# **Exploring Predictive Models for Diabetes Risk Assessment Using Data Mining Techniques**

## **Zeta**

## **Provide Team Members’ Names and UIS Emails**

| **Names** | **UIS Email ID** |
| --- | --- |
| **Bhavana Yadav Krishnappa** | **bkris2@uis.edu** |
| **Shreya Vishwas** | **shre2@uis.edu** |
| **Sohan Shylaja Mohan** | **sshyl2@uis.edu** |

February 11, 2023

# **Data Mining Project Overview**

The goal of the project is to identify and define a problem the team want to solve using data mining techniques. For example, using the data related to understanding the trends that which types of cars can be grouped together to categorise the car for customers. This will be helpful for the dealers to meet customer satisfaction. Also, the used car data can be analyzed to identify the most significant factors in predicting the price (numerical target) of the car or predicting whether the car is expensive or inexpensive (categorical target).

For a project, teams will identify the problem, obtain the data, and define the scope of the problem. Team needs to identify a data set with at least 8000 observations or records. There should be at least two problem questions. One of the questions needs to be addressed using predictive analysis (using supervised techniques) and the other using exploratory analysis (unsupervised techniques). The team will use visualization for data exploration and preparation. Then, the team will implement several supervised and unsupervised data mining techniques to solve the identified problem. Appropriate performance evaluation metrics will be used to evaluate the performance of these supervised and unsupervised models to establish the significance of the developed models with highest accuracy.  Finally, scoring the predictive models (supervised models) on new data will be performed. Scoring plays a crucial role in data mining as it allows in confirming the reliability of the developed model by deploying the model on new data.

To assist teams in solving a feasible and interesting problem using several data mining techniques, each team should obtain the instructor’s approval regarding the scope and nature of the problem, the data set, and the nature of the analysis. *Teams will use Decision Trees, Regression, and Neural Network supervised techniques for comparing and identifying the best predictive model (can be adapted from predictive analysis assignment)* The *unsupervised techniques of Cluster Analysis will be used by team for exploratory analysis (can be adapted from exploratory analysis assignment)*. Overall, integrating the learning outcomes from individual course Assignments (*data exploration and visualization, exploratory analysis, and predictive analysis*) will be helpful to implement the project. **The team will use SAS Enterprise Miner to implement the project (exploratory and predictive analysis). Tableau or Excel can be used for data preparation and visualization.**

# **Project Proposal**

To start with, teams must submit a project proposal by answering the questions below (as much as you know). Identifying a problem that interests you and you would like to address is crucial in data mining. *The goal of the project proposal is to identify the problem, obtain the data, and define the scope of the problem.* *This will also include identification of tentative predictor and target variables.* *Team needs to identify a data set with at least 8000 observations or records.*

# **Answer following questions as a part of your project proposal.**

## **What is (are) the problem(s) for investigation? Briefly discuss the problem you want to investigate (***You can refer to Chapter 21 of the Textbook (pp. 499) to see some examples of problem definitions).* **Additionally, discuss why this problem is important for a data mining project.**

Diabetes is an emerging chronic disease affecting people worldwide to address this public health concern this project aims to improvise the Behavioral Risk Factor Surveillance System Dataset to predict diabetes risk factors to empower proactive intentions.

The objective of this project diverges into 2 branches: predictive modeling which employs machine learning algorithms to develop model forecasting and exploratory analysis delves into the various demographic variables and lifestyle behaviors.

This granular understanding not only informs public health policies but promotes a healthier lifestyle. The significance of data mining in handling the diabetes epidemic cannot be overstated. Its advanced analytical techniques, including machine learning, clustering, and regression analysis, unlock hidden patterns. The project aligns seamlessly with ongoing efforts to improve the health sector. Through a collaborative and interdisciplinary approach, this project provides effective diabetes management and prevention.

## **Develop tentative question(s) you want to address in your project** [*Be very specific and* ***write the questions****. There should be at least two questions, one for predictive analysis and the other for exploratory analysis. These questions will evolve and can change while working on the project.*]

## ​​**Predictive Analysis**: Predict the likelihood of an individual having diabetes based on their demographic and health-related factors?

## **Exploratory Analysis**: What are the key risk factors associated with diabetes development among individuals surveyed in the BRFSS dataset?

## **Information about the dataset. [***Team can obtain the**secondary data set of their choice from various web sources. Kindly refer to* [*Appendix*](#_heading=h.gjdgxs) *for some recommended sources. If team member has any data set, they want to use, please feel free to use the data set with appropriate approval and usage protocol.***]**

* 1. **What is the source of the dataset?** [*Be specific. Provide the URL if applicable.*]

BRFSS dataset for the year 2015 obtained from Kaggle<https://www.kaggle.com>

* 1. **What is the dataset about**? **Explain briefly.** **Conduct some literature search to identify where and how the data set you have selected for your project has been used. This will help you to identify the gap and address the problem in a new and/or different way** [*Not more than two paragraphs. Include references and in-text citations using APA. The data and variable description are usually available at the data source in the form of data dictionary and/or codebook.]*

The BRFSS dataset (Diabetes Health Indicators Dataset, n.d.), comprises responses from over 400,000 Americans by exploring the intricate relationship between demographic factors, lifestyle behaviors, and the presence of diabetes. We focus on the ‘**diabetes\_012\_health\_indicators\_BRFSS2015.csv’ file** which contains 253,680 survey responses which consists of both survey questions as well as calculated variables from individual responses. The Target variable being “Diabetes\_012” showcasing three classes:

| Target variable class 0 | diabetes or only pregnancy |
| --- | --- |
| Target variable class 1 | prediabetes |
| Target variable class 2 | diabetes |

Utilizing the comprehensive features, we aim to analyze the data, answer the questions and understand the risk factors of diabetes.

