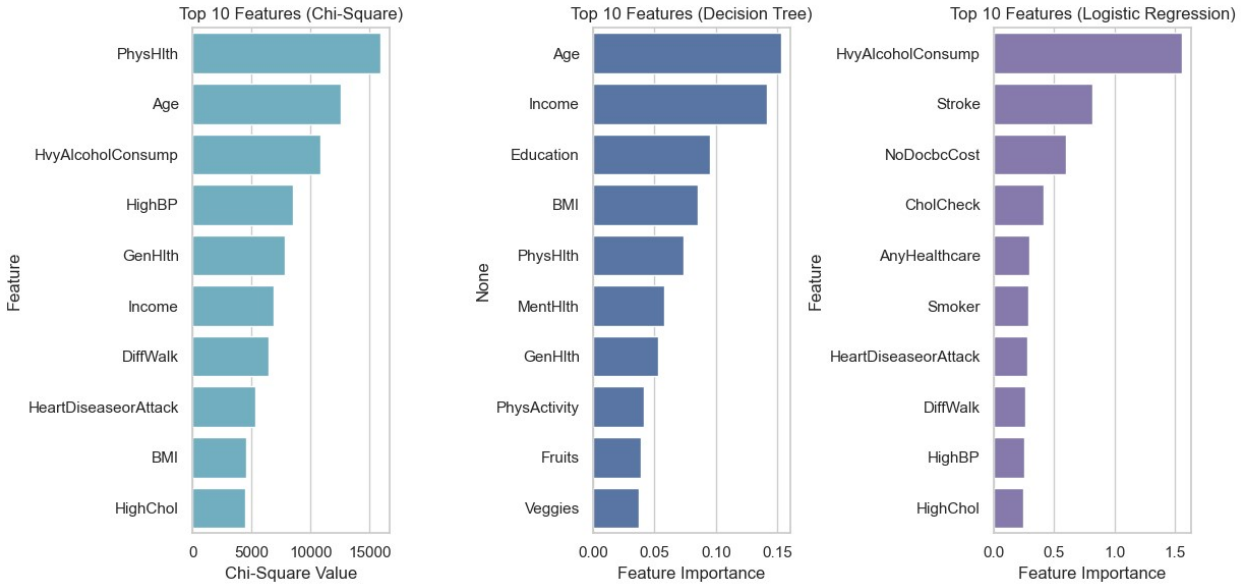


Figure 19. The Final Model's Top 10 Features Selected Based on Importance

The final models were applied to the test set and evaluation metrics were generated and are summarized in Figure 20. The Decision tree has the highest accuracy indicating that it is the model able to correctly predict positive and negative cases 63% of the time. Logistic Regression has the highest precision indicating that it is more accurate when predicting positive diabetes cases, reducing the risk of false alarms. Logistic regression also had the highest recall/sensitivity, making it the best model for ensuring we don't miss any diabetes diagnosis. Logistic also had the highest F1 showing it good balance between precision and recall. It also has the highest F2 score (which puts more emphasis on recall) indicating that it is a good model for identifying as many true positive diabetes cases. Logistic regression outperforms chi-square and decision when it comes to the sensitivity of the model (72%), further emphasizing it is the best model for catching as many positive cases as possible when looking not to miss any diagnosis. Specificity identifies how well the model identifies true negatives. Lastly, Logistic

regression also had the highest ROC AUC (0.67154) meaning it is the most effective model for distinguishing between the classes.

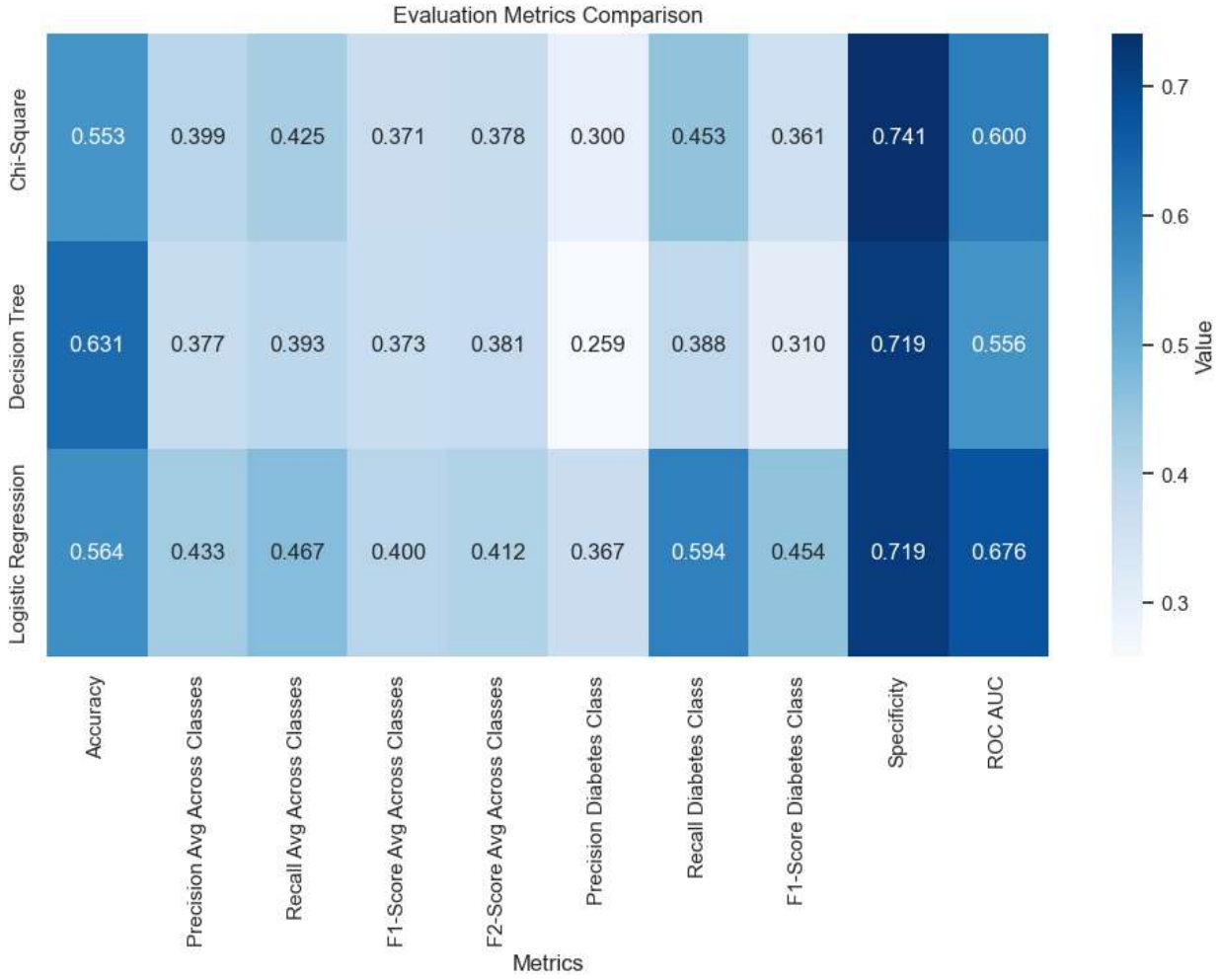


Figure 20. Heatmap of Evaluation Metrics for Final Models on Test Set

A side-by-side comparison of the initial evaluation metrics from the final model on the train set compared to those for the final model on the test set is also visualized in Figure 21. A significant drop in performance across all evaluation metrics can be noted for the decision tree

model once it was applied to the test set, suggesting that the model is unstable and that there may be overfitting. Recall for example dropped from 0.828 to 0.388 between the train and the test set for the decision tree model. The chi-square model applied on the test set also generally has lower performance metrics than when it was applied to the train set and experienced a drop of 0.619 to 0.453 for recall. The logistic regression model has a mix of metrics that are slightly higher, and slightly lower than those generated when applied to the train test overall indicating an increased generalizability compared to the other two models. The recall metric for the logistic regression model improved slightly from 0.548 on the train set to 0.549 on the train set. This metric of interest indicates an improvement in the model's ability to decrease false negatives,

ensuring that a diabetes diagnosis doesn't get missed.

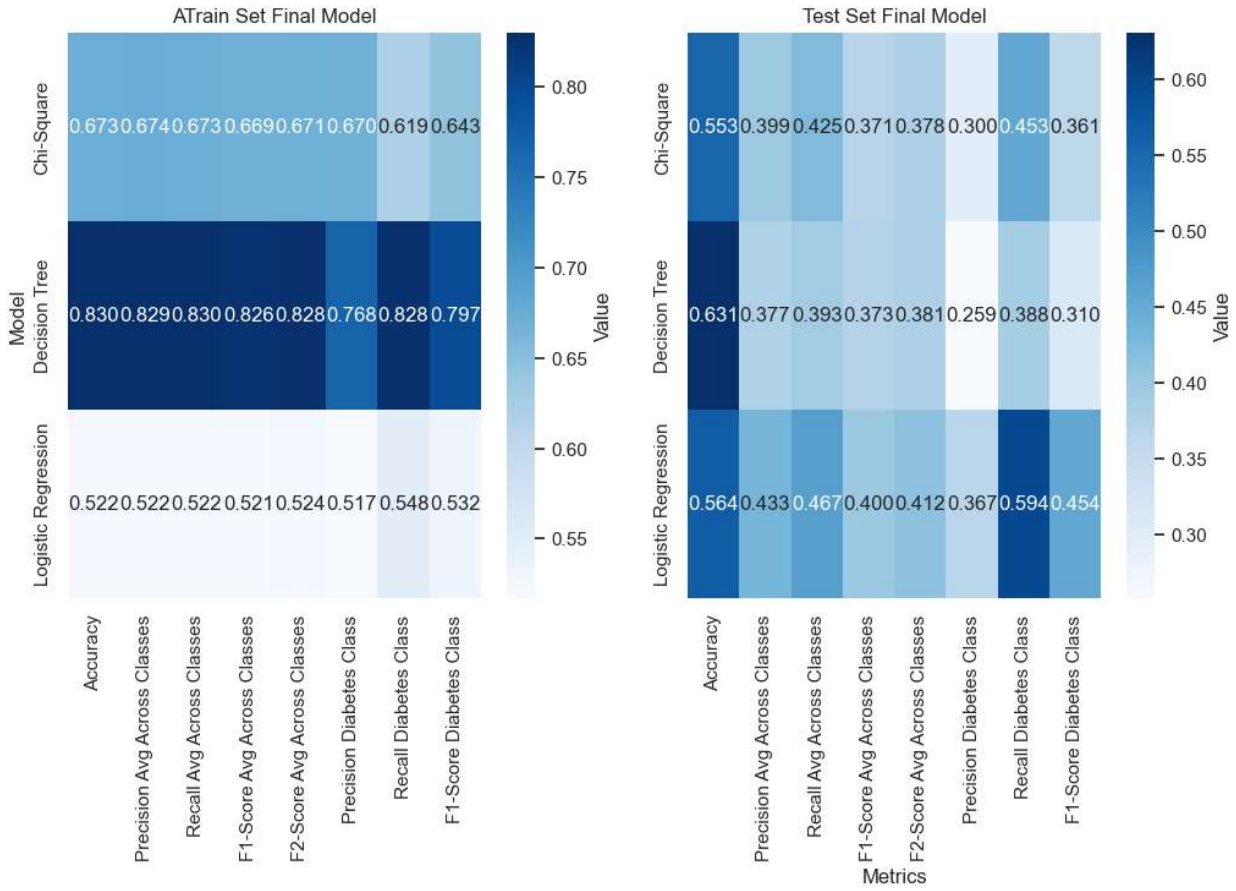


Figure 21. Side by Side Heatmap Comparison of Train set and Test set Evaluation Metrics

