The `diabetes.csv` comes from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements. All patients here are females at least 21 years old of Pima Indian heritage. The variables are

- `Pregnancies` The number of times pregnant
- `Glucose` Plasma glucose concentration at 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`
- `Age`
- `Outcome` (0 or 1)

Use `xgboost` library to construct a gradient boosting ensemble model to predict the patients outcome. Split the dataset into a train (66%) and test (33%) sets.

a) Find the best subset of features using variable importance

b) Track the test performance as more boosted trees are added to the ensemble. Plot the resulting learning curves and identify the best number of boosted trees.

c) Use a window of 10 rounds to find the best number of boosted trees to the ensemble.

Also use `GridSearchCV` and 10-fold cross validation to find the best number of trees, their depth, and the learning rate. Whenever is needed use `random_state = 1`.