West Texas A&M university
Paul and Virginia Engler College of Business

In partial fulfillment of the requirements in:
CIDM 5350 Business Intelligence and Decision Support Systems

Diabetes Prediction from Behavioral Risk Factors using Azure Machine Learning Studio and Power BI

Submitted by:

Vernice Tanquerido

Submitted to:

Dr. Jeffry Babb

Project Overview

Diabetes is a chronic illness that inhibits the body's capacity to process blood glucose or blood sugar. In the United States, diabetes is the eighth leading cause of death and is the major cause of kidney failure, lower limb amputations, and adult blindness (U.S. Centers for Disease Control and Prevention, 2024). This condition has deep impacts on the quality of life of those affected and poses significant challenges to the healthcare system.

In 2021, about 1 out of 11 people are diagnosed with diabetes in the United States (National Institute of Diabetes and Digestive and Kidney Diseases, 2024). By 2045, the International Diabetes Foundation projects approximately 783 million will be living with diabetes worldwide, an increase of 46% compared to 2021. This increase raises alarms and highlights the need for effective prevention and management strategies.

Over 90% of people diagnosed with diabetes have type 2 diabetes, which is influenced by a variety of demographic, socioeconomic, environmental, and genetic factors (International Diabetes Foundation, n.d.) Unlike type 1 diabetes, which is largely hereditary and unpreventable, type 2 diabetes can be significantly mitigated through lifestyle changes. Maintaining a healthy body weight, engaging in regular exercise, following a balanced diet, and avoiding smoking can dramatically lower the chances of developing this condition (World Health Organization, 2023).

The Behavioral Risk Factor Surveillance System (BRFSS) is a project of the Center for Disease Control and Prevention (CDC) in collaboration with the different states across the United States. THE BRFSS is a series of monthly telephone interviews to collect data on health risk behaviors, chronic diseases and conditions, access to health care, and health services (Center for Disease Control and Prevention, 2023). The survey collects over 350,00 responses each year. This data is crucial for understanding the factors contributing to the leading causes of death and disability in the United States.

This data collected through the BRFSS can be leveraged to predict the likelihood of diabetes based on individuals' lifestyle choices and risk behaviors. By analyzing this data, patterns, and correlations contributing to diabetes can be identified, which can raise awareness and encourage public health initiatives.

Ultimately, the goal of this study is to raise awareness about the risk factors associated with diabetes and to promote lifestyle changes to mitigate these risks and reduce the chances of developing the disease. Public health campaigns, informed by data-driven insights, can educate individuals about the importance of maintaining a healthy lifestyle. Additionally, healthcare providers can use predictive models to identify high-risk individuals and offer targeted interventions, such as personalized diet and exercise programs, smoking intervention, and regular monitoring of blood sugar levels.

This project specifically aims to analyze various factors listed in the survey and assess their accuracy in predicting diabetes using machine learning tools and developing models that will unravel relationships between the risk factors and eventually predict an individual's likelihood of developing diabetes. By taking advantage of the comprehensive data collected through the BRFSS

and applying machine learning techniques, this project aims to contribute to the ongoing efforts to combat the growing diabetes epidemic.

PHASE ONE: PROJECT PLANNING

Project Idea

The goal of this project is to identify patterns and analyze risk factors and behaviors using data from the BRFSS, ultimately predicting the likelihood of an individual developing diabetes. To achieve this, advanced AI techniques will be leveraged to deliver insights efficiently and interactively. The project will conduct two types of analyses on the dataset: diagnostic analysis to understand the current state and risk factors, and predictive analysis to forecast the probability of diabetes based on individual profiles.

The Key Influencer tool in Power BI will be utilized for the diagnostic analysis. Figure 1 gives an overview of the use case architecture.