The Decision tree model had the lowest average ROC AUC value (0.566) suggesting poor discrimination for the diabetes class (there is only a 56.6% probability that the model will correctly rank a randomly positive class compared to a randomly selected one. The Logistic Regression model had the highest average ROC AUC value (0.676) and the highest AUC of 0.78 for the diabetes class (diabetes_2) indicating that it is the feature selection model with the highest true positive rate. The ROC AUC curves for each model are shown in Figure 22. Lastly,

the selected features based on the order of importance were the same for each final model applied to the test set as the order of those selected for the train set listed in Figure 19. Based on these findings, the Logistic Regression Feature Selection model was selected to use as a feature selection technique in Question 3. It has the best model because it had the highest recall on the test set, it is the model with the best generalizability, and the strongest overall discrimination.

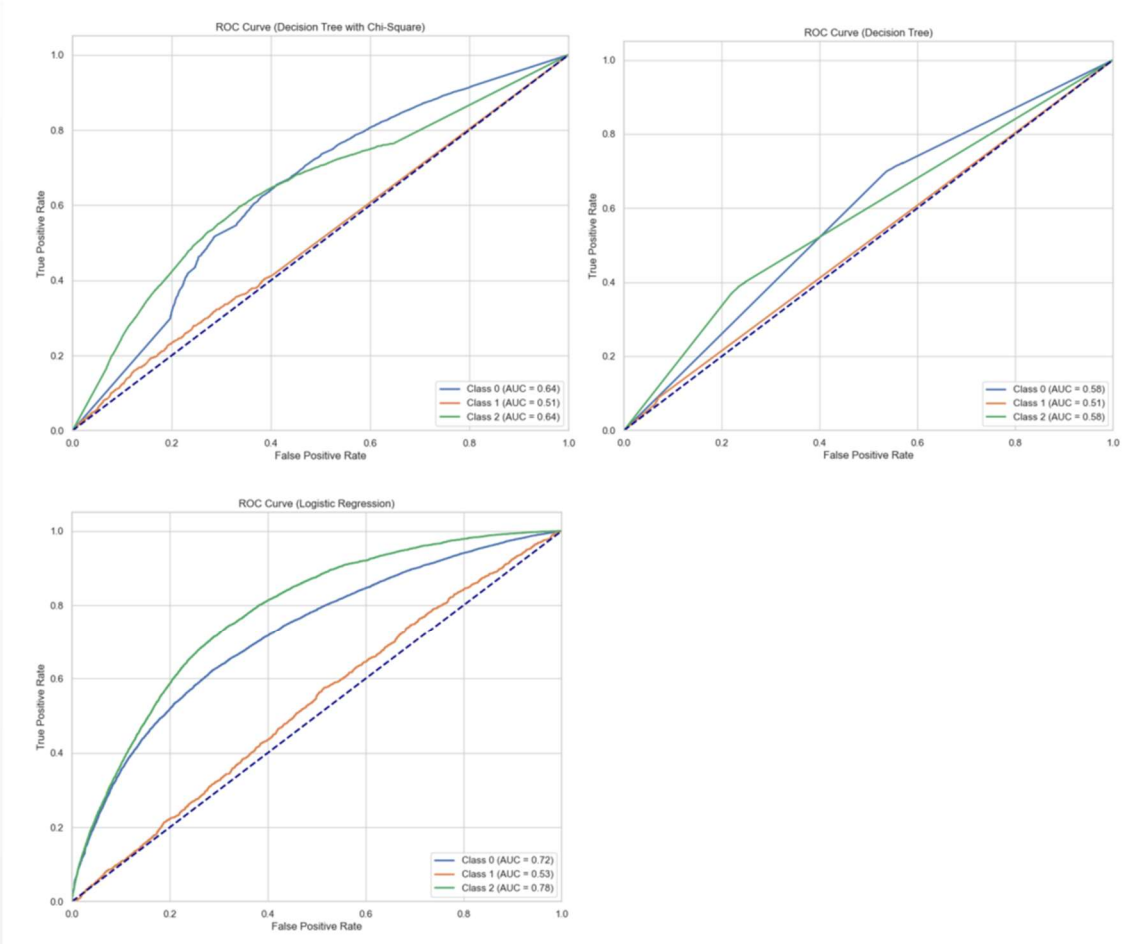
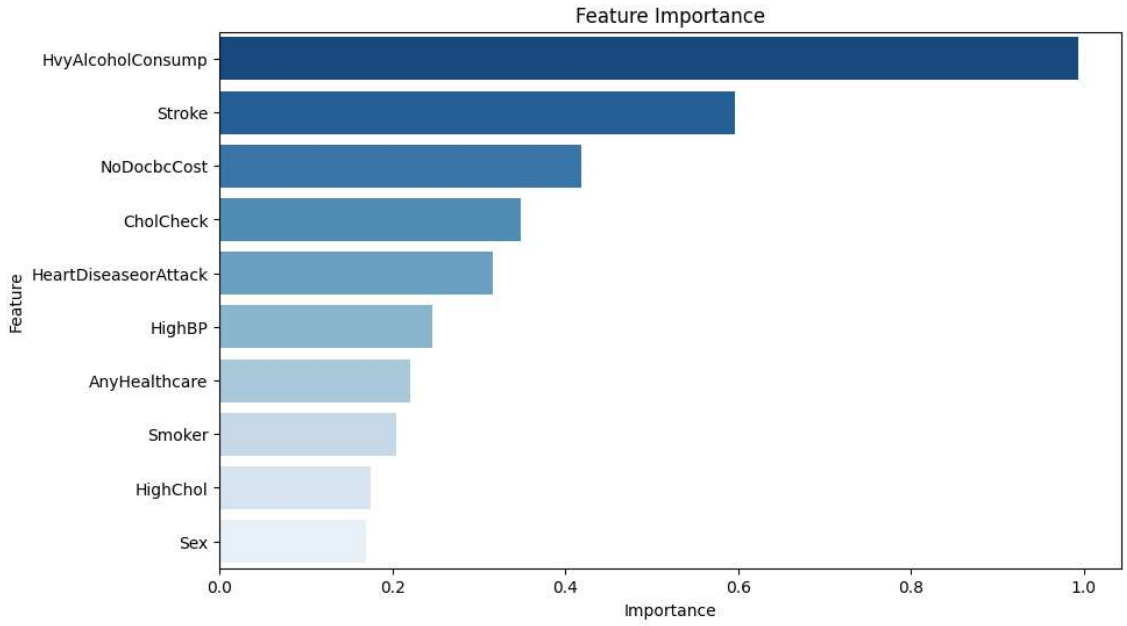


Figure 22. ROC AUC for each Feature Selection Model on the Test Set

Question 3

The final logistic regression model[11] from question 2 applied to the training set did not result in the same top 10 features selected by importance. The top ten features selected, in order from most important to least important based on the absolute value of the coefficients are listed in Figure 23. The performance metrics for each model trained on each k fold were visualized in Figure 24 to interpret the stability of each model. Based on the results, it can be concluded that the decision tree, logistic regression, and naive Bayes models appear to have stable performance metrics across each fold for all accuracy, precision, recall, f1 score, f2 score, and Roc Auc. The K-nearest neighbors model appears to be the least stable as it has the highest fluctuations across all performance metrics. Aggregate performance metrics are summarized in Table 5 and visualized in Figure 25.



[11] LogisticRegression(multi_class='multinomial', solver='lbfgs', C=0.01, penalty='l2')

