Predicting Diabetes and Identifying Important Health Indicators

CIND820 Literature Review, Data Description, and Approach

Stephanie Boissonneault
500616408
Supervisor: Tamer Abdou
October 27th, 2024

Ryerson University

Table of Contents

Abstract **2**
Introduction **3**
Literature Review **4**
Literature Search Strategy 4
Literature Search Results 5
Diabetes as a Common Disease for the Application of Predictive Modeling 7
Commonly Applied Machine Learning Algorithms 7
Commonly Selected Features for Diabetes Prediction 9
Commonly Reported Evaluation Metrics 10
Commonly Reported Challenges 10
Other Recommendations and Future Opportunities 11
Literature Search Summary 12
Proposed Research Questions **12**
Proposed Research Methodology **13**
Data Description 14
Data Cleaning and Preprocessing 14
Descriptive Statistics 15
Approach for Answering Research Questions **20**
Question 1 20
Question 2: 21
Question 3 24
Project Implications **25**
Link to the Working Dataset **26**
References **27**
Appendix A **30**
Appendix B **42**

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 also crucial for clinicians to interpret and validate the model's diagnosis and disease detection. Based on the findings of the literature review, this project proposes conducting a study investigating feature selection of diabetes health indicators and building various machine learning models for disease detection. Several validation metrics will also be reported during model comparison for future consideration and validation in clinical settings.

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 which

are 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.). 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.). Early diagnosis and disease management are therefore crucial for mediating further health complications and improving the health of Canadians. This research project proposes an investigation into the application of machine learning techniques for helping with the identification of diabetes and its risk factors for early disease detection and intervention using the publicly available Centers for Disease Control (CDC) 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 the PubMed database as it is 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 during 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. Studies focussing solemnly on deep learning techniques were excluded to isolate the search to machine learning techniques commonly used on smaller datasets.

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.

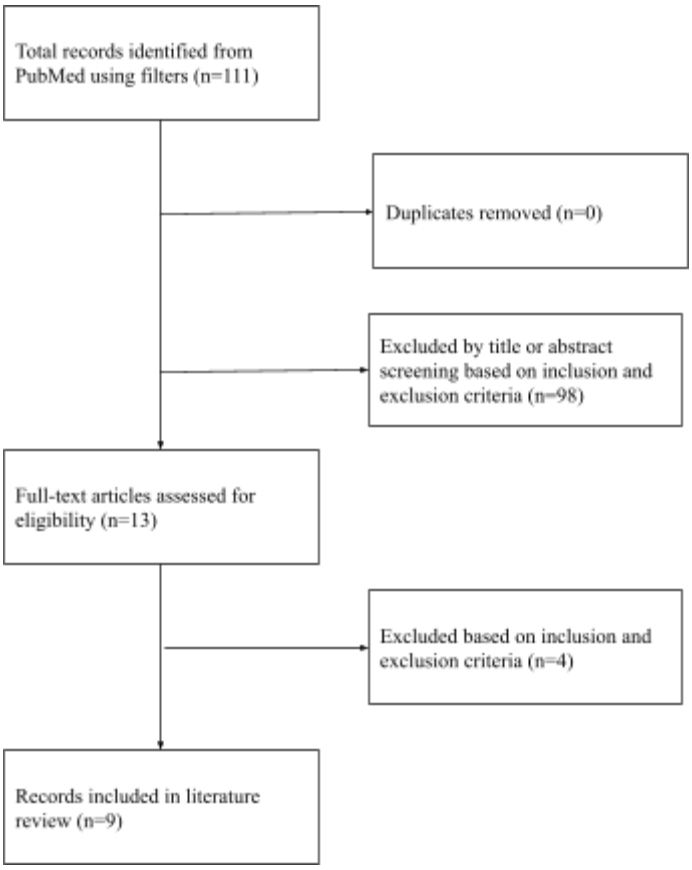

Figure 1. Search Strategy Diagram

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, decision tree models were commonly described as the most common and best performing models by multiple articles. 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) conducted a survey following 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 f 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.

Proposed Research Questions

The lack of consensus across studies for the models and features identified as the best predictors of diseases suggests a need for further exploration into 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 will also investigate “What machine learning models best predict diabetes mellitus and what health indicators are considered most important for disease prediction?’’ to recommend predictive models that could improve screening and diagnosis of the disease for early intervention and disease management. Lastly, this project will also investigate “What patterns can be uncovered in subpopulations based on age, sex, and income?” to help provide further insights for designing programs that meet individual needs and target most at-risk populations based on differences in the incidence and experience of disease across different demographics and socioeconomic factors. 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.

Proposed Research Methodology

To help answer the research questions, this project proposes applying machine learning techniques to a consolidated version of the 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 Jupyter Notebook, Visual Studio Code, Python programming language, and Python libraries will be 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 having no diabetes or only diabetes during pregnancy, having pre-diabetes, or 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 also generated using Python's Panda's library and was uploaded to 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. There were no missing values identified in the dataset. Although zeros were recorded for MentHlth and PhysHlth while these were ment to be measured on a scale of 1 - 30, these zeros are considered valid rather than missing data as they represent that there were no days in the past 30 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 and could be meaningful in evaluating the frequency of itemsets during the evaluation of association rules. Duplicates will therefore not be removed during initial data cleaning but will be aggregated before any application of machine

learning for predictive models to avoid biased results. Data types were standardized and transformed where appropriate. Cleaning and preprocessing steps are further described in the Jupiter notebook for the working dataset uploaded to the GitHub repository.