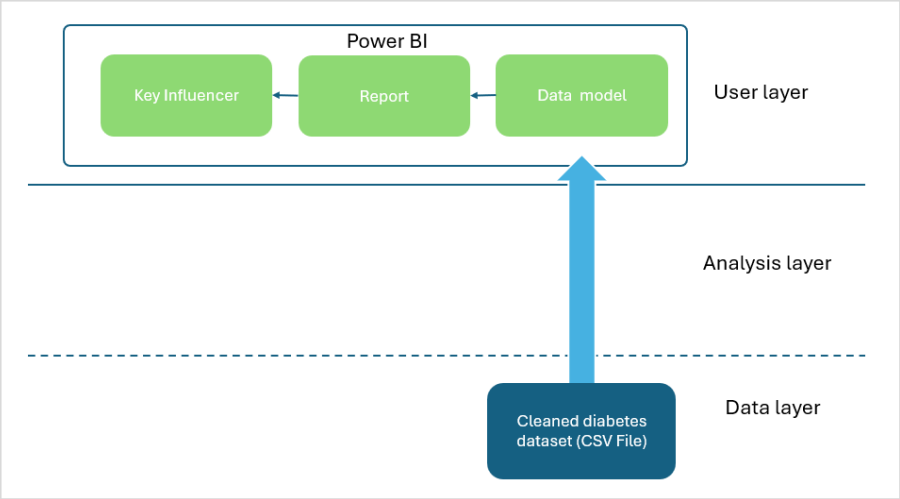

Figure 1. Key Influencer use case architecture

This analysis will be helpful in understanding the different factors that influence the likelihood of diabetes. The key influencers visual will list the top contributors to diabetes while the top segments visual will give an aggregated group that significantly impacts the target variable.

For the classification task, AutoML from Azure Machine Learning Studio will be utilized. The cleaned and balanced dataset consisting of 70,693 instances with 21 features will be the input to the AutoML Job. The best model will be selected based on the weighted precision scores of the models generated.

The selected model will be deployed as a web service. The model's endpoint URL will be integrated into an R script, making it accessible and usable within Power BI.

Figure 2 gives an overview of the use case architecture.

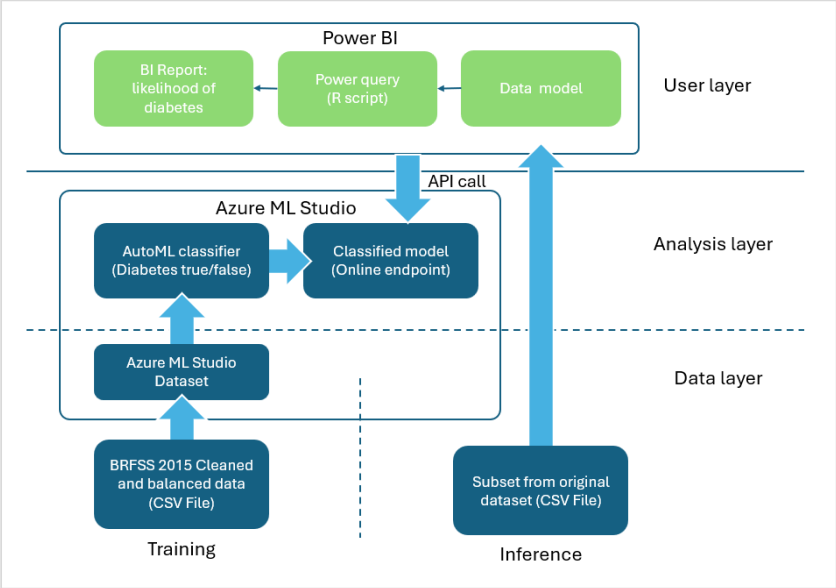

Figure 2. Classification use case architecture

Data Sources

This project used the cleaned dataset created by Alex Teboul, available on Kaggle, which is based on the 2015 BRFSS. The balanced dataset contains 70,693 instances with 21 risk factors that will be our features and a binary target variable indicating whether an individual is diagnosed with diabetes. The dataset and its source can be accessed using the links below:

- Cleaned dataset
- Behavioral Risk Factor Surveillance System

4V Model Analysis of the Data Source

Table 1 summarizes the assessment of the dataset using the 4V model.

