Name: Matthew Vu
Project Title: Experimentation of Machine Learning Techniques to Create a Diabetes Risk Prediction Application

Research Questions:

1. What behaviors and habits are associated with diabetes risk?
2. What machine learning algorithms (ML) out of the following is most effective at predicting diabetes risk: logistic regression, naïve bayes, support vector machines, k-nearest neighbors, or decisions trees
3. Will clustering the data using Gower's distance and the Partition Around Medoids (PAM) clustering algorithm improve model accuracy?

Motivation and Literature Review:

The motivation for the research and creation of this questionnaire application stems from the fact that the early detection and prediction of diabetes remains a relevant topic. Particularly, there exist several reasons for continued research. First, there exists high prevalence of diabetes, and many people do not know that they have it. In fact, according to the CDC's National Diabetes Statistics Report, over 37 million Americans have diabetes; however, about 1 in 5 of them do not know that they have it. In particular, the last part is concerning because the CDC would say that prolonged high blood sugar levels can lead to cardiovascular and kidney disease (CDC, 2022). Second, early intervention for individuals at risk of diabetes can help reduce the chances of becoming diabetic. For instance, a study conducted by the Diabetes Prevention Program Research Group discovered that subjects who were administered into lifestyle early intervention programs to address their risk of diabetes contracted diabetes 58 percent less than subjects who did not participate (Diabetes Prevention Program Research Group, 2002). Third, living with diabetes comes with financial burden. In fact, diabetic people spend approximately 2.3 times more on medical expenditures than people without diabetes (Petersen, 2018). With that said, creating my questionnaire application will give individuals a cost-effective resource to easily know their diabetes risk. This will allow them to take appropriate action, such as lifestyle changes or getting a full diabetes screening.

In addition to wanting to create an application to help individuals take action, I would like to fill gaps in past research. The first piece of research is from a book chapter titled "Likelihood Prediction of Diabetes at Early State Using Data Mining Techniques", which outlines experimenting with data mining techniques to understand if behavioral risk factors, such as polyuria and polydipsia, can be used to predict diabetes (Islam et al., 2019). Now, improvements I would make are experimenting with support vector machines and data describing a person's habits, such as diet, sleep, smoking, etc. In addition, the second piece of research is from an article titled "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques", which describes experimenting with ML algorithms to predict the risk of type 2 diabetes (Xie et al., 2019). They used the CDC's 2014 Behavioral Risk Factor Surveillance System Health Survey Dataset consisting of predictors describing behavioral and habitual risk factors associated with diabetes, such as smoking and lack of exercise. Now, an improvement I would make is to investigate predictors describing diet, high blood pressure, and difficulty walking. Finally, some aspects I would add to both pieces of research are more EDA (to prove predictor significance and check for multicollinearity), deploying their research with an application, and experiment with clustering to improve accuracy and k-nearest neighbors.

Why these research questions are important to me:

I would like to contribute to the conversation of using data science concepts within the medical field. Additionally, I would like to deploy concepts I learned in previous courses to create a practical application to help individuals make data-driven decisions. Moreover, by sharing this project on GitHub, I hope to get others involved in medical data science.

What do you hope to learn beyond more insight on the question:

I would like to review EDA concepts (like inference and association metrics), data preparation, supervised ML algorithms, and R programming. Also, I would like to learn more about Gower's distance and Partitioning Around Medoids (PAM) clustering.

Data and Metadata:

The dataset used to train, evaluate, and test the ML algorithms is from the CDC and titled "Behavioral Risk Factor Surveillance System: Public Health Surveys of 400k people 2015". It contains 441, 456 survey responses to the CDC's 2015 Behavioral Risk Factor Surveillance System Health Survey. Also, the 330 variables are the answers to these survey questions. Moreover, it comes in the form of a 516 MB .csv file with 441, 457 rows and 330 columns on Kaggle. I am using this dataset because it contains several useful behavioral and habitual risk factors associated with diabetes, such as diet, high blood pressure, and difficulty walking. Also, there are plenty of variables to choose from in case I need more variables. However, as an initial starting point to make the dataset more manageable, I selected predictors that are known to be associated with diabetes, such as age, difficulty walking, high blood pressure, BMI, etc. and cleaned the dataset similar to another dataset by Alex Teboul.

Methods of Analysis:

First, the data will be split into a 50-25-25 training, validation, and test set. Note, we can do this because the cleaned dataset still has over 200,000 observations. The training set will be used to train the algorithms; the validation set will be used assess the model's accuracy, choose which ML algorithm is the best, and whether to cluster or not; and the test set will only be used to finally calculate the generalization error. Second, logistic regression, naïve bayes, support vector machines, k-nearest neighbor, and decisions trees will be trained without PAM clustering and evaluated. Third, these same algorithms will be trained with PAM clustering and evaluated. Note, we use PAM clustering to accommodate the mix of numerical and categorical variables. Fifth, we pick and use the best performing combination of algorithm and clustering (or no clustering) in the application. Sixth, we use the test set to calculate the generalization error.