Figure 23. Top 10 Features Selected Based on Feature Importance

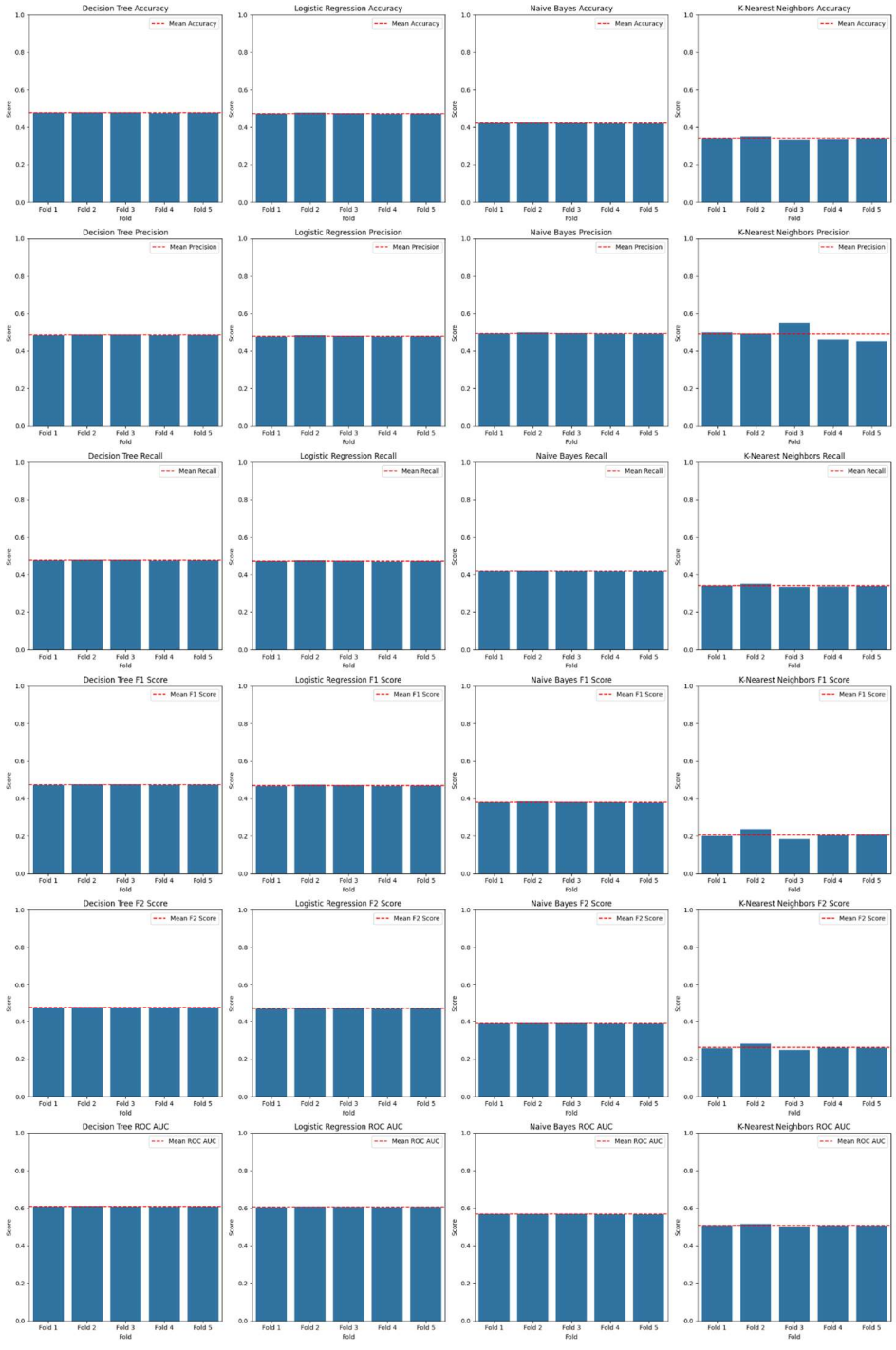


Figure 24. Performance Metrics for Each Model Across Each K Fold.

Table 5: Performance Metrics for Averages Across Classes for Each Model on Test Set

Model	Accuracy	Precision	Recall	F1 Score	F2 Score	ROC AUC
Decision Tree	0.478361	0.485829	0.478361	0.474352	0.474822	0.608770
Logistic Regression	0.474225	0.480251	0.474225	0.470400	0.474822	0.605669
Gaussian Naive Bayes	0.422686	0.494450	0.422686	0.381551	0.474822	0.567015
K-Nearest Neighbors	0.342521	0.491005	0.342521	0.206701	0.474822	0.506891

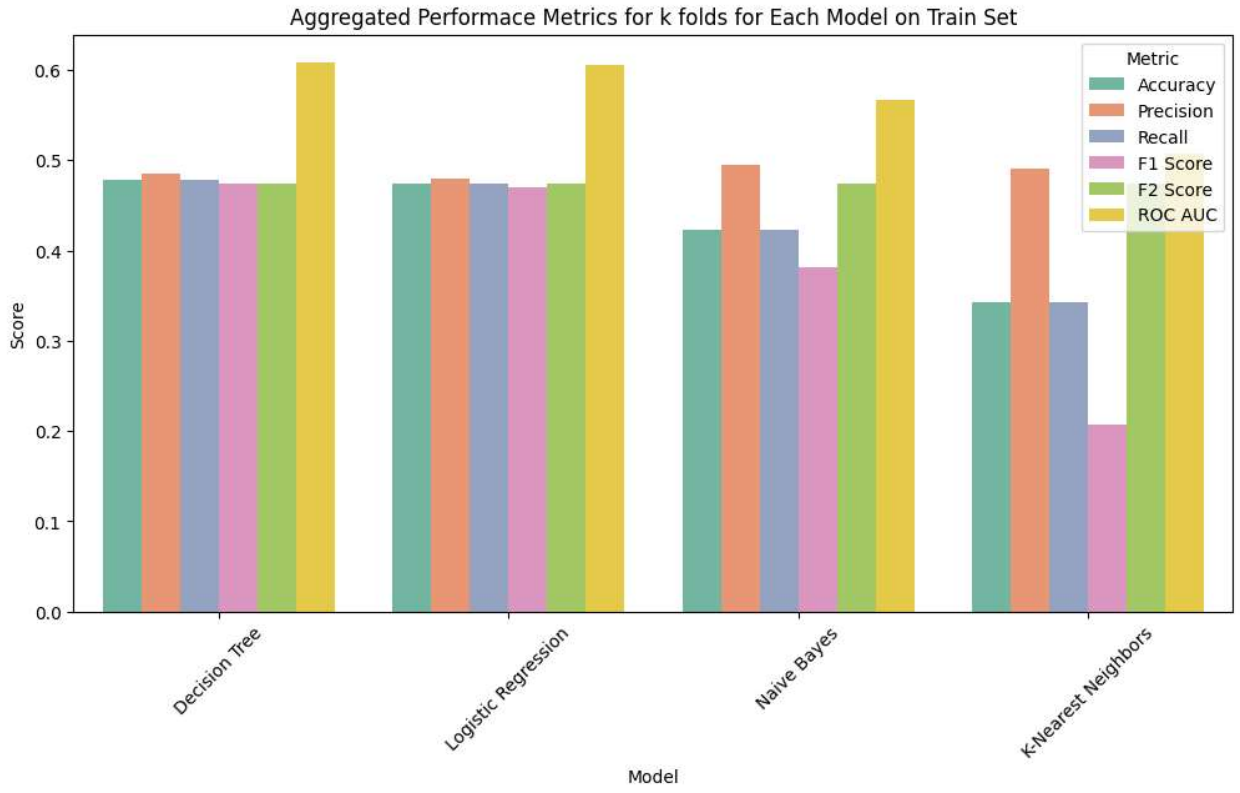


Figure 25. Aggregated Performace Metrics for k folds for Each Model

The final model for the decision tree based on the best hyperparameters from the were a criterion of 'gini', a max depth of 10, a minimum _samples_leaf of 1, and a minimum sample split of 10. The final model for the logistic regression has the parameters of 'C': 0.001, 'penalty': 'l2'. The best hyperparameters selected for the Naive Bayes model include 'var_smoothing': np.float64(0.11288378916846892). Lastly, the hyperparameters selected for the K-Nearest Neighbors model are 'n_neighbors': 3, 'p': 1, 'weights': 'uniform'.

The computation time, memory usage, and model size were evaluated during the application of each model to the test set and are summarized in Table 6. We can conclude from the results that the K-nearest neighbor is the least efficient model having the highest computation time and model size. Logistic regression and naive Bayes had similar computation times, with logistic regression having a slightly lower model size. The decision tree model had a short computation time of a little under half a second and a model size about 47 times that of the logistic regression and the Naive Bayes models.

Table 6: Efficiency Across Models

	Computation Time	**Memory Usage (bytes)**	**Model Size (bytes)**
Decision Tree	Computation Time: 0.4558 seconds	56	80881
Logistic Regression	Computation Time: 1.8290 seconds	56	1599
Gaussian Naive Bayes	Computation Time: 1.8261 seconds	56	1775
K-Nearest Neighbors	Computation Time: 1.9371 seconds	56	48795702

Results for the performance metrics of each model based on averages across all classes are summarized in Table 7 and visualized in Figure 26. While the KNN model had the highest accuracy of 0.77, it had the lowest performance for recall, precision, specificity, and ROC AUC scores. The Decision tree and the logistic regression models performed similarly across all evaluation metrics. While the Naive Bayes performed similarly to the decision tree and the logistic regression model for recall, precision, specificity and ROC AUC, it had the lowest performance in accuracy, F1-score, and F2-score compared to all other models. The decision tree model and the logistic tree model both had the highest recall of 0.42, the highest F2 Score of 0.35, and the highest ROC AUC of 0.62.

Table 7: Performance Metrics for Averages Across Classes for Each Model on Test Set

Model	Accuracy	Recall	Precision	F1 Score	F2 Score	Specificity	ROC AUC
Decision Tree	0.446600	0.423343	0.402784	0.334648	0.352213	0.741992	0.624778
Logistic Regression	0.451486	0.424390	0.402879	0.336974	0.354513	0.742294	0.628584
Gaussian Naive Bayes	0.250123	0.400384	0.415803	0.245504	0.255831	0.713311	0.617578
K-Nearest Neighbors	0.773690	0.344917	0.374780	0.334540	0.338929	0.677094	0.589180