Descriptive Statistics

First, the frequency distribution of the target variable was visualized and revealed a large class imbalance (see Figure 2). This target variable will therefore need to be balanced at a later stage since many predictive machine learning models can be biased towards the majority class. 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.

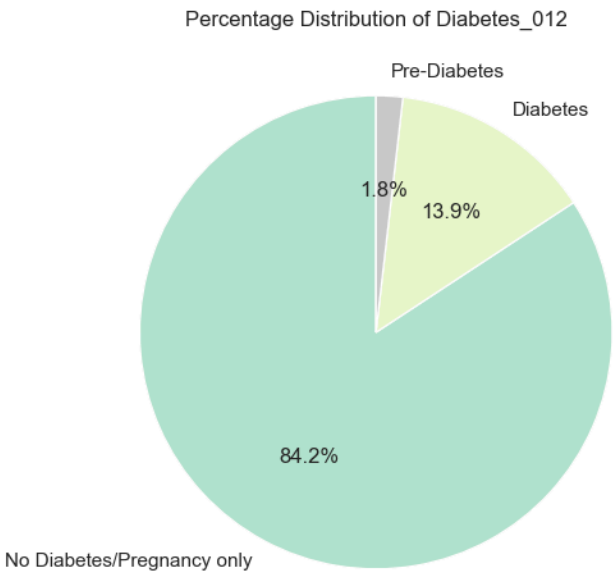

Figure 2. Frequency distribution of “Diabetes_012” target variable

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 values that may be relevant to the analysis. 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 remain between the 1 - 30 day range. It is important to note that some machine learning algorithms may however be sensitive to outliers. If this seems to be the case during analysis, the outliers could be capped at the 95th percentile.

