Task

Datasets for Parkinson's disease need to be found. Special attention should be paid to the parameters used there. It must be tabular data. The features in the dataset should be such that a person can measure them on their own or get them from tests in a hospital. If the dataset is not in csv format, you need to convert it to this format.

The result of the work is a word document of the following format:

- dataset name
- number of records
- list of features in the dataset
- description
- public dataset or not
- dataset link

Next, using one of the found datasets, build a model and give predictions. The result is an html version of the ipynb file with code and comments.

1. dataset name

   Parkinson's Disease Classification Data Set

2. number of records

   756 records

3. list of features in the dataset

   'id', 'gender', 'meanPeriodPulses', 'meanAutoCorrHarmonicity',

   'meanNoiseToHarmHarmonicity', 'meanHarmToNoiseHarmonicity',

   'meanIntensity', 'class', 'gender + class'

4. description

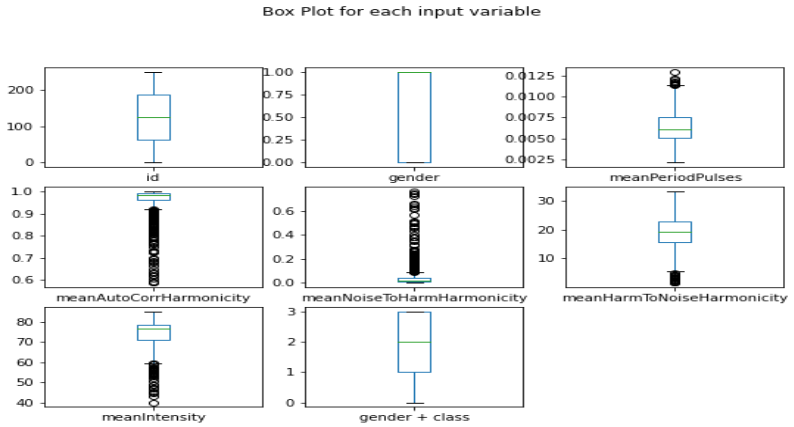
   The data is collected from 188 PD patients (107 men 81 women) with age range from 33 to 87 and 64 healthy individuals (23 men and 41 women) with age range from 41 to 82. I applied several machine learning models including logistic regression, decision tree, K-nearest neighbors, linear discriminant analysis, and gaussian naive bayes. I found the K-Nearest Neighbors' accuracy is the highest of all algorithms. Then I applied this algorithm to the features to get the prediction. I used the Decision boundary to visualize the data and have accurate predictions.

