Prediction of Early Onset Diabetes Using Classification Algorithms

By the Hardworking Procrastinators

Christopher Panlasigui[1], Jayashree Kapoor[2], Tia Whiteman[3], Wai Moon Chin[4]

[1,2,3,]4City University of N.Y. - Baruch College Zicklin School of Business

Data Mining for Business Analytics

Abstract

Over a third of the US population is afflicted by pre-diabetes or diabetes which leads to other health complications and oftentimes results in deaths. Early detection of the disease can avoid or delay complications. With the aid of data mining methodology and machine learning algorithms, we are going to construct a predictive model to detect the early onset of diabetes from the diabetes dataset obtained from UCI Machine Learning Repository.

Introduction

Data mining or data dredging as statisticians used to call it, started as a negative term to refer to the extraction of information that is not supported by data.[1] These days, it is used positively to refer to the process of creating a statistical model from data, often with the aid of machine learning algorithms.

Our goal is to create a predictive model that detects the early onset of diabetes using data mining techniques. According to the Centers for Disease Control (CDC), more than 37.3M Americans have diabetes, and 96M US adults have prediabetes. 20% of the Americans that have diabetes and 80% of the US adults that have prediabetes are unaware that they have it. The disease can lead to other health problems such as heart disease, nerve damage, vision loss, chronic kidney disease, hearing loss, mental health, and more. It is also the 7th leading cause of death and costs our economy 327B yearly.[2] The model that we hope to create can be used as a litmus "test" for individuals and families to determine if they are at risk of the disease before it develops. In addition, it can be added to guidelines for improving prognosis by healthcare professionals to help prevent or delay diabetes complications.

The nature of this project will involve classification since the response variable and all but one of our predictors in our dataset are all qualitative. Considering our goal and various sources about the disease, we are going to determine whether someone is at risk (positive) or not at risk (negative) for the early onset of diabetes based on a set of symptoms. More specifically, what combination of symptoms is a good indicator?

The rest of the paper is organized into the following sections: the data mining task section covers the specific tasks to be performed and goals for the tasks to answer specific questions about our goal and dataset; the dataset section provides the origin, overview, and attributes of our dataset; the methods and models section covers the data mining methods we plan to employ to achieve the goals we set in the data mining task section; the assessment section provides information on how we will create our training and test sets, and methodology we will use to validate our models; the presentation and visualization describe how our results will be presented and visualized; the roles section contains the roles each group member will have in the project; finally the schedules section contains the table of dates and tasks that we plan to complete by.

Data Mining Task

1. Descriptive
   a. Summarization to provide us a general overview of our data. For example, we wish to answer the questions, what is the distribution of Age for all the categorical groups? What is the distribution of Age for several combination of categorical groups? Does it make sense to convert our Age variable into fewer ordinal groups for analysis?
2. Predictive
   a. Classification can provide us the set or different combinations of attributes, if any, that are likely to predict the early onset of diabetes.

Data Set

We obtained our dataset from UCI Machine Learning Repository.[3] The data was originally collected using direct questionnaires from patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh.[4] It has 520 observations with 17 variables, which are all categorical except for Age.

The variable Age is numeric that has a minimum value of 16, mean value of 48.03, and maximum value of 90; gender comprises 328 Males and 192 Females; Polyuria, a term for excessive urination, has 258 Yes and 262 No; Polydipsia, a term for excess thirst, has 233 Yes and 287 No; sudden weight loss has 217 Yes and 303 No; Weakness has 305 Yes and 215 No; Polyphagia, a term for excessive eating has 237 Yes and 283 No; Genital thrush, a term for common yeast infection caused by fungus candida has 116 Yes and 404 No; visual blurring has 233 Yes and 287 No; Itching has 253 Yes and 267 No; Irritability has 126 Yes and 394 No; delayed healing has 239 Yes and 281 No; partial paresis or more commonly known as partial paralysis has 224 Yes and 296 No; muscle stiffness has 195 Yes and 325 No; Alopecia, an immune system-induced hair loss has 179 Yes and 341 No; Obesity has 88 Yes and 320 No; class, has 320 Positive and 200 Negative.