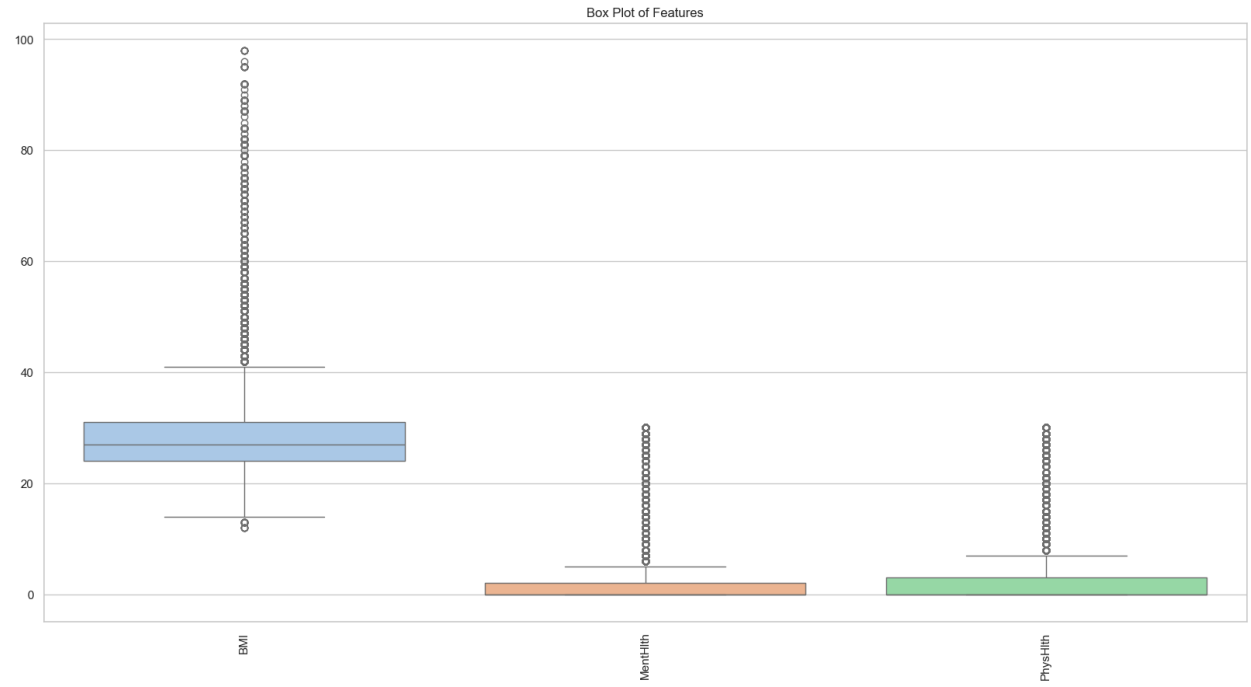

Figure 3. Boxplot of discrete and continuous data

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

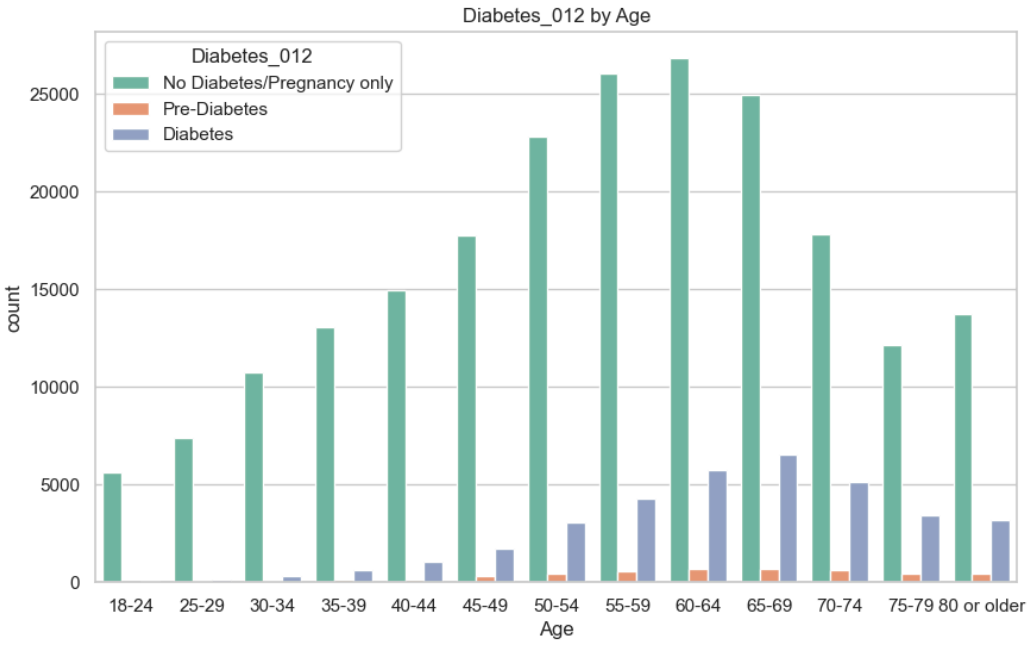

Figure 4. Frequency of diabetes diagnosis across age ranges

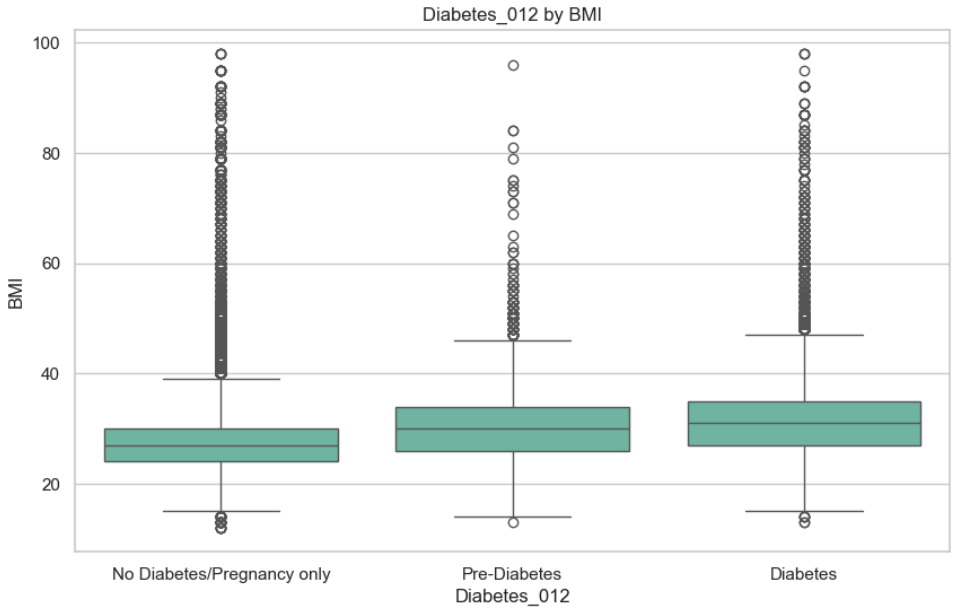

Figure 5. BMI across diabetes diagnosis

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).