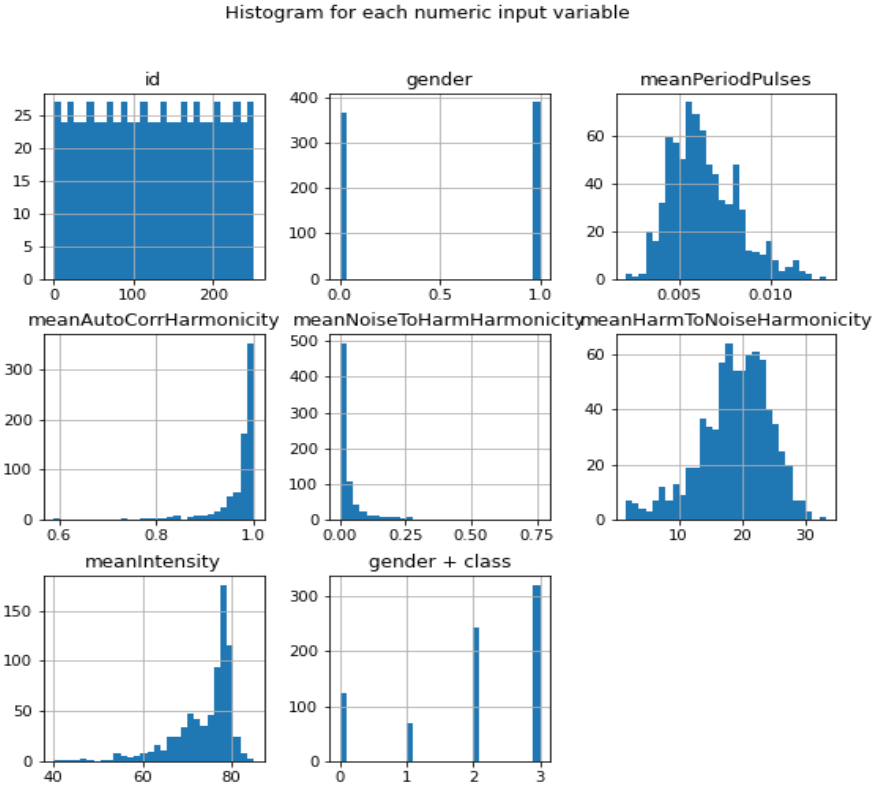
5. public dataset or not

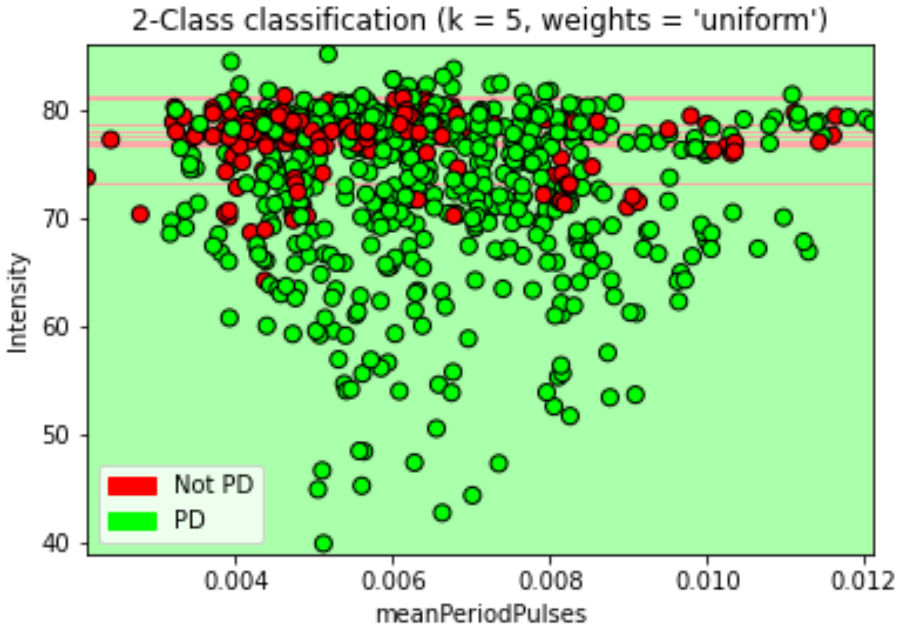
   it is public

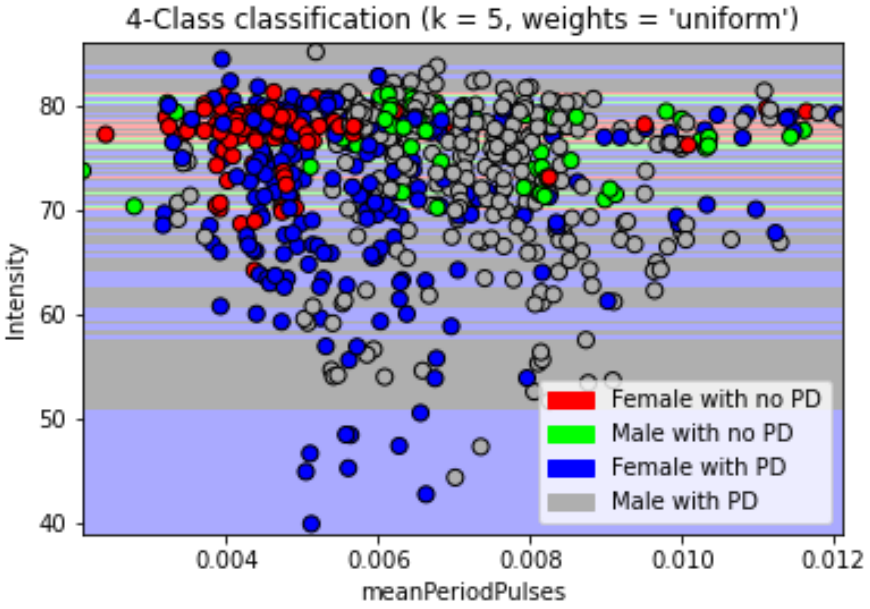
6. dataset link
   https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification#

7. Data









8. Prediction

According to the 2-classification decision boundary, healthy individuals tend to have a 70 to 80 intensity range, but the mean period pulses are scattered widely. Therefore, the intensity range could be the key to recognize the class. My test shows the accuracy of the model is over 80 %. Also, taking the gender difference into account, female PD patients have a wider range of intensity value compared to male patients. In general, male patients have a higher intensity value compared to female patients.