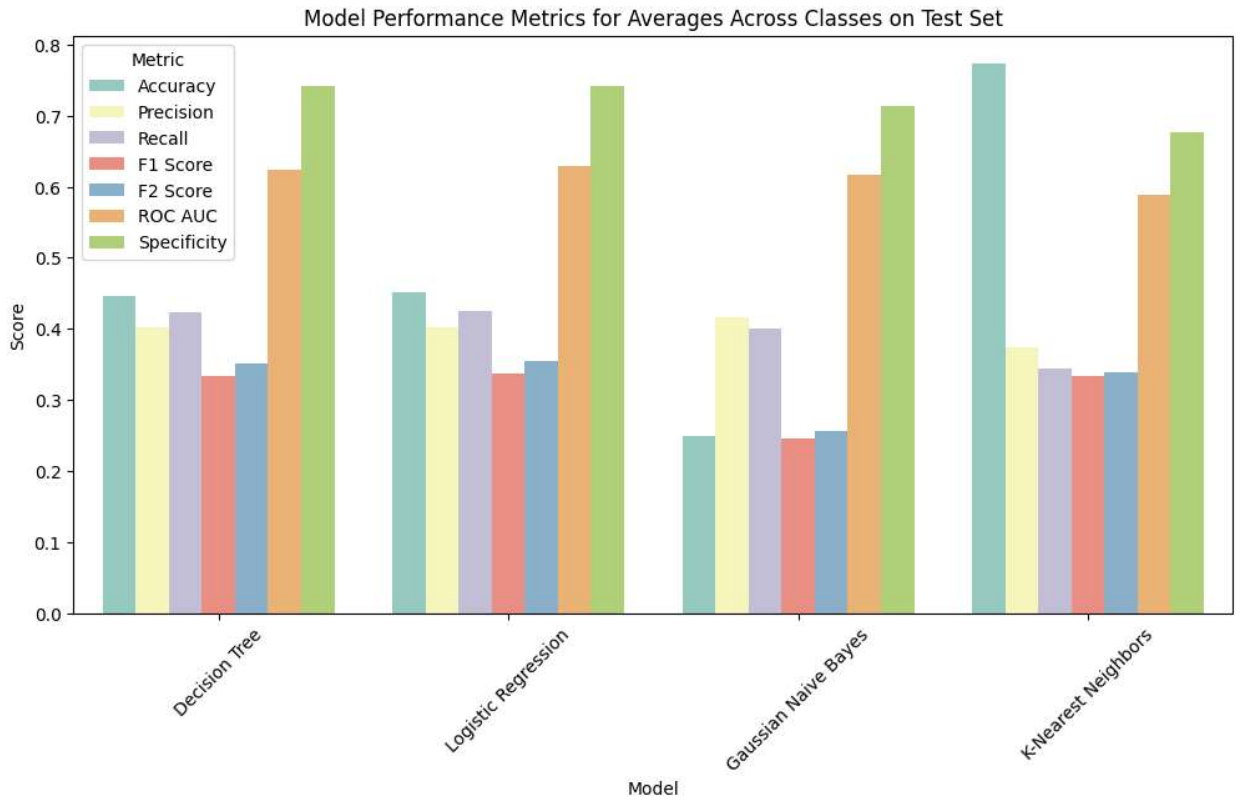


Figure 26. Model Performance Metrics for Averages Across Classes on Test Set

The difference between the performance metrics from the test set and the train set (test set - train set) were calculated and summarized in Table 8. All models experienced a slight decrease in accuracy, precision, recall, F1 score, and F2 score ranging between 0.01 and 0.2 during model application on the test set compared to the train set except for the KNN model which experienced an increase in accuracy of 0.43 and an increase in F1 score of 0.13. All models experienced a slight increase in ROC AUC ranging between 0.02 and 0.08. The least stable model is the KNN model as it experienced the greeted amount of performance metrics fluctuations between the train and test set. The decision tree model and the logistic regression model experienced the overall least amount of fluctuations on average and are the most stable. Lastly, because interpretability of the models is also of concern for clinical practice, a snapshot of the decision tree structure details as well as the logistic regression model are outlined in Figure 1 and Figure 2 of Appendix B.

Table 8. Differences Between the Test Set and Train Set Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score	F2 Score	ROC AUC
Decision Tree	-0.031761	-0.083045	-0.055017	-0.139705	-0.122610	+ 0.016007
Logistic Regression	-0.022740	-0.077372	-0.049835	-0.133426	-0.120309	+ 0.022915
Gaussian Naive Bayes	-0.172563	-0.078647	-0.022303	-0.136048	-0.218991	+ 0.050563
K-Nearest Neighbors	+ 0.431168	-0.116225	+ 0.002396	+ 0.127839	-0.135893	+ 0.082289

Question 4

There were four empty subsets for which there was no data to apply principle component analysis: Subset_1, Subset_13, Subset_25, and Subset_37. The number of principal components needed to reach an explained variance threshold of 90% for each subset is summarized in Figure 27. The subset with the lowest number of principal components needed to reach 90% variance was Subset_5 which consisted of overweight males ages 18 to 29 while Subset_36, consisting of obese females ages 50 to 64 had the highest number of principal components (27) needed to reach a 90% variance. It was determined that the average number of principal components needed to reach a 90% explained variance is 23.61. The top 5 features based on cumulative loadings selected for each subset using the number of principal components needed to reach a 90% threshold are summarized in a Heatmap in Figure 28.

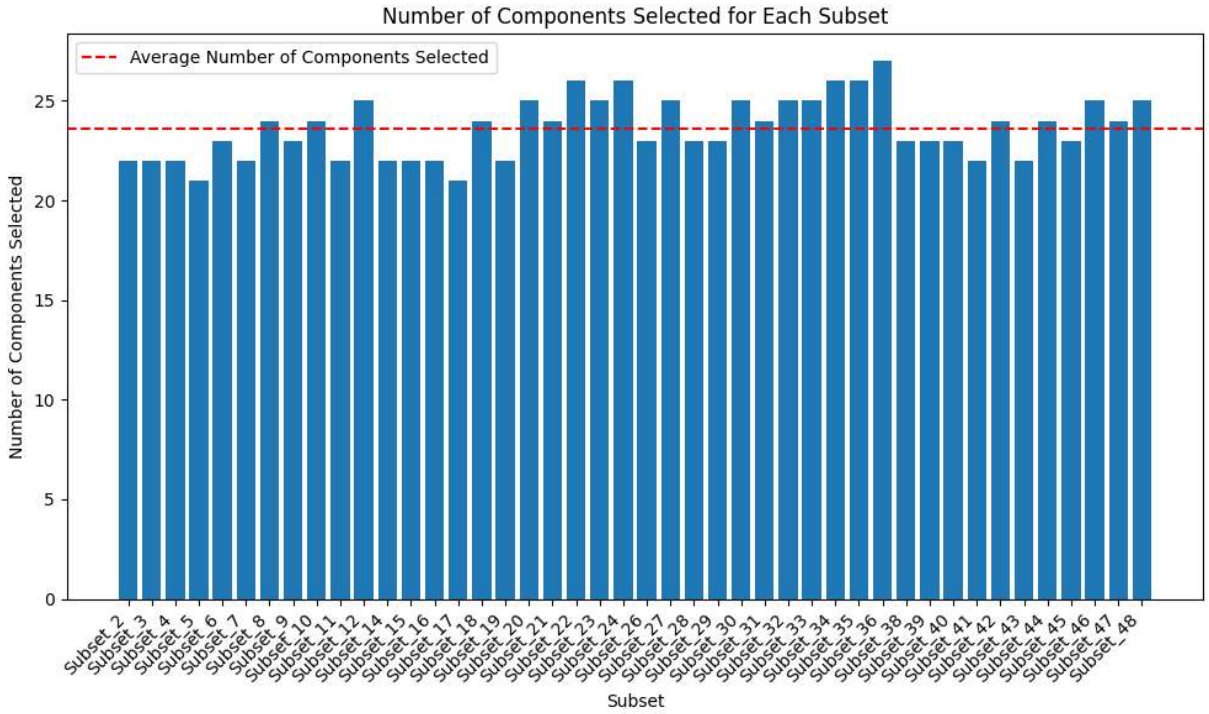


Figure 27. Number of PCs Selected for Each Subset based on a 90% Explained Variance

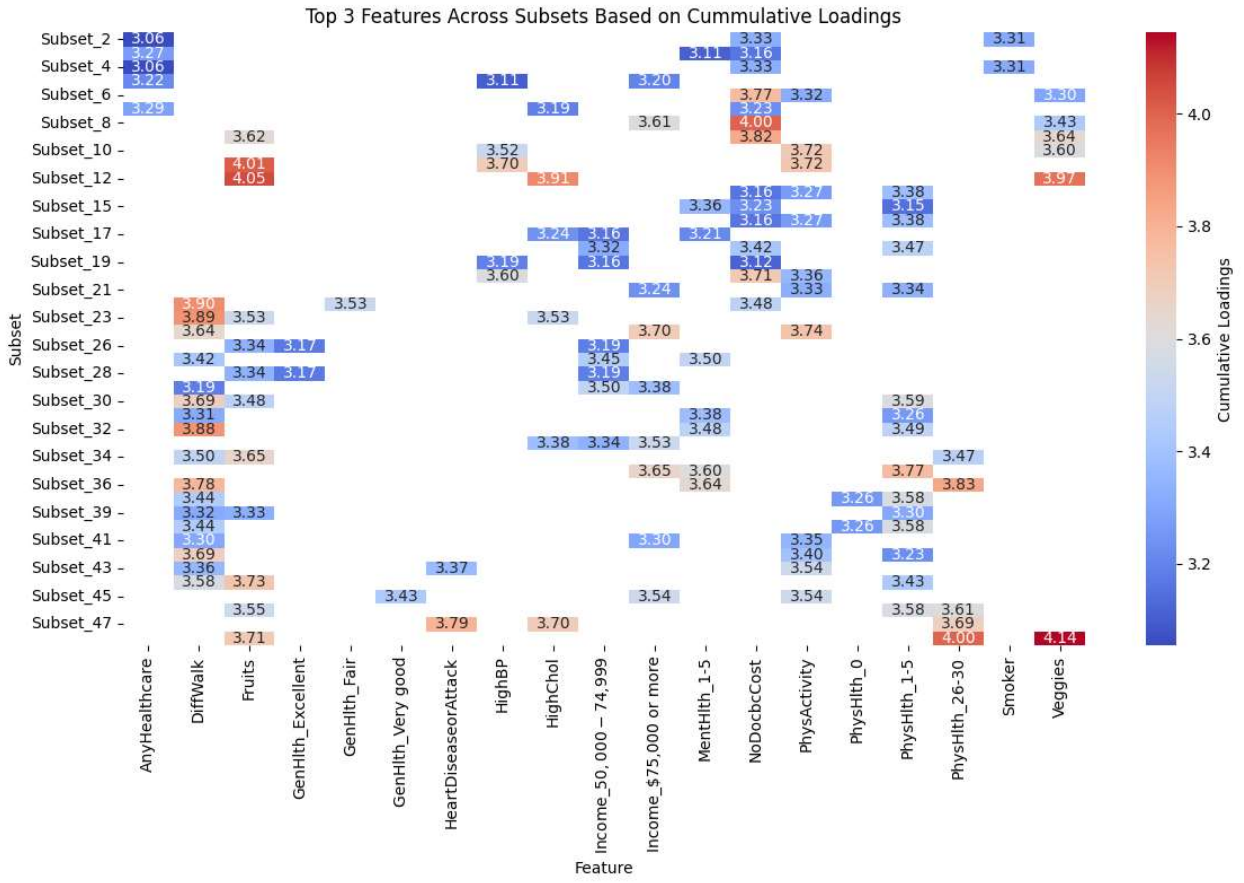


Figure 28. Top 5 Features Across Subsets (#PCs selected based on 90% variance threshold) Based on Cumulative Loadings

Select patterns can be seen based on the top 3 features with the overall highest loadings for the subsets using the number of principal components necessary to explain 90% of the variance in the subgroups. The selected top 3 features can therefore be said to have a stronger influence on the principle component, therefore having a stronger influence in explaining the variance at the 90% threshold. Some select features appear to have a greater influence based on age. AnyHealthcare, NoDocbcCost, and DiffWalk only appear as one of the top 3 characteristics between ages 18 to 29, ages 18 to 49, and ages 49 to 65+. HighBP was only selected as one of the top 3 features for subsets representing ages 18 to 49 and features for PhysHlth ratings were only selected for ages 30 plus. Out of the subsets that were selected Fruit was one of the top 3 features with the highest loading, 8/10 were female subsets across all

ages of which 6 represented a BMI category of overweight or obesity. On the other hand, 5/6 of the subsets with HighChol as a selected feature were for overweight and obese males across ages. Features like HeartDiseaseorAttack were only selected as one of the top three features for subsets 43 and 47 representing which were for males ages 65+ with obesity 1 and obesity 3. Smoker was only selected for subsets 2 and 4 representing females ages 18-29 who are a healthy weight or underweight. Income of 50,000 to 74,999 a year appears to have more influence on the principle components in ages 30 to 64 across gender and BMI category while Income of $75,000 or more has an influence on overweight and obese individuals across all ages, of which 6/9 were female.