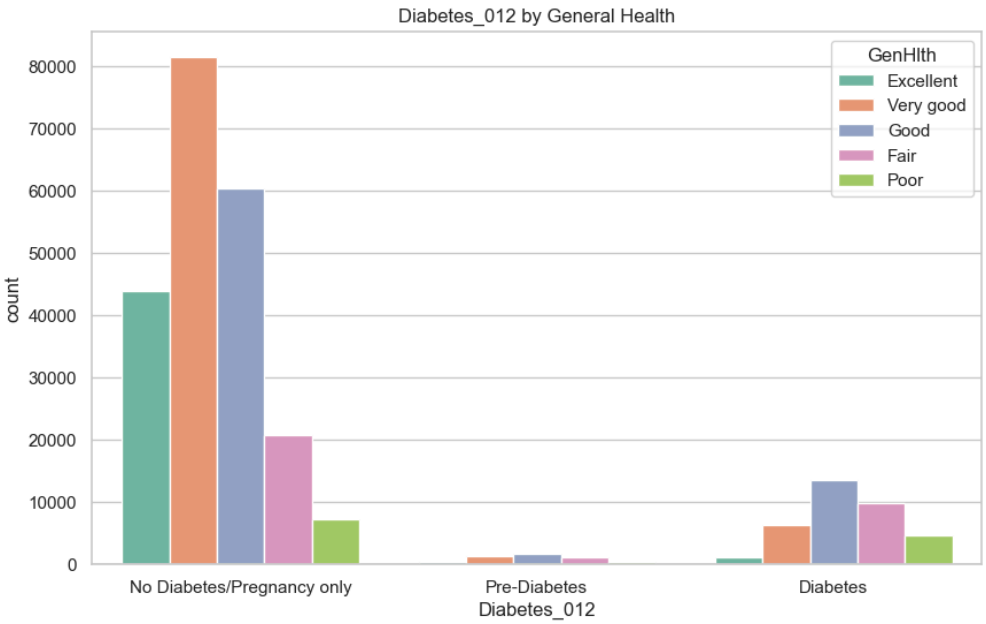

Figure 6. Diabetes by general health rating

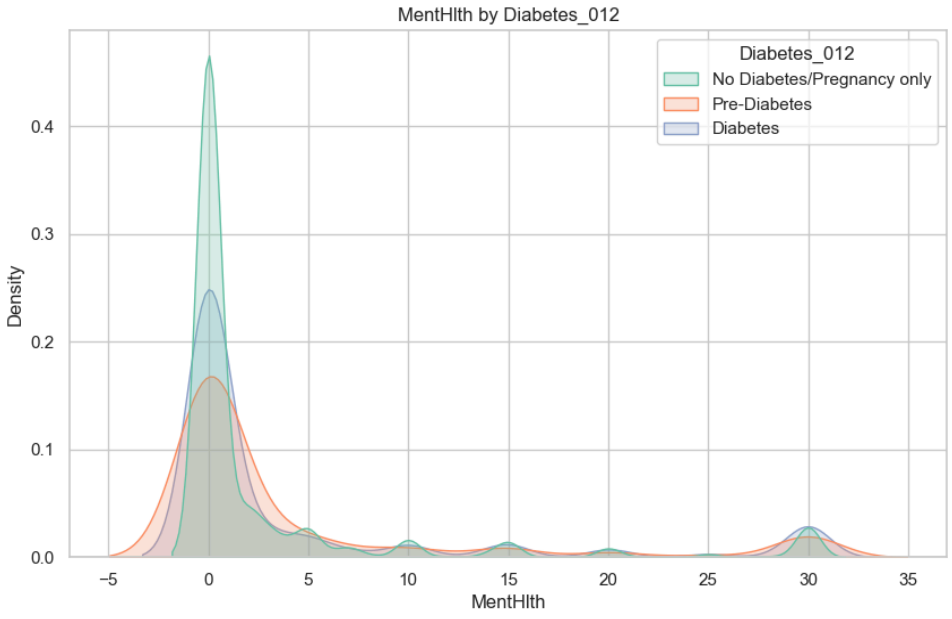

Figure 7. Days of reported poor mental health by diabetes diagnosis

Approach for Answering Research Questions

Question 1

To help identify “What health conditions and lifestyle factors commonly occur together in individuals with different diabetes diagnoses?”, this project proposes applying the apriori algorithm for knowledge induction as outlined in Figure 2. The original dataset will first be divided into three subsets based on diabetes diagnosis (no diabetes/gestational diabetes only, pre-diabetes, and diabetes). Relevant features for each subset will be extracted to create two transaction datasets; one for health conditions (HighBP, HighChol, Stroke, and HeartDiseaseAttack), and one for lifestyle factors (Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump). Each row of the transaction dataset will be given a transaction ID and list a combination of items, either health conditions or lifestyle factors, followed by the diabetes diagnosis as an outcome. The apriori algorithm will be applied separately to each transactional dataset for each subset of data to identify frequent item sets and generate sets of association rules based on a support threshold of 5-10% and a confidence threshold of 70%. Graph-based visualizations will be created to interpret the results. Results for commonly occurring health conditions and lifestyle factors will be interpreted for each diagnosis to extract meaningful insights and make recommendations for health indicators that could be assessed together in practice.