Table 1. 4V Model Analysis

Dimension	Assessment	Score (1-5, 1 being the lowest)
Volume	The BRFSS collects monthly data from individuals across the United States with an annual total of over 350,000 responses. The training dataset used specifically for this project contains 70,693. This volume of data is enough to train ML models.	5
Variety	The cleaned dataset includes responses from individuals aged 18 an older with diverse behaviors across all risk factors. However, the original dataset is imbalanced with the number of people not having diabetes outweighs the opposite. A 50-50 split balanced dataset was used for the purposed of this study.	3

Velocity	The CDC and BRFSS make data available for each year. The dataset used for this project only utilizes a year's worth of data. However, this study can be expanded to include more data before and after 2015.	3
Veracity	The dataset represent real-world responses from individuals across the United States and data collected is facilitated by the United States government through the different states and the Center for Disease Control and Prevention. A cleaned subset of this dataset was available on Kaggle and utilized in this study.	4

Figure 3 shows the net chart created from the assessment. This net chart indicates a feasible dataset to be utilized for our analysis.

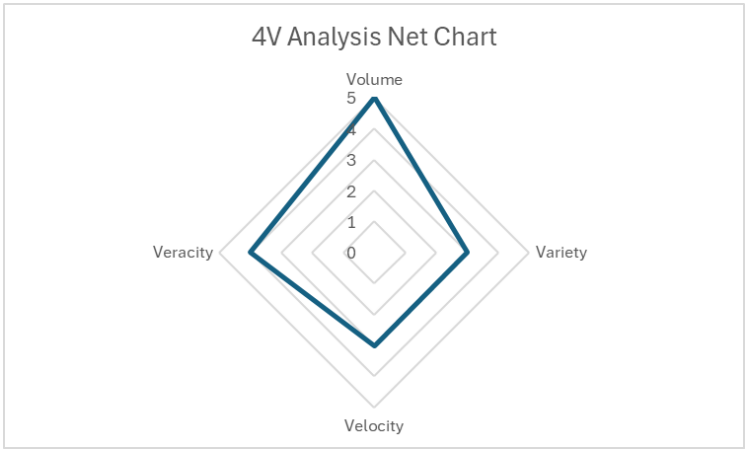

Figure 3. 4v Analysis Net Chart

Business Impact Storyboard

Predicting Diabetes using Behavioral Risk Factors

	SETUP	ACTIONS	OUTCOMES	RESULTS
Current implementation	• BRFSS uses Computer-Assisted telephone interview systems to conduct interviews regarding current health-related perceptions, conditions, lifestyle and behaviors. • Data is available on the CDC website for public use.	• CDC processes the data submitted by the states and staff run editing programs and data quality checks. • Data is sent to states and made publicly available for analysis.	• BRFSS data help states establish and track state and local health objectives, plan health programs, implement disease prevention and health promotional activities, and monitor trends.* • Data is also used to support health-related legislative efforts. *from cdc.gov	• Health priorities and emerging health problems are identified. • Trends in health risk behaviors and outcomes over time • Public health programs and policies to increase awareness regarding health and risk behaviors.
Future implementation	• More uniform data collection through the implementation of the Finest Online Records Management System (FORMS). • Detailed traffic safety analysis can be conducted	• Analyze historical data to determine key drivers for chronic illnesses such as diabetes. • Develop predictive models that calculate the likelihood of diabetes or other illnesses among individuals. • Explainable model to provide insights for causes of diabetes and other illnesses.	• Top factors that cause diabetes will be identified. • A model to predict the likelihood of diabetes given lifestyle choices and behavioral risks. • Results will be available in a BI dashboard.	• Awareness about the risk factors associated with diabetes and to promote lifestyle changes to mitigate these risks and reduce the chances of developing the disease

Figure 4. Business Impact Storyboard

PHASE TWO: DATA AND MODEL PREPARATION

The methodology of this study follows the 7-step process described by Tobias (2022).

1. Collect Historical Datasets

A clean and balanced subset of the BRFSS was obtained from Kaggle in CSV format. The dataset contains no missing values. The dataset is also tidy, meaning each observation is in its own row, each variable is in its own column and every measurement is a cell.

Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	0	0	1	12	1	0	0	1	0	0	0	1	0	2	15	0	0	1	11	6	7
0	1	1	1	13	1	0	0	0	0	0	0	1	0	4	30	30	1	0	10	5	1
0	0	1	1	13	0	0	0	1	1	1	0	1	0	1	0	0	0	1	11	6	8
0	0	0	1	13	1	0	0	1	0	1	0	1	0	1	0	0	0	1	4	6	8
0	0	0	1	13	0	0	0	1	1	1	0	1	1	5	30	30	0	0	6	4	8
0	0	0	1	13	0	0	0	1	1	1	0	1	0	3	0	0	0	0	11	4	1
0	0	0	1	13	0	0	0	1	0	1	0	1	0	4	3	10	1	0	7	6	5
1	0	1	1	13	0	0	0	0	0	1	0	1	0	4	0	2	1	1	6	4	8
1	1	0	1	13	1	0	1	1	0	0	0	1	0	5	0	8	1	1	9	5	2
0	0	0	1	14	1	0	0	1	0	0	0	1	0	5	10	30	1	0	8	3	1
0	0	0	1	14	0	0	0	0	1	0	0	1	0	1	0	0	0	1	8	5	5
0	0	0	1	14	1	0	0	1	1	1	0	1	0	4	0	18	1	0	10	6	8

Figure 5. Dataset features and target variable

2. Identify Features and Labels

The target variable in this dataset is binary, indicating whether an individual has diabetes or not. The dataset includes a total of 20 features, each representing a potential risk factor for diabetes. These risk factors encompass a range of demographic, behavioral, and health-related attributes. The following risk factors are included in this analysis:

1. blood pressure (high)
2. cholesterol (high)
3. smoking
4. diabetes
5. obesity
6. age
7. sex
8. race
9. diet (fruits intake)
10. diet (vegetable intake)
11. exercise
12. alcohol consumption
13. BMI
14. Household Income
15. Marital Status
16. Sleep
17. Time since last checkup
18. Education
19. Health care coverage
20. Mental Health

3. Training Data Split

This project utilizes AutoML which automatically handles the training-test data split.

4. Select Algorithms

A diagnostic analysis was done to explore underlying patterns in the observations. A key influencer analysis was done on the data to bring these patterns to the surface.

The key objective of this project is to predict whether an individual is likely to have diabetes based on their demographic, behavioral, and health-related attributes. Since this task involves categorizing individuals as diabetic or non-diabetic, it is framed as a classification problem. Consequently, the classification modeling option was selected in AutoML within Azure ML Studio.

5. Evaluate model

In the initial diagnostic analysis, the training dataset was analyzed in Power BI with the Key Influencers Tool. Each of the risk factors is added to the "explain by" box with the diabetes target variable in the analyze input box. The top 3 factors influencing diabetes in this case are:

1. When the individual has high blood pressure
2. When the individual has high cholesterol
3. When the individual describes their general health above 4. (From the BRFSS codebook is fair to poor)

