Use precision, recall, and f1 to evaluate. F1 first. Evaluate recall and precision if f1 equals

1. Preprocess - Project\_Preprocessing.ipynb
   1. Remove all missing values
2. Logistic - Project\_Logistic.ipynb
   1. Use LASSO to regularize model, found BloodPressure and Insulin have more than 500 out of 1000 times with coefficient of zero
   2. 2 data set: cleaned dataset with all 8 input variables, called dataset 1, cleaned dataset with BloodPressure and Insulin dropped, called dataset 2.
   3. Extra dataset: start from initial dataset, drop BloodPressure and Insulin, and remove missing values observations for other variables. End up with a dataset with the same number of columns as dataset 2 but more observations, called dataset 3.
   4. Build Logistic models for all 3 datasets 50 times with kfold 20 folds and tuning cut-off values (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65)
   5. Repeat above steps with cut-off (0.3, 0.35, 0.4, 0.45, 0.5, 0.55)
   6. Repeat above steps with cut-off (0.4, 0.45, 0.5)
   7. Repeat above steps with cut-off 0.45
   8. Logistic models are good on dataset 1 and dataset 2, cut-off 0.45
3. Multilayer Perceptron - Project\_MLP.ipynb
   1. Build MLP with different parameters
      1. 1 or 2 middle layers
      2. relu or tanh as activation fuction for middle layers
      3. 2 3 or 4 neurons for middle layers
      4. on dataset 1 2 or 3
      5. cut-off values (0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5)
   2. All models perform worse than Logistic models.
4. Support Vector Machine - Project\_SVC.ipynb
   1. Build SVC with different parameters
      1. kernel as rbf, linear, poly, or sigmoid
      2. on dataset 1 2 or 3
      3. cut-off values (0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5)
   2. Repeat with cut-off (0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7)
   3. Repeat with cut-off (0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55)
   4. SVC are good on dataset 1 and dataset 2, cut-off 0.35
5. Ensembles - Project\_Ensembles.ipynb
   1. Build Random Forest, AdaBoost, and Grandient Boosting models to extract feature importances. 3 variables are less important: Pregnancies, BloodPressure, and SkinThickness
   2. 1 new dataset: cleaned dataset with 5 important variables, called dataset 4.
   3. Extra dataset: start from initial dataset, drop Pregnancies, BloodPressure, and SkinThickness, and remove missing values observations for other variables. End up with a dataset with the same number of columns as dataset 4 but more observations, called dataset 5.
   4. Build Random Forest, AdaBoost, and Grandient Boosting models with different parameters
      1. On 5 datasets
      2. Number of trees: 100, 500, 1000, 1500, 2000
      3. Depth of tree: 3, 5, 7
      4. Cut-off values (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65)
      5. Leaning rate for Boosting models: 0.1, 0.01, 0.001
   5. Best random forest is on dataset 4, 500 trees, depth 3, cut-off 0.35
   6. Best AdaBoost model is on dataset 4, 500 trees, depth 3, learning rate 0.001, cut-off 0.3 or 0.35
   7. Chosen Gradient Boosting model is on dataset 4, 1000 or 1500 trees, depth 3, learning rate 0.001, cut-off 0.3 or 0.35
6. Final Tests - Project\_Final.ipynb
   1. Build Logistic Regression, SVC, Ensembles on same dataset to compare
   2. Parameters to tune and choose from
      1. Number of features, cleaned dataset with 5, 6, or 8 features
      2. Cut-off values (0.3, 0.35, 0.4, 0.45)
   3. Choose best models on the same dataset and same cut-off: Random Forest, SVC, and Logistic using 5 features and cut-off value 0.35
   4. Stack these 3 models and use a random forest to predict looks to help boost the quality.