There are no specific challenges in our dataset, nor it contains any missing values. However, our dataset restricts our data mining procedures to classification. We intend to use all variables.

Methods and Models

Since all but one of the variables are categorical, our predictive data mining task is restricted to classification methods. Thus, we intend to implement multiple Logistic Regression, Naïve Bayes classifier and Random Forest methods. We may perform binning data transformation on the Age variable to turn it into ordinal categorical data to improve model performance. Finally, categorical values will be recoded.

Multiple Logistic Regression

For the multiple logistic regression model, the R function glm() will be used to perform numerous univariate analyses, that is identify the relation of the response with each predictor one a time and investigate any association to scientific plausibility. We will compare a full model (with all variables) with 2 reduced models obtained from the aggregate of predictors from univariate analysis and from the backward stepwise selection. Since the predictors are categorical, we cannot think of the model outcome as a multiplicative model of odds. As an example, we are not going to be able to talk about the association of response increase with a single unit increase in X_1, holding other (X_2, … X_{16}) constant, because the results depend on the actual (numerical) value of X_1. Rather, the presence or absence of individual predictors in our model will provide the odds of being negative/positive for the disease.

Naïve Bayes

The naïve Bayes classifier is computationally efficient and able to handle categorical variables directly, so we do not need to re-code our categorical variables. It is oftentimes cited to provide good performance when the goal is classification or ranking. The function naiveBayes from the CRAN package e1071 will be used to perform the naïve base classifier. Since it is based on conditional probabilities assuming independence, our model can provide the likelihood of being positive/negative for the disease based on the existence or absence of different symptoms. Since it is best suited for categorical classifier, we hope that it would perform better than Logistic Regression.

Random Forest

Since we want to predict whether a person is positive for early onset of diabetes based on a combination of symptoms. This prediction will be based on a decision tree where the root node and the internodes represent the symptoms, and the terminal/leaf nodes represent the output or test whether someone is positive or negative for the disease. Random Forest is a machine learning model that uses an ensemble of decision trees to make its prediction. It randomly samples data and builds an ongoing series of decision trees on a subset. Since it is random, it reduces overfitting and bias.

Assessment

To assess our models, we will use stratified holdout, 80:20 resampling, and (k=10)-fold cross-validation. No baseline model will be used since we do not have any continuous variable except for Age. Then compare models' performances in each resampling method. In each test, the performance of each model will be evaluated by confusion matrix, and misclassification error.

Presentation and Visualization

Some of the results will be presented in tables as in the case of confusion matrices. Tree graphs will also be used as in the case of random forests. Finally, some results will be presented graphically using bar charts as in the case of visualizing the distribution of our dataset.

Roles

Roles of each member will be determine

Schedule

Date	Task to be Completed	Date Completed
10/24/2022	Log. Regression; add model/findings into draft	
10/31/2022	Naïve Bayes in R; add model/findings into draft	
11/07/2022	Random Forest in R; add model/findings into draft	
11/14/2022	Compare performance; draft Intro and Results; **SUBMIT PROGRESS REPORT**	
11/21/2022	Conclusion and abstract; **FINALIZE Project paper**	
11/28/2022	Power Point	
12/05/2022	SUBMIT FINAL REPORT, R code and Data	

Bibliography

1. Rajaraman A, Ullman JD. 2012. Mining of massive datasets. New York, N.Y. ; Cambridge: Cambridge University Press.
2. Diabetes facts. 30 Sept. 2022. Center for Disease Control and Prevention; [accessed 14 Oct. 2022]. https://www.cdc.gov/diabetes/basics/quick-facts.html.
3. Early stage diabetes risk prediction. 2020. UCI Machine Learning Repository; [accessed 14 Oct. 2022]. https://archive-beta.ics.uci.edu/ml/datasets/early+stage+diabetes+risk+prediction+dataset.
4. Islam MMF, Ferdousi R, Rahman S, Bushra HY. 2019. Likelihood prediction of diabetes at early stage using data mining techniques. Computer Vision and Machine Intelligence in Medical Image Analysis.