Discussion and Result Implications

Question 1

The itemsets that served as the highest predictors of no diabetes here were ‘’GenHlth_Excellent, PhysActivity, Fruits, Veggies,’’ and ‘’GenHlth_Excellent, PhysActivity, Fruits, Veggies, CholCheck’’ with both having a confidence level of 97.5, a support of 0.10, and a lift of 1.16. These findings indicate that individuals would have self-perceived their health as being excellent in combination with health behaviors involving partaking in physical activity within the past 30 days, consumming fruits and vegetables at least once or more times per day, and having had cholesterol levels checked by a doctor within the past 5 years. The low support of 0.10 and the high confidence level of 97.5 indicate that although this combination of health indicators does not occur very frequently in the population (it only appeared in 10% of the dataset), together they are highly associated with health and are high predictors of health (or having no diabetes or diabetes only during pregnancy). The lift above 1 is an indicator of a strong relationship since the association is more likely to occur together than by random chance. These insights can be used by clinicians when making recommendations as to what combination of lifestyle factors to adopt to stay in good health and avoid diabetes.

Results from the association rules for the diabetes diagnosis outcome were very limited. Only 3 association rules were generated revealing an association between diabetes and high blood pressure or high blood pressure in combination with a cholesterol check. The association has a very low confidence level of 24%, indicating that high blood pressure and the combination of high blood pressure and cholesterol check only had a diabetes diagnosis as the consequent 24% of the time. This indicates that the itemset is a weak predictor of the diabetes diagnosis. Further, the association rules also had a low support of 10% indicating it was only observed in the 10% of the sample studied. The high lift value of 1.7 does not provide any further insights that could be extracted as the high value is only an amplified result due to the combination of both low support and confidence. As a result, there are no significant insights to be drawn from the association rules to put into practice in clinical settings.

Question 2

Chi-square (filter method), decision tree (wrapper method), and multinomial logistic regression(embedded method) feature selection techniques were applied to a train set using k-folds cross validation (5 folds), and compared to their application on a test set to identify the best method based on performance metrics, stability, and generalizability. The decision tree and chi-square had selected the same order of top 10 features across all 5 folds during the application of the model on the train set indicating that they are highly stable models. Although the logistic regression model was consistent most of the time, there were slight variations between the top 10 selected features between the 5 folds indicating that the model may be less stable. The average performance metrics across classes for the feature selection models on the train set were greatest across the board for the decision tree model with scores ranging between 0.768 to 0.830, followed by the Chi-Square model with scores ranging between 0.619 to 0.674, and lastly the Logistic regression model with scores ranging between 0.517 to 0.548.

When the models were applied to the test set to check for generalizability, the performance metrics for the decision tree models dropped greatly across the board by about 0.5 and ranged between 0.259 and 0.393 except for accuracy which scored 0.631. This is a sign that the decision tree model may have overfitted the train model. This overfitting may have occurred as a result of the hyperparameter tunning only being conducted after training the model on each k-fold. As such, there were no set parameters for the decision tree which allowed it to memorize patterns during training. The Chi-square model also experienced a drop in performance metrics by just a little less than half of the original scores except for accuracy. The logistic regression model experienced a slight increase in scores for accuracy, and recall, and while other performance metrics did experience a slight drop, the model was the most stable. This increase in score for recall is of special interest because the selected model must seek features that are important to ensuring that a diabetes diagnosis (a true positive) does not go missed (as a false negative). Recall is therefore an important metric to consider for clinical practice as early diagnosis and detection are important for treating the disease to prevent complications that could arise from untreated or missed diagnosis. Because the logistic regression model's performance metrics experienced the least amount of fluctuation between the train set and the test set, it can be concluded that it is the model with the best generalizability. Generalizability is especially important to consider in this context because it is a representation of how the models will perform on unseen data such as in the real world in clinical settings. Logistic regression is also a model that selects features based on coefficients. The ability to determine the features selected as most important based on the absolute values of the coefficients means that the model can be interpreted by healthcare practitioners in clinical practice. Based on the overall generalizability, interpretability, and performance metrics of the

models based on the test set (representing a real-world scenario of unseen data), it can be concluded that the logistic regression model is the best feature selection model.

All final models selected the same top 10 features for the train set as the test set. Most of the selected features deemed as best predictors of the diabetes class align with findings from the literature review. Research by Fregoso-Aparicio et al (2021) revealed that a combination of lifestyle, socioeconomic, and diagnostic data generally produced better predictive models. While not all feature selection models selected the same features based on importance, they did all select a combination of lifestyle, socioeconomic, and diagnostic data that contribute to predicting the disease class. The literature review revealed BMI and obesity as a common diabetes predictor (Kodama et al.,2022; Kumar et al., 2023). While the BMI category was one of the top 10 selected features for the chi-square and the decision tree model, it was not one of the selected features by the multinomial logistic regression model. Age was identified as the top important feature by the decision tree model and the second most important feature by the chi-square model, further supporting findings by Kodama et al. (2022) who identified age as a common predictor. Age was however not one of the top selected features by the logistic regression model. Kumar et al. (2023) identified the illness as a common diabetes predictor, a finding further supported by all feature selection models that listed health conditions such as high blood pressure, heart disease or attack, high cholesterol, stroke, general and mental health as features important towards predicting the diabetes class. While Kodama et al. (2022) identified physical activity as a rarely selected diabetes predictor, physical activity or difficulty walking or going up the stairs was identified as a top 10 feature important towards predicting diabetes by all feature selection models. The role of physical activity and movement on diabetes diagnosis should therefore be further investigated in future studies.

Question 3

The features selected by the final multinomial logistic regression feature selection model to the train set in order of most important to least important included: Heavy Alcohol Consumption, Stroke, No Doctor visits because of cost, whether cholesterol levels were checked within the past 5 years, access to healthcare coverage, smoker, ever had a heart disease or attack, have serious difficulty walking or going up stairs, have high blood pressure, and have been diagnosed with high cholesterol. Unlike the common diabetes predictors discussed during the literature review, age and BMI were not included.

The Naive Bayes model was deemed the worst-performing model as it had poor stability across performance metrics from the k folds during training on the train set. It also had the highest computation time of 1.9 seconds and was the largest model size. The model experienced the highest fluctuation in performance metrics between the train set and the test set indicating poor generalizability. The model also had the lowest recall and ROC AUC score on the test set, indicating a higher number of false negatives (or missed diagnoses), and is less capable of distinguishing between positive and negative classes.

The decision tree model and the logistic regression model were both selected as the best diabetes prediction models. Both models had good stability based on the consistency of the performance metrics across each fold on the train set. Both models also had similarly high-performance metrics across all metrics for both the train set and the test set (except for KNN which had the highest accuracy on the test set) with only the occasional 0.01 difference in score. While the decision tree has a slightly lower computation time (0.45 seconds) and higher model size (80881 bytes), the logistic regression has a slightly higher computation time (1.83 seconds) and lower model size (1599 bytes). Both models overall have a good performance,

however the tradeoffs lie with the decision tree having better computation time while the logistic regression model is smaller in size. Both the decision tree model and the logistic regression model are transparent and interpretable, however the logistic regression model is more easily interpretable because the model isn;t as big as the decision tree and feature importance for predicting the diabetes class are based on the coefficients. A higher absolute value for the coefficient means the feature holds more weight in helping predict the diabetes class. The decision tree model is a much larger size and is a scrollable element that may take longer to go through and interpretate.

These findings are somewhat consistent with the literature review which often identified that decision trees were the most commonly identified as the best model (Abdulazeem et al., 2023; Fregoso-Aparicio et al., 2021; Olusanya et al., 2022). Unlike Zhang et al. (2022) who found that non-logistic regression models tend to perform better than logistic regression models, the results from this project indicate that logistic regression is still considered a good option for a machine learning model that requires transparency and has high interpretability.

While the decision tree and the logistic regression models had the highest pooled performance metrics across classes, the scores obtained for their performance appear lower than those obtained in other research studies. The pooled AUROC for both the decision tree and the logistic regression on the test set were 0.624778 and 0.628584 respectively, scores much lower than those obtained by Zhang et al. (2022) who observed a pooled AUROC of 0.889 for non-logistic regression models and pooled AUROC of 0.8151 when examining 25 diabetes predictive machine learning models. Olusanya et al. (2022) also observed a high pooled accuracy of 0.88 for non-linear predictive machine learning models such as decision

trees, a score much higher than the 0.446600 (decision tree) and the 0.45148 (logistic regression) obtained in this project. Performance metrics may vary greatly from those from past studies due to differences in the size and the types of features evaluated in the dataset.

While models were validated during training using k-fold cross-validation to help minimize bias through repeated training on different folds, this method does not validate the model's performance in a true clinical setting in practice. While feature selection and the model's decision process are transparent and interpretable for clinicianès use in clinical settings, the selected features may not be a true reflection of the health information regularly collected by physicians when making a diagnosis in clinical practice.

