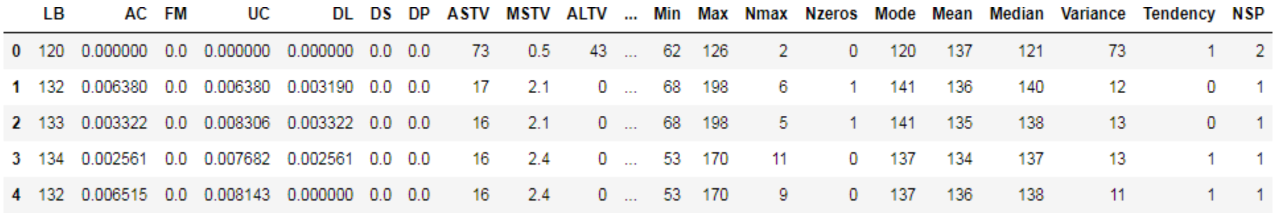
Random Forest on “Cardiotography” dataset

Problem Statement:

Consider yourself to be Sam, who is an established data scientist. You’ve been contacted by a medical company to come up with a model which would help in classifying whether the patient is ‘Normal’, ‘Suspected to have disease’ or in actuality has the ‘disease’.

Cardiotography Dataset:

The details regarding this dataset are present in the data dictionary



	LB	AC	FM	UC	DL	DS	DP	ASTV	MSTV	ALTV	...	Min	Max	Nmax	Nzeros	Mode	Mean	Median	Variance	Tendency	NSP
0	120	0.000000	0.0	0.000000	0.000000	0.0	0.0	73	0.5	43	...	62	126	2	0	120	137	121	73	1	2
1	132	0.006380	0.0	0.006380	0.003190	0.0	0.0	17	2.1	0	...	68	198	6	1	141	136	140	12	0	1
2	133	0.003322	0.0	0.008306	0.003322	0.0	0.0	16	2.1	0	...	68	198	5	1	141	135	138	13	0	1
3	134	0.002561	0.0	0.007682	0.002561	0.0	0.0	16	2.4	0	...	53	170	11	0	137	134	137	13	1	1
4	132	0.006515	0.0	0.008143	0.000000	0.0	0.0	16	2.4	0	...	53	170	9	0	137	136	138	11	1	1

Lab Environment: Jupyter Notebook

Domain: Medical

Tasks to be performed:

- Read the .csv file and understand the structure of the dataset.
- Make a scatter-plot between ‘ASTV’ & ‘MSTV’ columns
- Take the ‘ASTV’ column as the independent variable and ‘NSP’ column as the dependent variable
  - Divide the data into ‘train’ and ‘test’ sets with test size to be 30%
  - Build the random forest classifier on the train set, where the numbers of estimators are 300. Then predict the values on the test set
  - Build a confusion matrix and also find out the accuracy of the model built.
- Take ‘LB’, ‘ASTV’, ‘MSTV’ and ‘Variance’ as the independent variables and ‘NSP’ as the dependent variable
  - Divide the data into ‘train’ & ‘test’ sets with test size to be 30%
  - Build the random forest classifier on the train set, where the numbers of estimators are 100. Then predict the values on the test set
  - Build a confusion matrix and also find out the accuracy of the model built