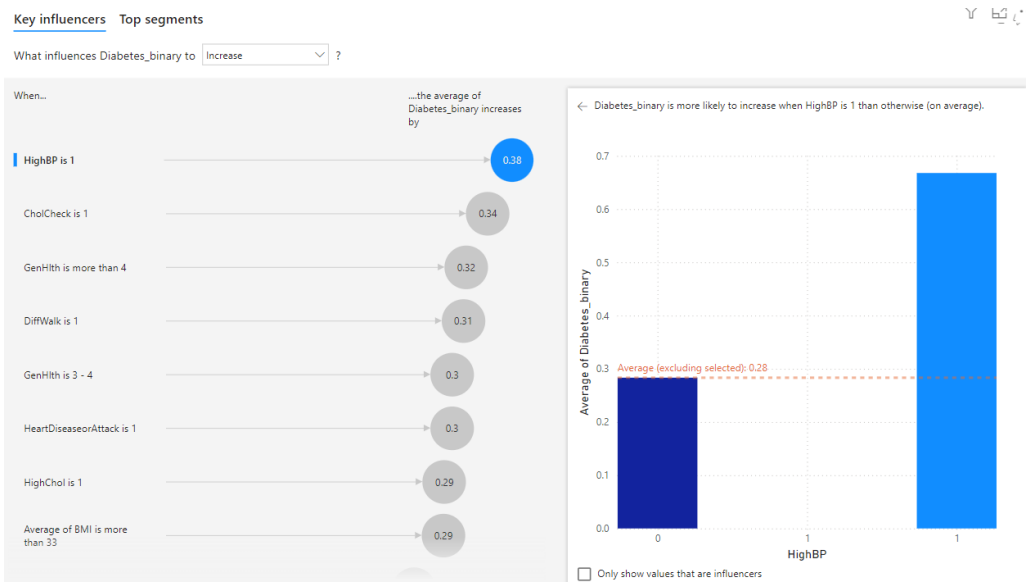

Figure 6. Key Influencers

For the classification task in AutoML, the best model identified was VotingEnsemble with a weighted precision score of 0.762 and accuracy of 0.761.

Algorithm name	Explained	Responsible AI	Precision scor... ↓	Sampling	Created on
VotingEnsemble	View explanation		0.76293	100.00 %	Jul 23, 2024 11:05 PM
MaxAbsScaler, LightGBM			0.76074	100.00 %	Jul 23, 2024 10:05 PM
SparseNormalizer, XGBoostClassifier			0.75746	100.00 %	Jul 23, 2024 10:28 PM
SparseNormalizer, LightGBM			0.75696	100.00 %	Jul 23, 2024 10:48 PM
SparseNormalizer, XGBoostClassifier			0.75656	100.00 %	Jul 23, 2024 11:01 PM
StandardScalerWrapper, XGBoostClassifier			0.75559	100.00 %	Jul 23, 2024 10:05 PM

Figure 7. Models generated in AutoML

A similar analysis with the key influencers chart on Power BI was generated by the AutoML. On this platform, the top 3 features that had the most influence on the likelihood of diabetes were:

1. High blood pressure
2. General health
3. BMI

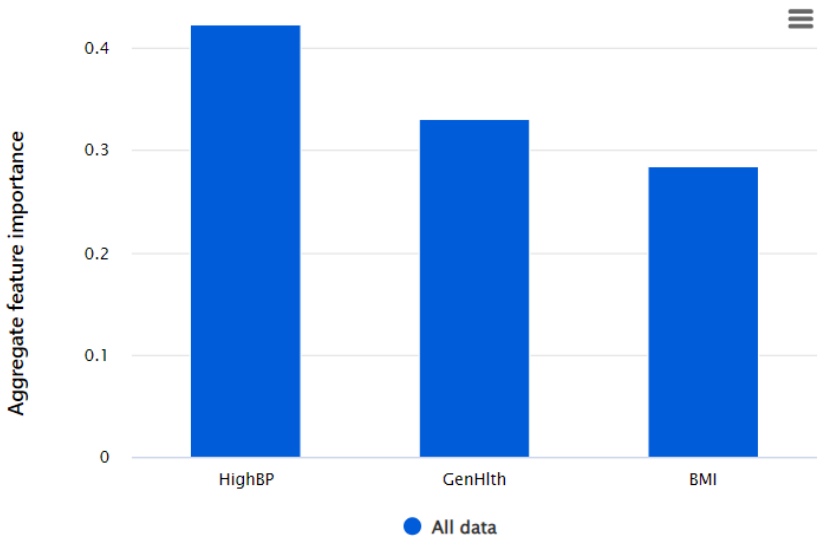

Figure 8. Aggregate feature importance on Azure

With this model, Below, is the normalized confusion matrix generated in Azure Machine Learning Studio. In this model, 71.78% of positive cases were correctly identified while 80.41% of negative cases were correctly identified.