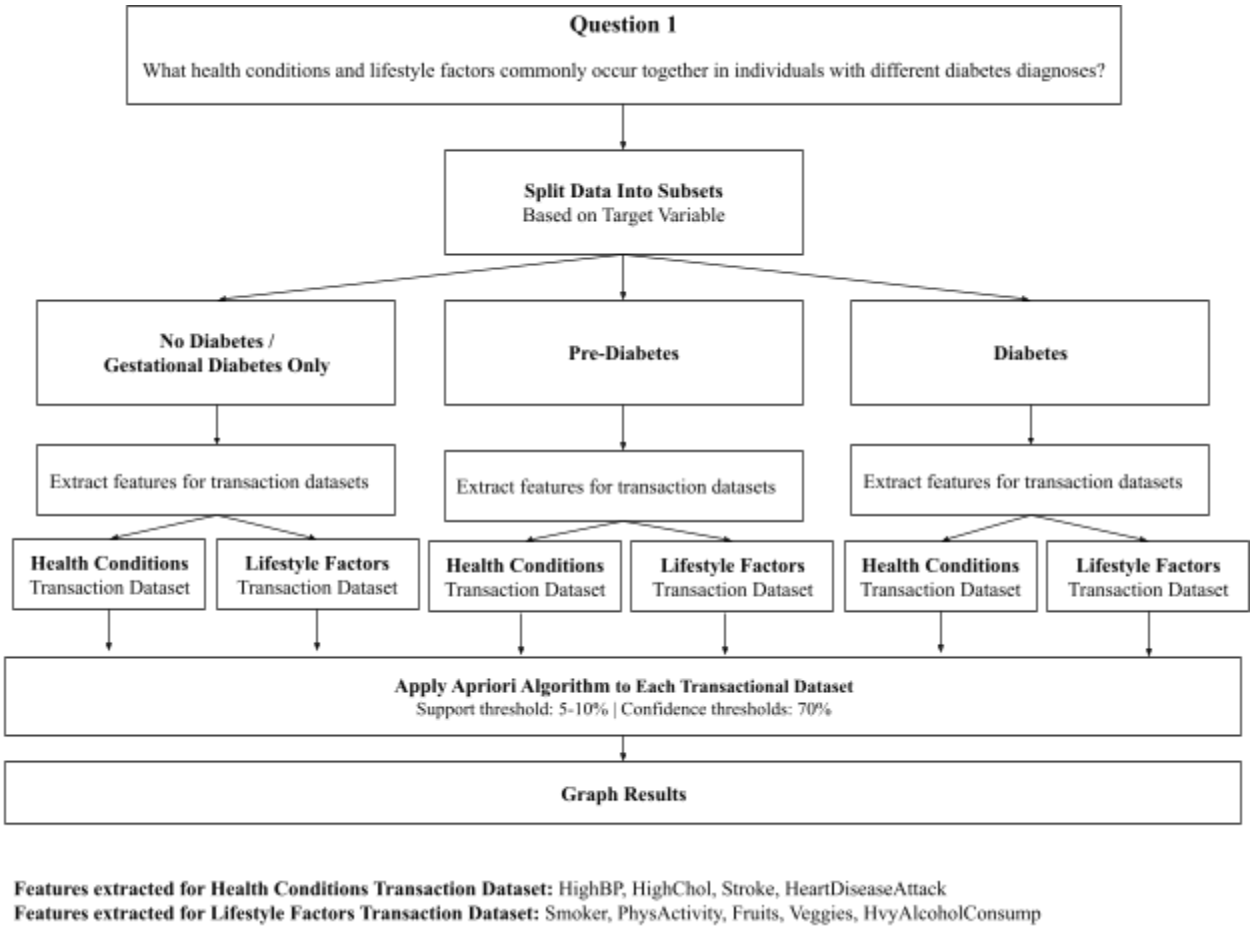

Figure 2: Apriori Algorithm Approach for Question 1

Question 2:

This project will apply and evaluate different machine learning models to answer the question “What machine learning models best predict diabetes mellitus and what health indicators are considered most important for disease prediction?”. First, duplicate values in the dataset will be aggregated to avoid biased results during the application of machine learning techniques. Oversampling using Synthetic Minority Over-Sampling Technique (SMOTE) will be

applied to the initial dataset to balance the data and improve machine learning performance. Feature selection will be conducted on the balanced dataset as advised by findings from the literature review and as outlined in Figure 3. The balanced dataset will be standardized and then split into a training set (70%) and a test set (30%). K-fold cross-validation will be applied to the training set. The feature selection techniques chi-square, decision tree, and multinomial logistic regression will be trained multiple times on the training sets to help tune hyperparameters and assess model performance. These feature selection techniques have been selected because they are interpretable, increasing transparency for clinical use and validation by practitioners in healthcare settings. Metrics from each fold for each feature selection technique will be aggregated to determine the best hyperparameters. The final model will be trained on the entire training set and evaluated to extract insights on feature importance. A k-fold cross-validation will be conducted to evaluate the generalizability of the model. Reported evaluation metrics for each model will include accuracy, sensitivity, specificity, precision, F1-Score, and F2-Score to enable benchmarking between different models from different studies. The features with the best model interpretability and performance will be selected to train multiple predictive machine-learning models.

A dataset with the selected features from the previous analysis will be used to train multiple machine-learning models to predict a diabetes diagnosis as outlined in Figure 4. First, the data will be split into a training set (70%) and a test set (30%). K-fold cross-validation will be applied to the training set. The machine learning algorithms decision tree, random forest, multinomial logistic regression, Naive Bayes, and k-nearest neighbors will each be applied to each k-fold of the training set to train the models. Hyperparameters will be tuned before training the final model on the entire training set. The models will then be applied to the test set to

evaluate their predictive ability. Several evaluation metrics will be reported including accuracy, sensitivity, specificity, precision, F1-Score, and F2-score to enable benchmarking between models from different studies. Combining K-folds and train-test split will increase the reliability and generalizability of the models, ensuring that bias is minimized.