3.3 **Which variable in your dataset will be your target variable for predictive analysis?** **Provide a brief description (at least one line) of your target variable.** [*Be specific. You can only have ONE target variable. The target variable is what you are going to predict in your analyses. If you are going to do unsupervised analysis, which does not have a target variable, please say so*].

| **Target Variable Name** | **Variable Data Type *[e.g. Numeric, Nominal, Ordinal, Binary]*** | **Description *[can be obtained from* *data dictionary and/or codebook]*** |
| --- | --- | --- |
| ***Diabetes\_012*** | ***Categorical*** | ***Diabetes\_012, which categorizes individuals into three groups: 0 for no diabetes, 1 for prediabetes, and 2 for diabetes*** |

**3.4 Which variables in your dataset will be your predictor variables for predictive and exploratory analysis?** **Provide brief description (at least one line) of each predictor variable.** [*Be specific. The predictor variables are the variables that will help you predict the target variable. These predictor variables might evolve over time but based on your preliminary analysis of data chose some predictor variables. The same or different predictors/input variables can be used to perform the exploratory analysis*].

| ***Input/Predictor Variable(s) Name*** | ***Used for Exploratory Analysis or Predictive Analysis or Both*** | ***Variable Data Type***  ***[e.g. Numeric, Nominal, Ordinal, Binary]*** | ***Description [can be obtained from data dictionary and/or codebook]*** |
| --- | --- | --- | --- |
| ***HighBP*** | *Both* | *Binary* | ***Indicates presence of high blood pressure (0 for no, 1 for yes).*** |
| ***HighChol*** | *Both* | *Binary* | ***indicates the presence of high cholesterol (0 for no, 1 for yes).*** |
| ***BMI*** | *Both* | *Numeric* | ***Body Mass Index*** |
| ***Smoker*** | *Both* | *Binary* | ***Indicates smoking status (0 for non-smoker, 1 for smoker)*** |
| ***PhysActivity*** | *Both* | *ordinal* | ***indicates level of physical activity (0 for low, 1 for moderate, 2 for high)*** |
| ***Fruits*** | *Predictive analytics* | *numeric* | ***frequency of fruit consumption (servings per day)*** |
| ***Veggies*** | *Predictive analytics* | *numeric* | ***frequency of vegetable consumption (servings per day)*** |
| ***Age*** | *Both* | *numeric* | ***represents age of the individual*** |
| ***GenHlth*** | *Both* | *ordinal* | ***Represents if individuals self rated general health on scale of 1 to 5*** |
| ***HvyAlcoholConsumption*** | *Both* | *binary* | ***Indicates if the individual is heavy drinker or not.*** |
| ***Stroke*** | *both* | *binary* | ***Indicates of the individual ever had stroke.*** |
| ***Sex*** | *both* | *binary* | ***Indicates indivdual’s sex*** |

## **Provide the timetable (Project Plan) for your project.**

[*Provide milestones and plan the activities and show them in a tabular format. For example, milestones will be project proposal and project report; and activities will be to define problem, obtain data, data exploration, data cleaning and preparation, variable selection-dimension reduction, developing model based on different data mining techniques (supervised and/or unsupervised), performance evaluation, model comparison, finalizing model, interpreting, and disseminating the result. Refer to Figure 2.1 (Shmueli et al. 2018, pp. 15) to identify the schematic of data modeling process*].

| **Define Purpose** | **Obtain Data** | **Explore and clean Data** | **Determine Data Mining Task** | **Choose Data Mining Method** | **Apply the method and select a final method** | **Evaluate Performance** | **Deploy** | **Disseminate Findings** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **completed** | **completed** | **Currently working** | **Currently working** | **Will complete by week 7** | **Will complete by week 7** | **Will complete by week 7** | **Will complete by week 8** | **Will complete by week 8** |

## **References**

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., and Lichtendahl Jr, K. C. 2018. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons.

*Diabetes Health Indicators Dataset*. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

# **Appendix: Recommended Data Sources**

Potential data sources can include, but are not limited to the Internet (e.g., search for “datasets for data mining”), current or past employers (with permission), public databases, university datasets, datasets that you collected in the past. Some sources of publicly accessible datasets are as follows. You might need to create a profile on some of these websites to access the content. Using university email ids to create the profile might provide additional benefits.

* <https://www.kaggle.com/datasets>
* <https://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free>
* <http://www.kdnuggets.com/datasets/index.html>
* <https://www.kdnuggets.com/datasets/government-local-public.html>
* <https://aws.amazon.com/datasets?_encoding=UTF8&jiveRedirect=1>
* [Google Dataset Search](https://toolbox.google.com/datasetsearch)
* <https://www.cdc.gov/data.html>
* <https://data.world/datasets/open-data>
* <https://www.nhtsa.gov/research-data/crash-injury-research>