The application of machine learning for predicting diabetes mellitus has policy implications as it could help identify the incidence and monitoring of the disease based on risk factors. It is a tool that could further relieve pressure on the healthcare system by limiting the overprescription of invasive clinical tests, which are often costly and time-consuming (Olusanya et al., 2022). Gaining a deeper understanding of the interactions between some of the identifying disease predictors could help further refine screening tools and program design for early intervention for identifying individuals at high risk of the disease.

Question 4

Results from the principle component analysis revealed that health indicators such as access to health coverage or accessed healthcare based on cost appear to have a greater influence on explaining 90% of the variance in the diabetes diagnosis outcomes in younger populations ages 18 to 29 and 18 to 49 respectively. Whether individuals have been told they have high blood pressure also had a greater influence between the ages of 18 and 49. These

observations may be explained by whether or not younger individuals access preventative care to catch and identify early signs and symptoms of the disease before it progresses to become a greater issue. Programs and campaigns targeted at younger populations to help increase their access to preventative and diagnostic services such as pop-up free clinics or free online consultations could help relieve some of the financial burden experienced by Canadians who don't have access to the provincial care plans. In Canada however, the majority of Canadians do have access to their provincial healthcare plans. This concern for Canadians can therefore be translated to the lack of access to health professionals due to long waiting lists for family doctors. Increasing incentives for doctors and nurses to have more professionals and open more clinics would help accommodate the long waiting lists of individuals with health insurance to receive preventative and diagnostic services. Ensuring that all healthcare professionals, including dental professionals or other facilities regularly visited, take blood pressure and advise when blood pressure is high to see a doctor, could help encourage younger individuals to seek care early.

Health indicators such as reported difficulty walking or climbing stairs, as well as reported levels of physical health appear to have a greater influence on explaining the variance in the diabetes outcomes in populations above the age of 30. This observation may be explained by the common increase in health complications or accumulated injuries that arise later in life. Some health indicators proved to be most influential in explaining 90% of the variance in the diabetes outcome in specific combinations of age, sex, and BMI category subsets. Income levels of $75,000 or more tend to have a greater influence on obese and overweight individuals across ages. Whether blood cholesterol is determined to be high is more influential in overweight and obese males. Whether an individual has experienced heart disease

or a heart attack proves to be most influential in males ages 65 and over with obesity. Answers to whether or not an individual smokes only seemed to have a greater influence in explaining variance in the diabetes outcome in healthy weight and underweight females ages 18 to 29. Health professionals in clinical practice can include these age, sex, and BMI class-appropriate probing questions when interviewing individuals that fall into the population subset, and be mindful of the influence that their answers may have on their health outcomes.

Limitations and Recommendations for Future Studies

These observations may be explained by whether or not younger individuals access preventative care to catch and identify early signs and symptoms of the disease before it progresses to become a greater issue. Programs and campaigns targeted at younger populations to help increase their access to preventative and diagnostic services

The dataset used for this study contained data on health indicators from an American population and does not appropriately reflect health indicators in the Canadian population to draw insights for Canadians. Future studies should be conducted on data from Canadian populations to determine any differences in population health indicators and needs that reflect Canadians. Features from the selected dataset were also limited to health conditions, lifestyle factors, demographics, and socioeconomic factors and did not consider environmental factors, family history of diabetes, or other genetics that may also influence disease diagnosis. Because diabetes mellitus is multifaceted, future studies must apply machine learning algorithms on datasets that consider a combination of factors such as genetics, lifestyle, and environmental factors to provide insight into complex and dynamic disease prediction. Ethical considerations

such as confidentiality, informed consent, and transparency should also be considered for personal health data collection on Canadian populations per the Personal Information Protection and Electronic Documents Act.

A lack of data also limited results when applying the apriori algorithm and principal component analysis. The lack of association rules generated for the diabetes class Question 1 may be attributed to the insufficient data due to the large class imbalance in the dataset. There was also a lack of data to represent subsets of underweight males ages 18 to 64 and healthy-weight males age 50 to 64 for principal component analysis in Question 4. Further, a very large number of principal components was needed to reach a 90% threshold for the explained variance of the components on the class variable. greatly limited the interpretability of the model for extracting meaningful insights. Insights were therefore only based on averages of loadings for each feature across principle components, therefore lacking transparency for using similar models in clinical settings. Because the literature review did not reveal any studies that similarly applied the apriori algorithm to generate and examine association rules or the application of PCA for the analysis of population subsets, it is recommended that future studies apply the apriori algorithm and PCA on datasets with a larger diabetes class and that represent a large population in each subset to extract insights on itemset's ability to predict the diabetes class and to uncover patterns in the population subsets.

As discussed during the literature review, there is a lack of consensus in the literature regarding the best models for diabetes prediction. To help strengthen model validation and consensus on key diabetes predictors. Future research should consider increasing the number of reported parameters to include evaluation metrics such as accuracy, sensitivity, specificity, precision, F1-score, F2-score, and ROC AUC to increase opportunities for benchmarking and

comparison between models (Fregoso-Aparicio et al., 2021). Special attention should be given to the recall, F2-score, and ROC AUC scores given that the goal of diabetes model prediction use in clinical practice settings is to ensure that models do not miss a diabetes diagnosis. Assessing model transparency and interpretability should also be included in discussions for selecting the best model as these are both important considerations for the application of the model in clinical settings.

Analysis techniques applied throughout this report can only inform on associations between the health indicators and the diabetes class variables, not causation. Although model validation techniques such as k-fold cross-validation have been applied in this study to the feature selection and machine learning models, there is a lack of validation of the models in clinical settings as this study did include a component for the application and testing of the models in a clinical setting. As discussed by Zhang et al. (2022), clinical needs rather than selected performance measures should be considered during feature selection to help build models that ensure the features used can be easily obtained during routine medicine. It is therefore recommended as part of this project continuity and for future studies that partnerships be formed with clinicians to put into practice the machine learning algorithms in healthcare settings to further validate whether models meet clinical needs based on a combination of evaluation metrics, transparency, and model interpretability as these are all necessary factors to build trust with clinicians for application of models in clinical settings.

Conclusion

There is a high prevalence of diabetes mellitus among Canadians. Early detection and diagnosis is the disease to help prevent diabetes progression and complications. While studies investigating machine learning models for predicting diabetes are widespread, there is no consensus on which models are best for use in clinical practice. This project investigated the health indicators and their associations with diabetes for predicting the disease. It was determined that a combination of lifestyle factors such as doing physical activity, eating fruits and vegetables, and regularly having cholesterol checked would help maintain health. High blood pressure was commonly observed in 10% of individuals with diabetes. All feature selection models selected a combination of lifestyle, demographic, and health conditions as diabetes predictors, findings that were consistent with the literature review. Multinomial Logistic Regression was identified as the best feature selection model, and the decision tree model and the logistic regression model were selected as the best predictive machine learning models based on overall effectiveness, efficiency, stability, and interpretability. Lastly, screening questions for use in clinical practice were recommended based on results identified from the principal component analysis for different combinations of age, sex, and BMI class population subsets. While findings for the best model were overall consistent with the literature review, more research is recommended on Canadian datasets with follow-up validation in clinical settings to further investigate diabetes prediction models for use in clinical practice.

Link to GitHub and the Working Dataset

https://github.com/stephbois/Big_Data_Analytics_Project.git

References

Abdulazeem, H., Whitelaw, S., Schauberger, G., & Klug, S. J. (2023). A systematic review of clinical health conditions predicted by machine learning diagnostic and prognostic models trained or validated using real-world primary health care data. *PloS one*, *18*(9), e0274276. https://doi.org/10.1371/journal.pone.0274276

Centers for Disease Control and Prevention. (2015, September). Behavioral risk factor surveillance system: Overview BRFSS 2014. CDC. https://www.cdc.gov/brfss/annual_data/2014/pdf/Overview_2014.pdf

Centers for Disease Control and Prevention. (2024, March). Adult BMI categories. CDC. https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

Das, A., & Dhillon, P. (2023). Application of machine learning in measurement of ageing and geriatric diseases: a systematic review. *BMC geriatrics*, *23*(1), 841. https://doi.org/10.1186/s12877-023-04477-x

Diabetes Canada. (n.d.a.). Diabetes in Canada. https://www.diabetes.ca/research-(1)/advocacy-reports/national-and-provincial-backgrounders/diabetes-in-canada#:~:text=Adult%20men%20are%20more%20at,type%202%20diabetes%20(11).

Diabetes Canada. (n.d.b.). What is diabetes?. https://www.diabetes.ca/about-diabetes-(3)/what-is-diabetes

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review.

Diabetology & metabolic syndrome, *13*(1), 148. https://doi.org/10.1186/s13098-021-00767-9

Google for Developers. (n.d.) Classification: ROC and AUC. Machine Learning. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Government of Canada. (2024a, June 1). Sustainable development goal 3: Good health and well-being. https://www.canada.ca/en/employment-social-development/programs/agenda-2030/health-well-being.html

Government of Canada. (2024b, June 17). Canada and the sustainable development goals. https://www.canada.ca/en/employment-social-development/programs/agenda-2030.html#sdg

Kodama, S., Fujihara, K., Horikawa, C., Kitazawa, M., Iwanaga, M., Kato, K., Watanabe, K., Nakagawa, Y., Matsuzaka, T., Shimano, H., & Sone, H. (2022). Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis. *Journal of diabetes investigation*, *13*(5), 900–908. https://doi.org/10.1111/jdi.13736

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of ambient intelligence and humanized computing*, *14*(7), 8459–8486. https://doi.org/10.1007/s12652-021-03612-z

Olusanya, M. O., Ogunsakin, R. E., Ghai, M., & Adeleke, M. A. (2022). Accuracy of Machine Learning Classification Models for the Prediction of Type 2 Diabetes Mellitus: A

Systematic Survey and Meta-Analysis Approach. *International journal of environmental research and public health*, *19*(21), 14280. https://doi.org/10.3390/ijerph192114280

Statistics Canada. (2023, November 29). Diabetes among Canadian adults. https://www.statcan.gc.ca/o1/en/plus/5103-diabetes-among-canadian-adults

Teboul, A. (2021). Diabetes health indicators dataset. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

UC Irvine Machine Learning Repository. (2023, September 9). CDC Diabetes Health Indicators. https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

Varga, T. V., Liu, J., Goldberg, R. B., Chen, G., Dagogo-Jack, S., Lorenzo, C., Mather, K. J., Pi-Sunyer, X., Brunak, S., Temprosa, M., & Diabetes Prevention Program Research Group (2021). Predictive utilities of lipid traits, lipoprotein subfractions and other risk factors for incident diabetes: a machine learning approach in the Diabetes Prevention Program. *BMJ open diabetes research & care*, *9*(1), e001953. https://doi.org/10.1136/bmjdrc-2020-001953

Zanelli, S., Ammi, M., Hallab, M., & El Yacoubi, M. A. (2022). Diabetes Detection and Management through Photoplethysmographic and Electrocardiographic Signals Analysis: A Systematic Review. Sensors (Basel, Switzerland), 22(13), 4890. https://doi.org/10.3390/s22134890

Zhang, Z., Yang, L., Han, W., Wu, Y., Zhang, L., Gao, C., Jiang, K., Liu, Y., & Wu, H. (2022). Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis. *Journal of medical Internet research*, *24*(3), e26634. https://doi.org/10.2196/26634

Appendix A

Table 1: Articles examining the use of machine learning for disease prediction, which included diabetes as one of the diseases.