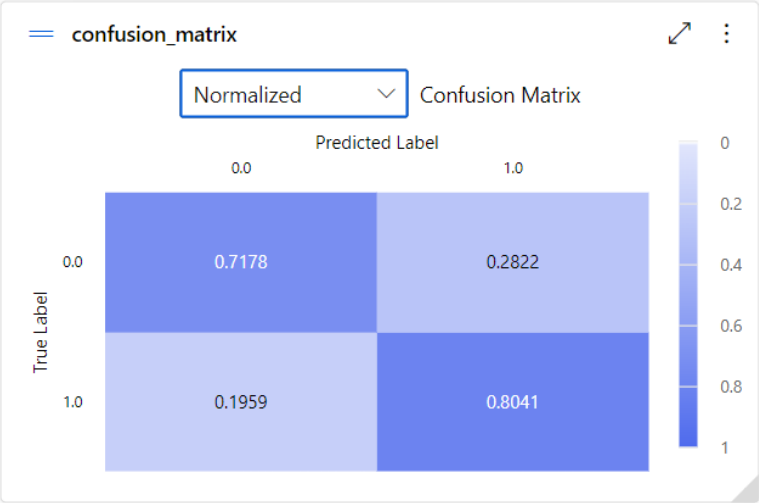

Figure 9. Confusion matrix of voting ensemble model

The metrics calculated for our model are as follows:

1. Average precision score macro – 0.822
2. Average precision score micro – 0.829
3. Average precision score weighted – 0.822
4. F1 score macro – 0.761
5. F1 score micro – 0.761
6. F1 score weighted – 0.761

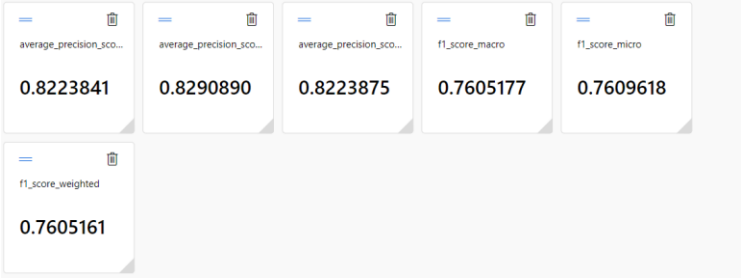

Figure 10. Metrics calculated from voting ensemble model

6. Deploy

For the predictive analysis, after evaluating the model, it was deployed to a web service. The REST endpoint URL was recorded to be used for predictions.

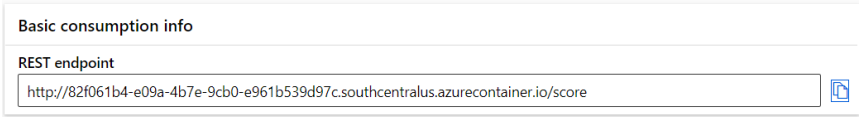

Figure 11. REST endpoint

To get inferences to the Power BI dashboard, the R script used by Tobias in case 7 classification was modified to fit this case. For the purposes of this project, authentication was disabled.

```r
# SECTION 0: Setup and Variables ----

# Make sure these packages are installed
library("httr")
library("rjson")
library("dplyr")

API_KEY = ""
API_URL = "http://82f061b4-e09a-4b7e-9cb0-e961b539d97c.southcentralus.azurecontainer.io/score"


# SECTION 1: API Request Function ----

inference_request <- function(HighBP, HighChol,CholCheck,BMI,Smoker,Stroke,HeartDiseaseorAttack,PhysActivity,Fruits,Veggies,HvyAlcoholConsump
,AnyHealthcare,NoDocbcCost,GenHlth,MentHlth,PhysHlth,DiffWalk,Sex,Age,Education,Income) {

  # Bind columns to dataframe
  request_df <- data.frame(HighBP, HighChol,CholCheck,BMI,Smoker,Stroke,HeartDiseaseorAttack,PhysActivity,Fruits,Veggies,HvyAlcoholConsump
,AnyHealthcare,NoDocbcCost,GenHlth,MentHlth,PhysHlth,DiffWalk,Sex,Age,Education,Income)

  req = list(
    Inputs = list(
      "data"=apply(request_df,1,as.list)
    ),
    GlobalParameters = list(
      'method' = "predict"
    )
  )

  # POST request - send JSON to API
  result <- POST(
    url = API_URL,
    add_headers(.headers = c('Content-Type' = "application/json", 'Authorization' = paste('Bearer', API_KEY, sep=' '))),
    body = enc2utf8(toJSON(req))
  )
  return(result)
}


# SECTION 2: Data preprocessing ----
# Fetch data from Power Query workflow
df <- dataset


# SECTION 3: Get Predictions ----
result <- inference_request(df$HighBP, df$HighChol, df$CholCheck, df$BMI, df$DSmoker, df$Stroke, df$HeartDiseaseorAttack, df$PhysActivity,
df$Fruits, df$Veggies, df$HvyAlcoholConsump, df$AnyHealthcare,df$NoDocbcCost,df$GenHlth,df$MentHlth,df$PhysHlth,df$DiffWalk,df$Sex,df$Age
,df$Education,df$Income )

# SECTION 4: Data postprocessing ----
result <- unlist(content(result))
df$Diabetes_Pred <- result

# SECTION 5: Format output for Power BI ----
output <- df
```

