P6886 – Final Project Outline

Onset of Type-II Diabetes Mellitus (T2DM) Predictions

Qihang Wu

- Primary Question

[Main research question] In this final project, we aim to propose several machine learning algorithms that provide adequte predictive accuracy for the onset of Type-II diabetes mellitus (T2DM) in at-risk populations. More specifically, we would like to answer the following questions:

1) Predictive Accuracy: Can we accurately predict the onset of T2DM in at-risk populations via logistic regression (with penalties), random forests, support vector machines (SVMs), etc.?
2) Model Comparison: What are the strengths and limitations of different modeling techiniques?
3) Optimal Modeling and Predictors: Which model is preferred to make predictions of T2DM? Based on which metric(s) or measure(s)? What are the most significant predictors in this population? Any limitations?

[Significance & Potential Impact] T2DM is the most prevalent form of diabetes and arises from the body's ineffective use of glucose. Over time, this results in excessive glucose levels in the bloodstream, which can eventually lead to serious complications affecting the nervous and immune systems. Accurate prediction of T2DM using advanced statistical and machine learning models can significantly achieve the public health benefits. Early detection facilitates timely interventions and personalized treatment strategies, enhancing patient outcomes. Consequently, this can decrease the overall incidence of T2DM and mitigate its complications.

- Study Design and Data

The dataset used for the analysis is collected from Mendeley Data, a free and secure cloud-based data repository. The data is currently available online and is originally from the Vanderbilt datasets. It contains demographic factors (e.g., age, gender, height, and weight, etc) and some lab results (e.g., total cholesterol, glucose level, and HDL, etc) from a total of 390 (228 females and 162 males) rural African Americans in Virginia, USA. The details about the variables will be provided in the next Section. In addition, we would consider intergrating supplementary datasets that could offer additional variables or enable cross-validation of our proposed models in the final project.

- Primary Response and Predictors

The table of primary response and predicors are shown below.

	Variable Name in Data Set	**Description**	**Values**
Primary Response	Diabetes	Yes (60), No (330)	Nomial
Predictors:	Cholesterol	Total cholesterol	mg/dl; Numerical

	Glucose	Fasting blood sugar	mg/dl; Numerical
	HDL	HDL or good cholesterol	mg/dl; Numerical
	Chol/HDL	Ratio of total cholesterol to good cholesterol. Desirable result is < 5	Numerical
	Age	All adult African Americans	Numerical
	Gender	228 females, 162 males	Nomial
	Height	Measured in inches	inches; Numerical
	Weight	Measured in lbs	lbs; Numerical
	BMI	703 × weight (lbs)/[height (inches)]^2	Numerical
	Systolic BP	The upper number of blood pressure	mmHg; Numerical
	Diastolic BP	The lower number of blood pressure	mmHg; Numerical
	Waist	Measured in inches	inches; Numerical
	Hip	Measured in inches	inches; Numerical
	Waist/hip	Ratio is possibly a stronger risk factor for heart disease than BMI	Numerical

- Proposed Methods for Developing Predictive Models

The proposed models include logistic regression, random forests, and SVMs. Before employing the above models, we will need to ensure the data scaling for SVMs and possibly logistic regression. This is because the predictors with the largest range will completely dominate in the computation of kernel matrix in SVMs. Meanwhile, we will consider a grid search to find the optimal parameters for each model, particualy random forests. The strengths and weaknesses of each model are described as follows.

1) Logistic Regression: As one of the most simple machine learning models, this method is easy to understand, interpretable, and can give decent results as expected. In general, it performs well on binary classification tasks and is computationally efficient. However, this model assumes the linear relationship between response and predictors, which may not be the case. Here we consider the penalized logistic regression to reduce the possibility of overfitting and improve the generalizability.
2) Random Forests: This method is known for high accuracy and robust performance across diverse nature of predictors. It combines the predictions of multiple decision trees and reduce the risk of overfitting. However, it is difficult to interpret and can be computatially expensive.
3) SVMs: This method works relatively well when there is a clear margin of separation between diabetes and no diabetes. On the other hand, this approach may fail if too many overlapping cases.

- Proposed Methods for Evaluating Your Predictive Models

The evaluation metrics/measures include accuray, precision and recall, area under the ROC curve (AUC-ROC), confusion matrix, 10-fold cross validation (CV), and more. Some metrics are trivial to obtain, for example, the accuracy will be obtained by calculating the proportion of total correct predictions over all data points. Precision (i.e., positive predictive value) is a metric that positive classifications are actually positive, while recall is also known as sensitivity. Last but not least, we will evaluate the models using 10-fold CV to ensure their robustness and generazalibility.

- Potential Challenges

  Several challenges are stated as follows.

  1) Data imbalance: With a relatively high number of negative (no diabetes), model will tend to predict the majority class and develop a bias.
  2) Possibility of overfitting: With only 390 observations, the risk of overfitting is high, especially when the complex model like random forests is used. More careful tuning strategy should be apply.
  3) Difficulty to interpret: The complex nature of some proposed models make it hard to interpret, which is a serious issue in clinical settings as we are not restricted to the predictions.