Source	Research Focus	Population and Sample Size	Study Design	Methods, Measures, and Statistical Analysis	Main Outcomes Stated by Author	Implications for Discussion/Conclusion	Researcher Notes (Limitations and Follow-Up)

Abdulazeem et al., 2023	Examine health conditions being targetted by Machine Learning prediction models within primary healthcare.	106 primary studies across 7 databases addressing machine learning diagnostic or prognostic predictive models on 42 health conditions supplied by primary health care data	Systematic review	Quality assessment using PROBAST Risk of bias assessment applied	Diabetes mellitus is one of the most frequent health conditions targetted by machine learning prediction models Common models included random forest, SVM, extreme, light, and adaptive boosting, decision tree, Naive Bayes, k-nearest neighbors, and LASSO, and neural networks	Models trained by historical data can reinforce outdated practices non-linear models, SVM, and decision tree models provide more insight into complex and dynamic disease prediction	High-risk of bias for many of the studies selected, often due to lack of model validation No statistical analysis of model attributes

Das & Dhillon, 2023	Examine the application of machine learning towards aging-related concerns, including chronic diseases	70 articles across 2 databases including healthy aging individuals 45 and up	Systematic Review	Followed PRISMA and used JBI critical appraisal tool for quality assessment of study	Risk prediction was the most common machine-learning approach Logistic regression, random forest, XG Boost were frequently used methods applied to a variety of datasets including population-based surveys, hospital records, and digitally traced data	Machine learning has been extensively applied in the detection, prediction, and identification of risk factors for diabetes Common models include logistic regression, XG Boost, decision tree Risk factors identified using clustering algorithms like principle component analysis, logistic regression	Future research should look to algorithms that are fair, transparent, and validated before clinical implementation

						classifier	

Kumar et al., 2023	Examined different diseases and their diagnostic measures using machine and deep learning classification	158 studies over 6 databases	Survey directed according to preferred reporting items for systematic review and Meta-Analysis guidelines	Several quality evaluation constraints were applied for inclusion criteria.	Commonly identified diabetes predictors include glucose, insulin, BMI, stress, illness, medication, amounts of sleep, periodic heart rate Proposed classification methods for diabetes detection include preprocessing, feature extraction, machine learning, and classification. Precision, recall, accuracy, and F score are	Random forest classifiers, logistic regression, fuzzy logics, gradient boosting machines, decision tree, k nearest neighbor, and support vector machine are primarily used for disease detection Computationally effective feature selection is needed to increase accuracy. Models need to be validated on multiple sites	Insufficient data and small sample sizes are common challenges throughout selected articles

					commonly used for model evaluation.	to improve generalizability.	

Table 2: Articles examining use of machine learning for predicting diabetes mellitus and identifying its risk factors.

Source Citation	Research Focus	Population and Sample Size	Study Design	Methods and Measures	Main Outcomes Stated by Author	Implications for Discussion/Conclusion	Researcher Notes (Limitations and Follow-Up)

Fregoso-Aparicio et al., 2021	Identify opportunities for improving type 2 diabetes prediction via the selection of machine learning techniques	90 studies across 2 search engines.	Systematic review	Review followed PRISMA and the Guidelines for performing a Systematic Literature Review in Software Engineering. Applied quality assessment, data extraction, and assessed risk of bias	18 different types of models were included in the review There is no consensus on the type or the amount of features across studies, Lifestyle, socioeconomic and diagnostic data types generally produce better models SVM, RF, GBT and DNN were the most popular machine learning techniques	Heterogeneity among techniques used and lack of transparency of features reducing their interpretability Reporting five or more parameters (accuracy, sensitivity, specificity, precision, and F1-score) can enable benchmarking studies and models Machine learning models worked best on well-structured balanced	Difficulty with comparisons between models due to heterogeneity in the population and selected sample Also no concensus on evaluation metrics for reporting

					Decision tree and random forest wer the top performing models Most studies used metrics from confusion matrix to report on performance	dataset containing a mix of different types of features K nearest neighbors, and Support vector machines are frequently preferred for prediction	

Kodama et al., 2022	Investigate Machine Learning algorithm's ability to predict type 2 diabetes mellitus	12 studies comparing ML classification with actual diabetes incidence from 1950-2020 found in Medline and Embase	Systematic review of longitudinal studies	Extracted data and used QUADUS-2 to evaluate study quality Data was synthesized for each study using a confusion matrix to determin pooled sensitivity, specificity, positive likelihood ratio, and negative likely ratio.	Included classifiers were decision tree, neural network, k-nearest neighbor, logistic regression, support vector machine, random forest, reverse engineering and forward simulation The most frequently selected features were age, obesity, and blood glucose. Physical activity and family history were rarely selected	Future research should compare the ability to predict type 2 diabetes mellitus among ML algorithms, previously established risk models	Limited to comparing studies from high-income countries and limited to comparing studies that specified consistency across evaluation metrics

					feature		

Olusanya et al., 2022	Investigate soft-computing and statistical learning models ability to predict type 2 diabetes mellitus by pooling data of machine learning estimates	34 studies conducted in different countries between 2010 -2021 from 3 different search engines	Meta-Analysis	Searched Web of Science, Scopus, and PubMed and extracted and summarized data from selected studies Assessed methodological quality of studies based on the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool A chi-squared test used for heterogeneity Total heterogeneity/ variability	Learning algorithms applied by the studies included linear regression (8/34), decision tree (8/24) diabetes patient data, 6/34 applied neural networks, 3/34 applied random forest, 5/34 used support vector machine Mon-linear ML models outperformed linear models. Decision tree models were the most frequently used model	There's been an increase in the application of ML for diabetes diagnosis, but no consensus and high heterogeneity between studies. Decision trees are machine learning models with high accuracy for predicting type 2 diabetes mellitus ML models have policy implications for evaluating and monitoring disease	There is high heterogeneity variance of estimates across studies Heterogeneity between study populations and samples, and burden of disease across the different groups

				among studies assessed using Higgins I-squared (I^2) publication bias assessed using Begg & Eggers test and funnel plot	for predicting type 2 diabetes mellitus and had a pooled accuracy of 0.88 [95% CI: 0.82, 0.92]	Recommend comparing different ML models for diabetes prediction	

Varga et al., 2021	Evaluate the predictive ability of standard lipid measurements and NMR-measured lipoprotein size and concentration for incident diabetes	Data from the Diabetes Prevention Program (DPP) and the Diabetes Prevention Program Outcomes Study (DPPOS)	Secondary analysis of randomized control trial	10 Models internally validated using nested cross-validation framework Evaluated models: Apple Logistic regression, Cox proportional hazards model, gradient boosting, random forest, support vector machines with linear kernel (SVM-	The best predictive models included measures of glycemia Standard lipids or NMR-based lipoprotein size and concentration measures did not augment the predictive utility of models incorporating glycemia	Machine learning algorithms provided no meaningful improvement for discrimination compared with logistic regression, which suggests a lack of influential latent interactions among the analytes assessed in this study	Small sample size is a limitation of this study Analysis performed on data from individuals with pre-diabetes where risk factors may differ than that of general population Data from blood samples may be skewed due to long storage times

				L), polynomial kernel (SVM-P) and radial kernel (SVM-R), and artificial neural network (ANN)			

Zanelli et al., 2022	Provide detailed overview of published methods used for detecting and managing diabetes using PPG and ECG signals	78 studies across 4 databases focused on glucose estimation and diabetes detection	Systematic Review	Several previously applied machine learning approaches to distinguish between diabetic and healthy subjects, notably random forest, logistic regression, and decision trees with high-performing specificity, sensitivity, and accuracy	Both traditional and machine learning approaches require the feature extraction step to be done no feature extraction algorithm can work if the input signal is corrupted.	Machine learning techniques are promising in helping to detect and manage diabetes by analysing PPG and ECG The knowledge of why and how a pathology was detected is fundamental in order to validate the diagnosis There is a need to better explore signal processing and feature extraction, splitting of the	Lack of standardizatio n of the results. Difficult to compare studies due to heterogeneity

						data to minimize bias, increase parameters used when evaluating performance, and need for clinical validation of the models	

Zhang et al., 2022	To conduct a thorough meta-analysis and compare machine learning models for predicting the risk of gestational diabetes mellitus to universal and selective screening models	25 studies from 4 databases including women from the general population aged 18 years and over without a history of vital disease	Meta-Analysis	Prediction Model Risk of Bias Assessment Tool (PROBAST) to assess models' risks of bias. Sensitivity analysis, meta-regression, and subgroup analysis to limit heterogeneity. Models described using primary outcome measures of discrimination and calibration Meta-DiSc software used	Non-logistic regression models (AUROC of 0.8891) performed better than logistic regression models (AUROC 0.8151). Features most commonly selected by models include maternal age, family history of diabetes, BMI, and fasting blood glucose	Machine Learning models achieved high accuracy in early detection of gestational diabetes mellitus and could be used by clinicians to assist in early screening Researchers should test more models Important considerations include feature selection based on clinical need rather than accuracy (features	Some risk of bias present in selected studies. Comparison between models difficult due to differences in features, sample sizes, and distributions. Few models underwent external validation

				to measure pooled estimates of AUROC, sensitivity, specificity, PLR, NLR, and DOR		selected should be easily obtained during routine medicine)	

Appendix B

Table 3: Feature Descriptions and Data Types.