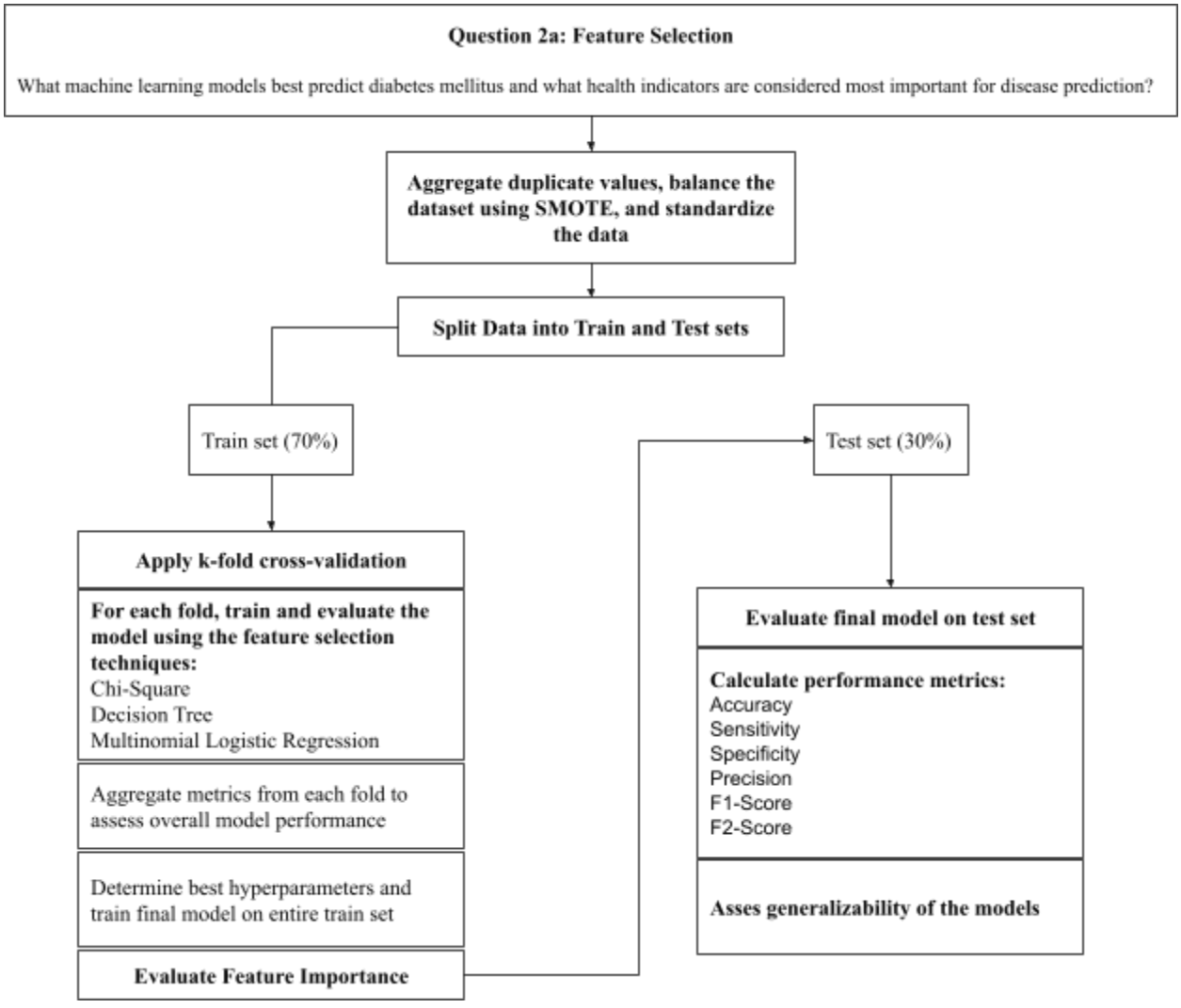

Figure 3. Feature Selection Approach for Question 2a

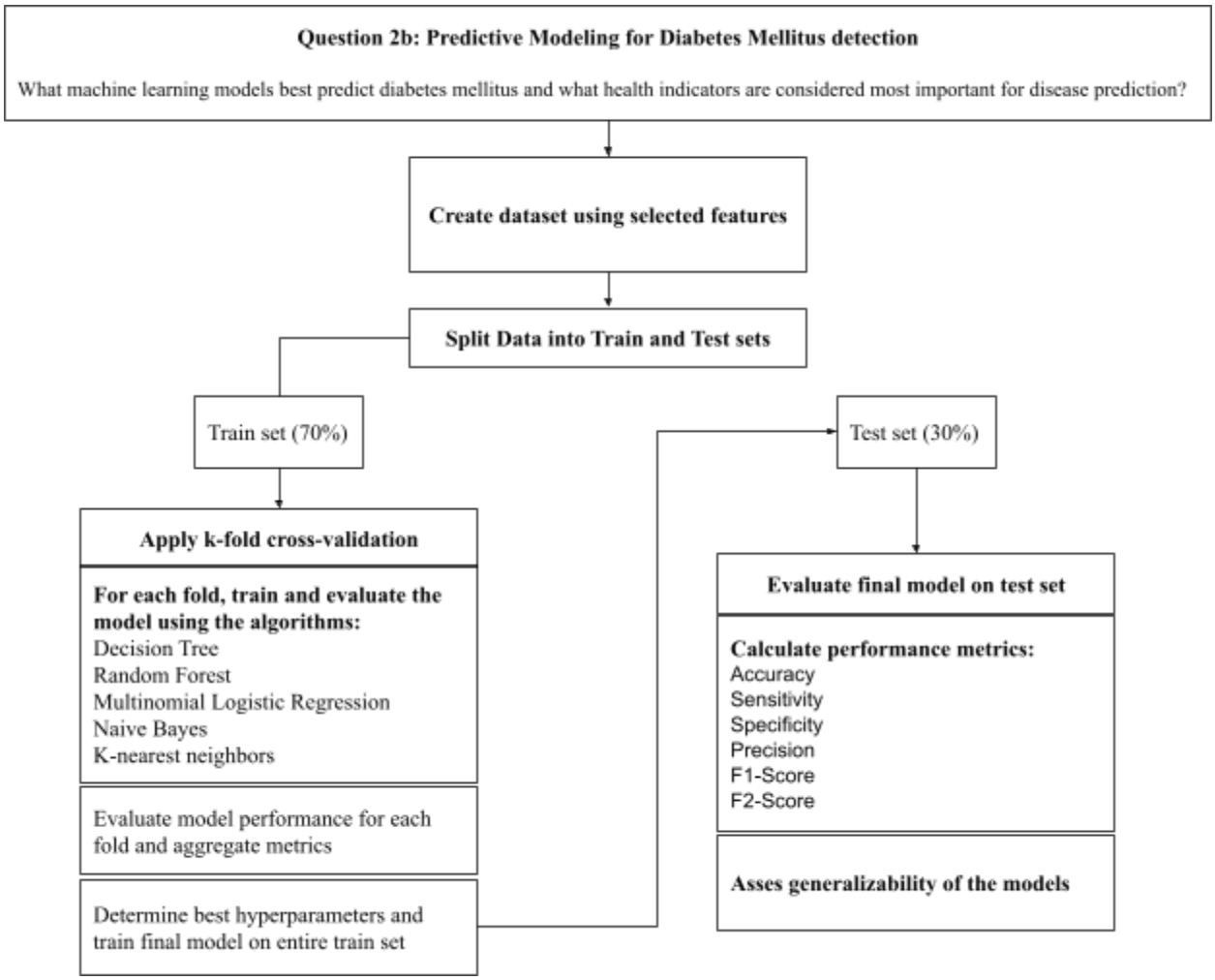

Figure 4. Building Predictive Models for Question 2b

Question 3

The literature review revealed age and obesity as common predictors of diabetes. Principle component analysis (PCA) will be used to learn “What patterns can be uncovered in subpopulations based on age, sex, and BMI” as outlined in Figure 5. First, a new column will be

created to categorize BMI data following health guidelines. The dataset will then be filtered based on the new categories. PCA will be applied to each subpopulation for analysis. Results will be graphed to create visualizations and extract insights.