Software:

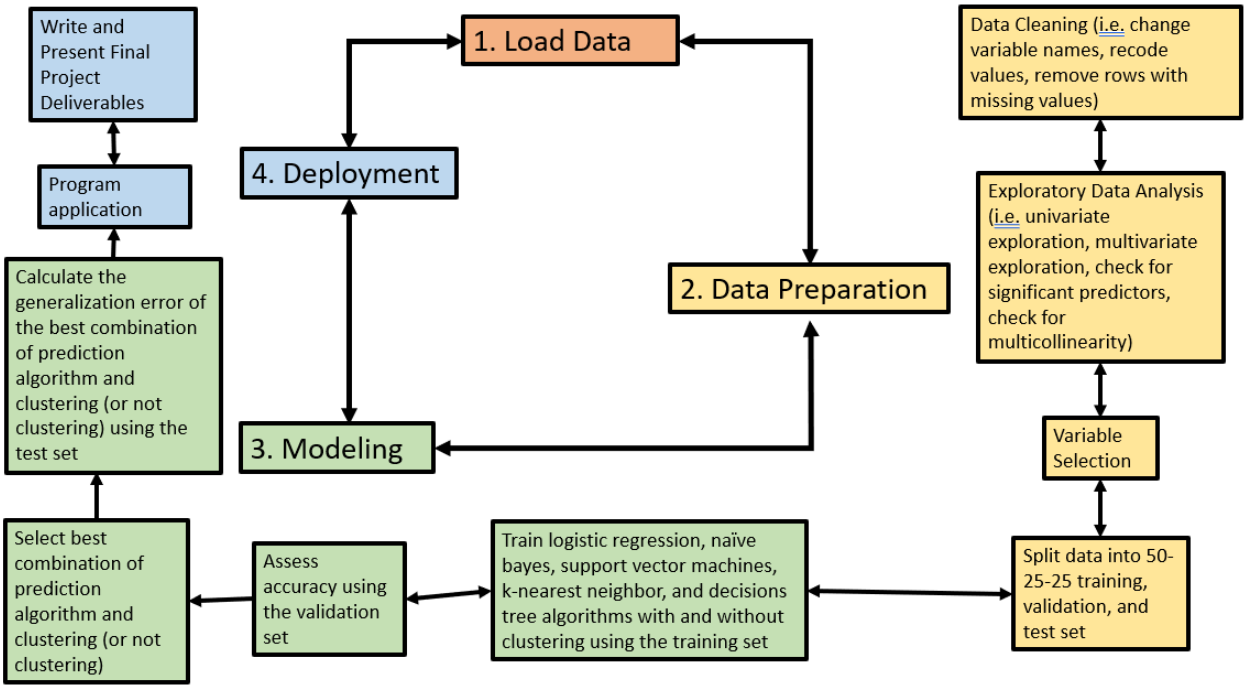
The software and packages that will be used for analysis and application creation are R, RStudio, Tidyverse, ISLR2, Daisy, and ShinyR.

Progress to Date and Tentative Deadlines:

As of March 1st, I have done initial data cleaning, variable selection, and EDA, such as plotting to explore relationships between diabetes and some predictors. Also, I created Cramer's V matrices to check for multicollinearity. Now, as for tentative deadlines, by Week 7, I should have useful predictors selected, data cleaned, visualizations made, and logistic regression trained. Week 8-9, I should experiment with the rest of the ML algorithms, clustering, and non-

clustering. Week 10-11, I should create the application and get started on the final deliverables. Week 12-15, I should finish the final deliverables.

Workflow Diagram:



References:

CDC. (2018). *Behavioral Risk Factor Surveillance System* [Data set]. *kaggle.* https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv

CDC. (2022, June 29). *National Diabetes Statistics Report*. The Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/data/statistics-report/index.html

Islam, M.M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *In Computer Vision and Machine Intelligence in Medical Image Analysis* (pp. 113–125). Springer Singapore. https://doi.org/10.1007/978-981-13-8798-2_12

Knowler, W., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., & Nathan, D. M. (2002). Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. *The New England Journal of Medicine*, *346*(6), 393–403. https://doi.org/10.1056/NEJMoa012512

Petersen, M. (2018). Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care*, *41*(5), 917–928. https://doi.org/10.2337/dci18-0007

Teboul, Alex (2019). *Diabetes Health Indicators Dataset* [Data set]. *kaggle.* https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease, 16*. https://doi.org/10.5888/pcd16.190109