Pima Indians Diabetes Prediction

Biol DB and Datamining: Manpreet S. Katari

Final due: May 18th 11:55pm

The Dataset

This dataset is a collection of samples obtained from the National Institute of Diabetes and Digestive and Kidney Disease.

The goal is to predict whether an individual has diabetes or not.

The attributes are :

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1) where 1 means they have diabetes.

Things to do

- Load the **PimaIndianDiabetes.csv** dataset.

Q1: Which variables are numeric and which are categorical? Explain why. (20pts)

Make sure to convert the variables to the correct type.

Q2: Use appropriate statistical analysis to determine which of the variables (on their own) are most helpful in predicting the outcome? (20pts)

- For each variable draw an appropriate graph (boxplot for numerical values and barplot for categorical values). Explain your answer.

Q3: Are there any missing values? and how do we fix this ? (20pts)

- Notice that some of the values are missing and replaced with 0. For example a blood pressure of 0 doesn't make sense and it can have an impact on the models. So how do we identify which 0 reoresents the value 0

and which represents NA? Explain your answer.

Q4: Building the Model (20pts)

- You are free to use either : leave out 30% or 10xfold Cross Validation for your testing purpose.
- You may pick any of the following to perform your modeling (randomForest, SVM, KNN, or neural nets)
- Explain your approach (for example, if you are only picking one, why did you pick it? if you are going to do all, how will you decide which is best?)

Q5: The Final Model (20pts)

- Determine which model worked the best and why you think so.
- Create a plot using colors to show which points are were predicted to have Diabetes and which were not. Use shapes to show which points were actually Diabetes and which weren't.

Submission

- You are expected to submit ALL your code with ALL the comments and discussion in the code.