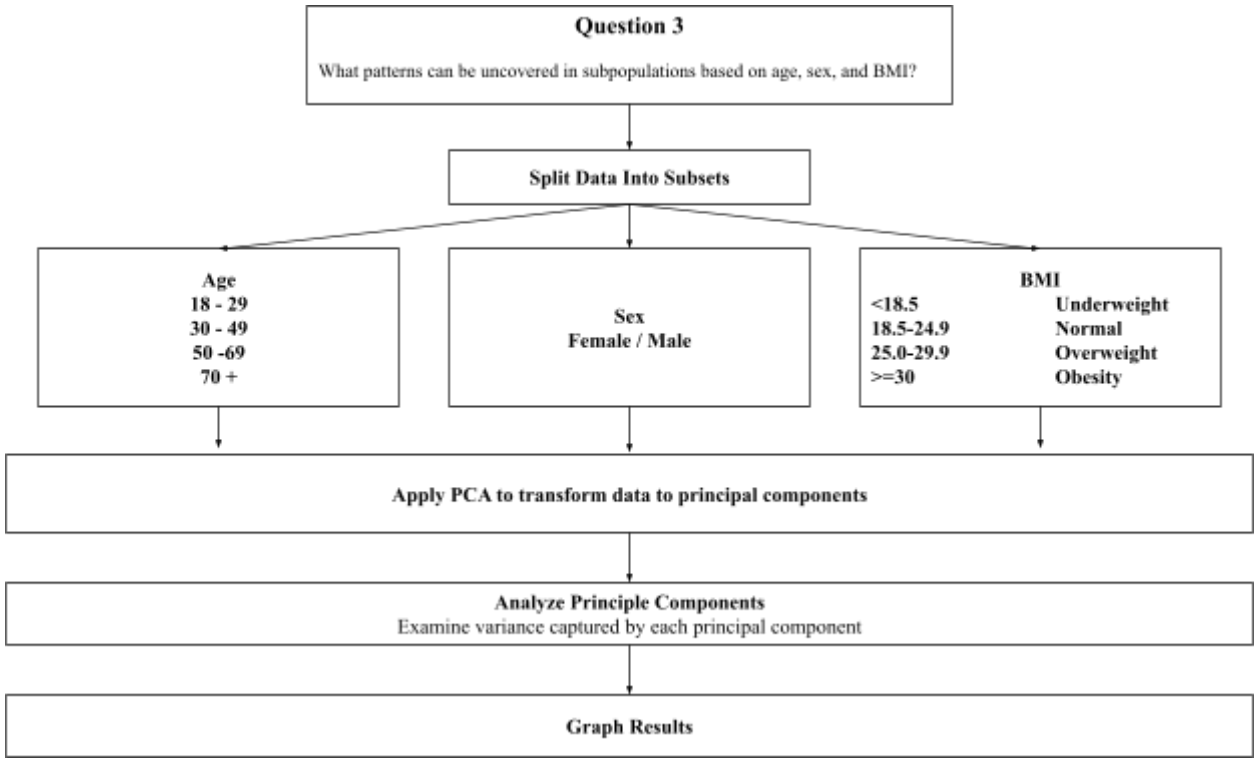

Figure 5. Applying PCA for Question 3

Project Implications

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.

Link to the Working Dataset

---

https://github.com/stephbois/Big_Data_Analytics_Project/tree/main/project_files/working_dataset

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. https://www.cdc.gov/brfss/annual_data/2014/pdf/Overview_2014.pdf

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.). 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

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,	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

					extreme, light, and adaptive boosting, decision tree, Naive Bayes, k-nearest neighbors, and LASSO, and neural networks		
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-ba	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	Future research should look to algorithms that are fair, transparent, and validated before clinical implementation

					sed surveys, hospital records, and digitally traced data	Risk factors identified using clustering algorithms like principle component analysis, logistic regression 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	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	Insufficient data and small sample sizes are common challenges throughout selected articles

					diabetes detection include preprocessing, feature extraction, machine learning, and classification. Precision, recall, accuracy, and F score are commonly used for model evaluation.	Computationally effective feature selection is needed to increase accuracy. Models need to be validated on multiple sites 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	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	Difficulty with comparisons between models due to heterogeneity in the population and selected sample Also no concensus on evaluation metrics for reporting

					SVM, RF, GBT and DNN were the most popular machine learning techniques Decision tree and random forest wer the top performing models Most studies used metrics from confusion matrix to report on performance	Machine learning models worked best on well-structured balanced 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	12 studies comparing ML classification with actual diabetes incidence	Systematic review of longitudinal studies	Extracted data and used QUADUS-2 to evaluate study quality	Included classifiers were decision tree, neural network, k-nearest	Future research should compare the ability to predict type 2	Limited to comparing studies from high-income countries and limited to

	diabetes mellitus	from 1950-2020 found in Medline and Embase		Data was synthesized for each study using a confusion matrix to determin pooled sensitivity, specificity, positive likelihood ratio, and negative likely ratio.	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 feature	diabetes mellitus among ML algorithms, previously established risk models	comparing studies that specified consistency across evaluation metrics

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	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	frequently used model 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	The best predictive models included measures of glycemia Standard lipids or NMR-based lipoprotein size and concentration measures did not augment the predictive	Machine learning algorithms provided no meaningful improvement for discrimination compared with logistic regression, which suggests a lack of influential latent	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

				proportional hazards model, gradient boosting, random forest, support vector machines with linear kernel (SVM-L), polynomial kernel (SVM-P) and radial kernel (SVM-R), and artificial neural network (ANN)	utility of models incorporating glycemia	interactions among the analytes assessed in this study	Data from blood samples may be skewed due to long storage times
Zanelli et al., 2022	Provide detailed overview of published	78 studies across 4 databases focused on	Systematic Review	Several previously applied machine	Both traditional and machine learning	Machine learning techniques are promising	Lack of standardizatio n of the results.

	methods used for detecting and managing diabetes using PPG and ECG signals	glucose estimation and diabetes detection		learning approaches to distinguish between diabetic and healthy subjects, notably random forest, logistic regression, and decision trees with high-performi ng specificity, sensitivity, and accuracy	approaches require the feature extraction step to be done no feature extraction algorithm can work if the input signal is corrupted.	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 data to minimize bias, increase parameters	Difficult to compare studies due to heterogeneity

						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	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	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 consideration	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

				outcome measures of discrimination and calibration Meta-DiSc software used to measure pooled estimates of AUROC, sensitivity, specificity, PLR, NLR, and DOR	fasting blood glucose	s include feature selection based on clinical need rather than accuracy (features 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	Scale of 1-30 days	discrete

	your mental health not good?		
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	categorical: ordinal (multivariate)

		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)	
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)