Figure 12. R script used as input in Power BI

A subset of the original dataset, consisting of 500 instances, was used as input in the Power BI dashboard. This csv file is named “diabetes-data-powerbi.csv.” The figure below displays the final Power BI dashboard:

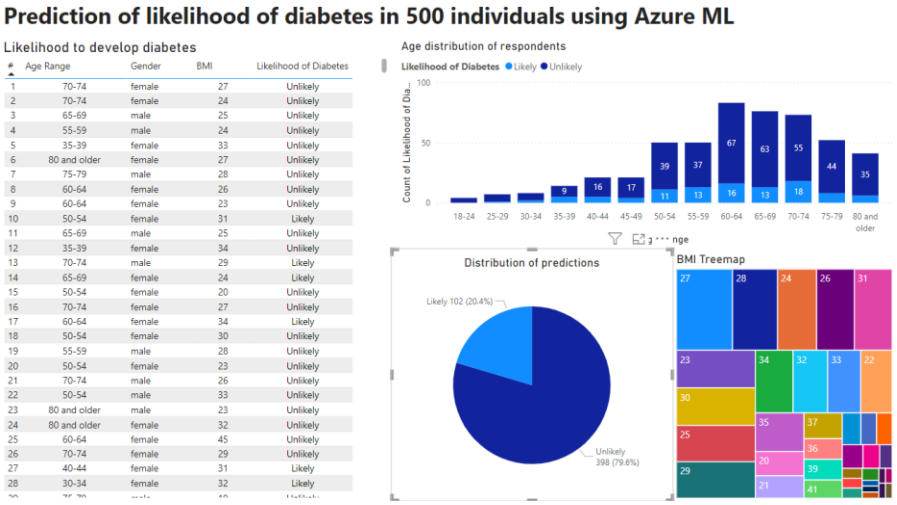

Figure 13. Power BI Dashboard

7. **Perform Maintenance**

As this is a prototype no further steps were initiated.

References

Center for Disease Control and Prevention. (2023, December 2). *Behavioral Risk Factor Surveillance System*. Retrieved from CDC: https://www.cdc.gov/brfss/annual_data/annual_2022.html

International Diabetes Foundation. (n.d.). *About Diabetes Facts and Figures*. Retrieved from International Diabetes Foundation: https://idf.org/about-diabetes/diabetes-facts-figures/

National Institute of Diabetes and Digestive and Kidney Diseases. (2024, January). *Diabetes Statistics*. Retrieved from National Institute of Diabetes and Digestive and Kidney Diseases: https://www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics

U.S. Centers for Disease Control and Prevention. (2024, May 15). *Diabetes Basics*. Retrieved from CDC: https://www.cdc.gov/diabetes/about/index.html

World Health Organization. (2023, April 5). *Diabetes*. Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Prevention,not%20smoke%20tobacco.