Name	Description	Value	Data Type
Diabetes_012	Been told they have diabetes.	0 = no diabetes or only during pregnancy 1 = prediabetes 2 = diabetes	categorical: ordinal (multivariate)
HighBP	Been told they have high blood pressure by a doctor, nurse, or other health professional.	0 = no high BP 1 = yes high BP	categorical: nominal (binary)

HighChol	Ever been told by a doctor, nurse, or other health professional that your blood cholesterol is high	0 = no high cholesterol 1 = yes high cholesterol	categorical: nominal (binary)
CholCheck	Have you had your cholesterol checked within the past 5 years	0 = no cholesterol check within the past 5 years 1 = yes cholesterol check within the past 5 years	categorical: nominal (binary)
BMI	Calculated Body Mass Index	1 - 9999 (has 2 implied decimal places)	continuous
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	0 = no 1 = yes	categorical: nominal (binary)
Stroke	(Ever told) you had a stroke	0 = no 1 = yes	categorical: nominal (binary)
HeartDiseaseorAttack	Coronary heart disease (CHD) or myocardial infarction (MI)	0 = no 1 = yes	categorical: nominal (binary)
PhysActivity	Physical activity in the past 30 days - not including job	0 = no 1 = yes	categorical: nominal (binary)
Fruits	Consumed fruit 1 or more times per day	0 = no 1 = yes	categorical: nominal (binary)

Veggies	Consumed vegetables 1 or more times per day	0 = no 1 = yes	categorical: nominal (binary)
HvyAlcoholConsump	Over 14 drinks per week for adult men and over 7 drinks per week for adult women	0 = no 1 = yes	categorical: nominal (binary)
AnyHealthcare	Have any kind of healthcare coverage, including health insurance, prepaid plans such as HMP, etc.	0 = no 1 = yes	categorical: nominal (binary)
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?	0 = no 1 = yes	categorical: nominal (binary)
GenHlth	How would you rate your general health?	1 = excellent 2 = very good 3 = good 4 = fair 5 = poor	categorical: ordinal (multivariate)
MenHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?	Scale of 1-30 days	discrete

PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?	Scale of 1-30 days	discrete
DiffWalk	Do you have serious difficulty walking or climbing stairs?	0 = no 1 = yes	categorical: nominal (binary)
Sex		0 = female 1 = male	categorical: nominal (binary)
Age	Age category	1 = 18-24 2 = 25-29 3 = 30-34 4 = 35-39 5 = 40-44 6 = 45-49 7 = 50-54 8 = 55-59 9 = 60-64 10 = 65-69 11 = 70-74 12 = 75-79 13 = 80 or older	categorical: ordinal (multivariate)

Education	Education level	1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)	categorical: ordinal (multivariate)
Income		1 = less than $10,000 2 = $10,000 to $15,000 3 = $15,000 to less than $20,000 4 = $20,000 to $25,000 5 = $25,000 to $35,000 6 = $35,000 to $50,000 7 = $50,000 to $75,000 8 = $75,000 or more	categorical: ordinal (multivariate)

```
# Access tree structure details
tree = dt_final.tree_

# Loop through each node
for i in range(tree.node_count):
    print(f"Node {i}:")
    # Check if it's a leaf node
    if tree.children_left[i] == -1 and tree.children_right[i] == -1:
        print(f"  Leaf node")
    else:
        feature_index = tree.feature[i]
        if feature_index != -2:  # -2 indicates no feature for leaf nodes
            feature_name = X_test_selected.columns[feature_index]
            threshold = tree.threshold[i]
            print(f"  Splitting on feature '{feature_name}' at threshold {threshold}")
        else:
            print("  No split (leaf node)")
```

✓ 0.0s

```
Node 0:
  Splitting on feature 'HighBP' at threshold 0.5
Node 1:
  Splitting on feature 'HvyAlcoholConsump' at threshold 0.5
Node 2:
  Splitting on feature 'CholCheck' at threshold 0.5
Node 3:
  Splitting on feature 'Sex' at threshold 0.5
Node 4:
  Splitting on feature 'AnyHealthcare' at threshold 0.5
Node 5:
  Splitting on feature 'Smoker' at threshold 0.5
Node 6:
  Splitting on feature 'DiffWalk' at threshold 0.5
Node 7:
  Splitting on feature 'NoDocbcCost' at threshold 0.5
Node 8:
  Leaf node
Node 9:
  Leaf node
Node 10:
  Splitting on feature 'NoDocbcCost' at threshold 0.5
Node 11:
  Leaf node
Node 12:
...
Node 853:
  Leaf node
Node 854:
  Leaf node
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Figure 1. Decision Tree Model Transparency and Interpretability

```python
import pandas as pd
import numpy as np

# Extract the coefficients and feature names
coefficients = lr_final.coef_  # Coefficients for each feature
intercept = lr_final.intercept_  # Intercept(s)
feature_names = X_test_selected.columns

# Organize the coefficients into a DataFrame for better readability
coeff_df = pd.DataFrame(coefficients.T, columns=lr_final.classes_, index=feature_names)
coeff_df.index.name = "Feature"
coeff_df.reset_index(inplace=True)

# Add intercepts if it's a binary classification (logistic regression only has one intercept)
if len(intercept) == 1:
    coeff_df["Intercept"] = intercept[0]
else:  # For multiclass, add intercepts for each class
    for i, class_label in enumerate(lr_final.classes_):
        coeff_df[f"Intercept (Class {class_label})"] = intercept[i]

# Sort features by the largest magnitude of coefficients
coeff_df["Max_Coefficient"] = coeff_df.iloc[:, 1:].abs().max(axis=1)  # Magnitude for sorting
coeff_df = coeff_df.sort_values(by="Max_Coefficient", ascending=False)

# Display the results
print("Logistic Regression Coefficients:")
print(coeff_df)
```

✓ 0.0s

```
Logistic Regression Coefficients:
              Feature         0         1         2  Intercept (Class 0)  \
9     HvyAlcoholConsump  1.131108 -0.614397 -0.516712             0.550532
0                   Sex  0.181391 -0.275220  0.093829             0.550532
1         AnyHealthcare  0.140984 -0.175116  0.034132             0.550532
2              DiffWalk -0.208306 -0.184804  0.393110             0.550532
4                HighBP -0.525241 -0.041256  0.566497             0.550532
3                Smoker  0.187697 -0.214654  0.026957             0.550532
5  HeartDiseaseorAttack  0.049129 -0.361343  0.312214             0.550532
6             CholCheck -0.595702  0.142326  0.453376             0.550532
7          NoDocbcCost  0.448873 -0.248617 -0.200257             0.550532
8                Stroke  0.453637 -0.589053  0.135416             0.550532

   Intercept (Class 1)  Intercept (Class 2)  Max_Coefficient
9             0.391609             -0.94214         1.131108
0             0.391609             -0.94214         0.942140
1             0.391609             -0.94214         0.942140
2             0.391609             -0.94214         0.942140
4             0.391609             -0.94214         0.942140
3             0.391609             -0.94214         0.942140
5             0.391609             -0.94214         0.942140
6             0.391609             -0.94214         0.942140
7             0.391609             -0.94214         0.942140
8             0.391609             -0.94214         0.942140
```

Figure 2. Logistic Regression Model Transparency and Interpretability

Table 4: Numerical Categories Transformed to Categorical

Name	Description	Value	Data Type
BMI	Calculated Body Mass Index	Underweight = 0 to 18.4 Healthy Weight = 18.5 to 24.9 Overweight = 25 to 29.9 Class 1 Obesity = 30 to 34.9 Class 2 Obesity = 35 to 39.9 Class 3 Obesity = 40 to 100	categorical
MenHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?	0 days 1 to 5 days 6-10 days 11-15 days 16-20 days 21-25 days 26-30 days	categorical
PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?	0 days 1 to 5 days 6-10 days 11-15 days 16-20 days 21-25 days 26-30 days	categorical

Table 5: Population Subsets and their index

Subset Index	Age range, sex, BMI category
Subset 1	subset_18_29_male_underweight,
Subset 2	subset_18_29_female_underweigh
Subset 3	subset_18_29_male_healthy_weight
Subset 4	subset_18_29_female_healthy_weight,
Subset 5	subset_18_29_male_overweight,
Subset 6	subset_18_29_female_overweight
Subset 7	subset_18_29_male_obesity_1
Subset 8	subset_18_29_female_obesity_1,
Subset 9	subset_18_29_male_obesity_2,
Subset 10	subset_18_29_female_obesity_2,
Subset 11	subset_18_29_male_obesity_3,
Subset 12	subset_18_29_female_obesity_3,
Subset 13	subset_30_49_male_underweight,
Subset 14	subset_30_49_female_underweight,
Subset 15	subset_30_49_male_healthy_weight,

Subset 16	subset_30_49_female_healthy_weight,
Subset 17	subset_30_49_male_overweight,
Subset 18	subset_30_49_female_overweight,
Subset 19	subset_30_49_male_obesity_1,
Subset 20	subset_30_49_female_obesity_1
Subset 21	subset_30_49_male_obesity_2,
Subset 22	subset_30_49_female_obesity_2,
Subset 23	subset_30_49_male_obesity_3,
Subset 24	subset_30_49_female_obesity_3,
Subset 25	subset_50_64_male_underweight,
Subset 26	subset_50_64_female_underweight
Subset 27	subset_50_64_male_healthy_weight,
Subset 28	subset_50_64_female_healthy_weight
Subset 29	subset_50_64_male_overweight,
Subset 30	subset_50_64_female_overweight
Subset 31	subset_50_64_male_obesity_1,

Subset 32	subset_50_64_female_obesity_1
Subset 33	subset_50_64_male_obesity_2,
Subset 34	subset_50_64_female_obesity_2
Subset 35	subset_50_64_male_obesity_3,
Subset 36	subset_50_64_female_obesity_3
Subset 37	subset_65_plus_male_underweight
Subset 38	subset_65_plus_female_underweight,
Subset 39	subset_65_plus_male_healthy_weight,
Subset 40	subset_65_plus_female_healthy_weight,
Subset 41	subset_65_plus_male_overweight,
Subset 42	subset_65_plus_female_overweight
Subset 43	subset_65_plus_male_obesity_1,
Subset 44	subset_65_plus_female_obesity_1,
Subset 45	subset_65_plus_male_obesity_2,
Subset 46	subset_65_plus_female_obesity_2,
Subset 47	subset_65_plus_male_obesity_3,

Subset 48	subset_65_plus_female_obesity_3