- Homework 3 Spec sheet -

Please load and use the "diabetes.csv" data file. This dataset contains information from over 250,000 people obtained by the CDC.

The first row represents the column header.
Each row after that represents the information of one person.

Columns represent (in order):

1) Diabetes status (1 = has been diagnosed with diabetes, 0 = has not)
2) High blood pressure (1 = has been diagnosed with hypertension, 0 = has not)
3) High cholesterol (1 = has been diagnosed with high cholesterol, 0 = has not)
4) Body Mass Index (weight / height^2)
5) Smoker (1 = person has smoked more than 100 cigarettes in their life, 0 = has not)
6) Stroke (1 = person has previously suffered a stroke, 0 = has not)
7) Myocardial issues (1 = has previously had a heart attack, 0 = has not)
8) Physically active (1 = person describes themselves as physically active, 0 = does not)
9) Eats fruit (1 = person reports eating fruit at least once a day, 0 = does not)
10) Eats vegetables (1 = person reports eating vegetables at least once a day, 0 = does not)
11) Heavy Drinker (1 = consumes more drinks than the CDC threshold/week, 0 = does not)
12) Has healthcare (1 = person has some kind of healthcare plan coverage, 0 = does not)
13) NotAbleToAffordDoctor (1 = person needed to see the doctor within the last year, but could not afford to, 0 = did not)
14) General health: Self-assessment of health status on a scale from 1 to 5
15) Mental health: Days of poor mental health in the last 30 days (self-assessed)
16) Physical health: Days of poor physical health in the last 30 days (self-assessed)
17) Hard to climb stairs (1 = person reports difficulties in climbing stairs, 0 = does not)
18) Biological sex (1 = male, 2 = female)
19) Age bracket (1 = 18-24, 2 = 25-29, 3 = 30-34, 4 = 35-39, 5 = 40-44, 6 = 45-49, 7 = 50-54, 8 = 55-59, 9 = 60-64, 10 = 65-69, 11 = 70-74, 12 = 75-79, 13 = 80+)
20) Education bracket (terminal education is 1 = only kindergarten, 2 = elementary school, 3 = some high school, 4 = GED, 5 = some college, 6 = college graduate)
21) Income bracket (Annual income where 1 = below $10k, 8 = above $75k)
22) Zodiac sign (Tropical calendar, 1 = Aries, 12 = Pisces, with everything else in between)

We/you will want to use these variables in prediction models.

This data is carefully curated, so there should not be too much missing data (if any).
Diabetes is major metabolic disorder, with over 400 million people suffering from this illness worldwide. It is also a leading cause of death, with over 1.5 million annual deaths. It has been suggested that a large proportion of these are due to lifestyle choices. Therefore, for both prevention and treatment, it is critically important to find good predictors of diabetes. We will attempt to do so in this exercise.

Mission command approach: As per §4.5 of the Sittyba, we will tell you what to do ("answer these questions"), not how to do it. That is up to you. However, we want you to:

a) Do the homework yourself. Do not copy answers from someone else.
b) Restrict your methods (for now) to what was covered in the lecture/lab (in other words, logistic regression, SVM, decision trees, random forest, adaboost))
c) Include the following elements in your answer (so we can grade consistently):

Each answer should contain these elements:

1) A brief statement (~paragraph) of what was done to answer the question (narratively explaining what you did in code to answer the question, at a high level).
2) A brief statement (~paragraph) as to why this was done (why the question was answered in this way, not by doing something else. Some kind of rationale as to why you did x and not y or z to answer the question – why is what you did a suitable approach?).
3) A brief statement (~paragraph) as to what was found. This should be as objective and specific as possible – just the results/facts. Do make sure to include numbers and a figure (=a graph or plot) in your statement, to substantiate and illustrate it, respectively.
4) A brief statement (~paragraph) as to what you think the findings mean. This is your interpretation of your findings and should answer the original question.

Note: Brief actually means "brief". There is no need to write a dissertation. There is value to being concise. A couple of pages should be sufficient for the entire report. Do – however – write a report. A data and code-dump is not very useful or valuable in practice. People who pay you so they can ask you questions usually want them answered.

Please answer the following questions in your report:

1. Build a logistic regression model. Doing so: What is the best predictor of diabetes and what is the AUC of this model?
2. Build a SVM. Doing so: What is the best predictor of diabetes and what is the AUC of this model?
3. Use a single, individual decision tree. Doing so: What is the best predictor of diabetes and what is the AUC of this model?
4. Build a random forest model. Doing so: What is the best predictor of diabetes and what is the AUC of this model?
5. Build a model using adaBoost. Doing so: What is the best predictor of diabetes and what is the AUC of this model?

Extra credit:

a) Which of these 5 models is the best to predict diabetes in this dataset?
b) Tell us something interesting about this dataset that is not already covered by the questions above and that is not obvious.

Hint: There are many ways to assess what the best predictor is. The most straightforward thing to do here is probably to see which predictor drops model performance the most, if that predictor is not included in the model / if its labels are shuffled randomly.