Muzaffer Estelik

Cardiac Arrhythmia

Data Science Career Track Capstone-1 Project

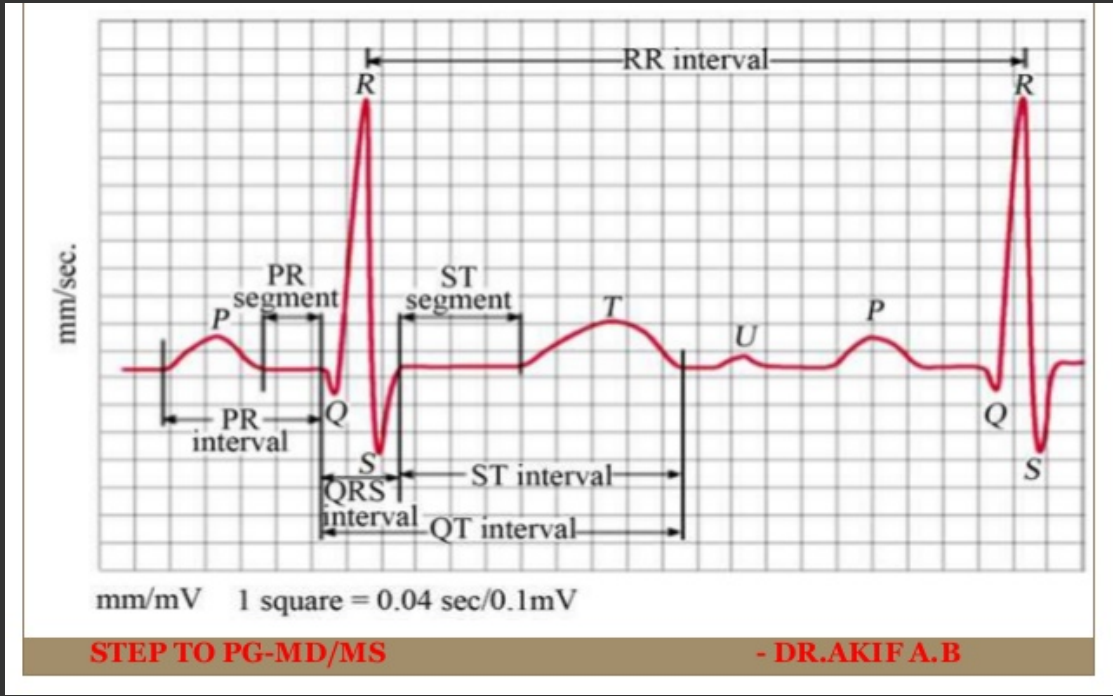
Aug 19th, 2018

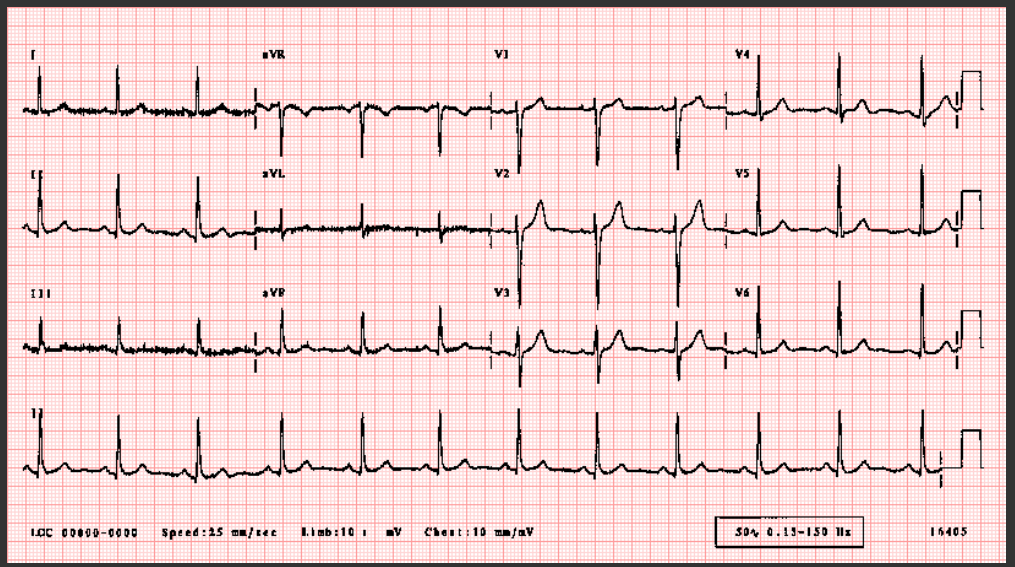
Table of Contents

Problem

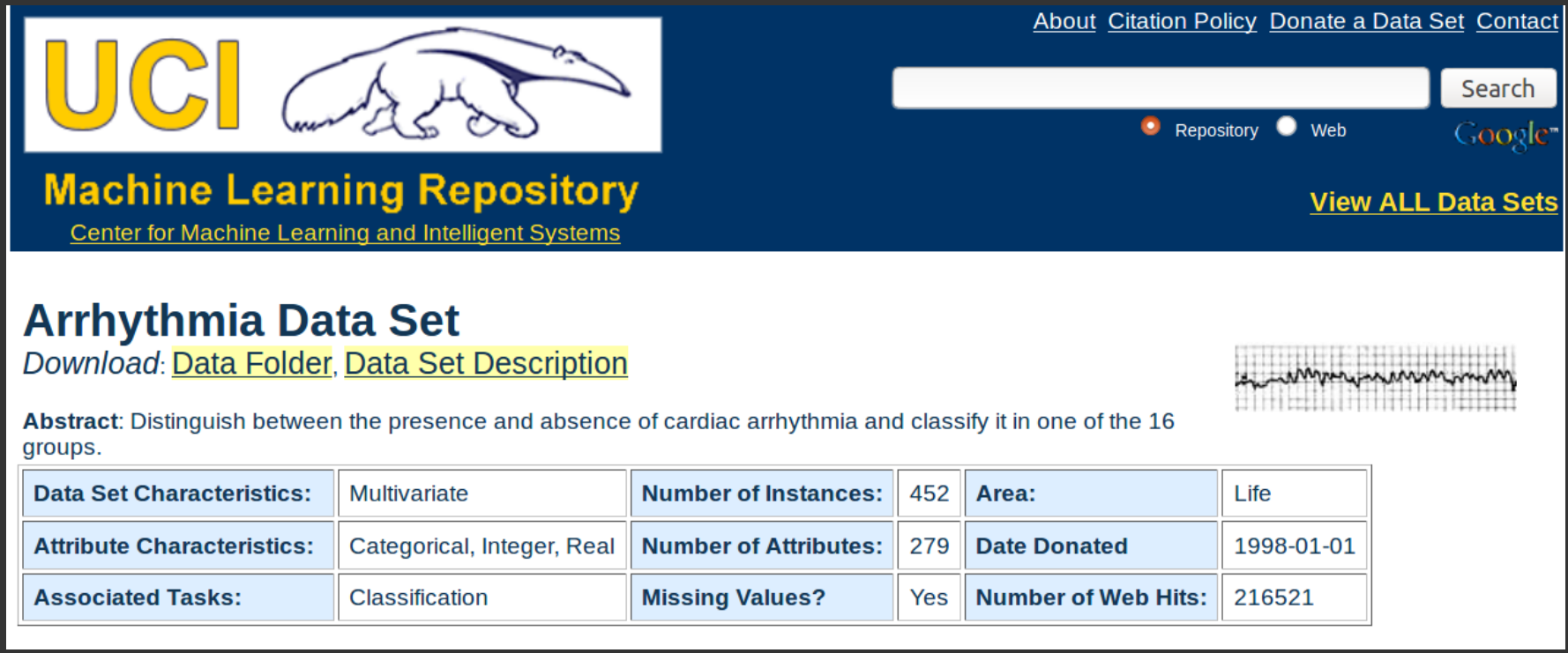
..famous about giving bad news..

Problem


RR interval
R
R
PR segment
ST segment
P
T
U
P
Q
Q
S
S
PR interval
QRS interval
ST interval
QT interval
mm/sec.
mm/mV
1 square = 0.04 sec/0.1mV
STEP TO PG-MD/MS
- DR.AKIF A.B



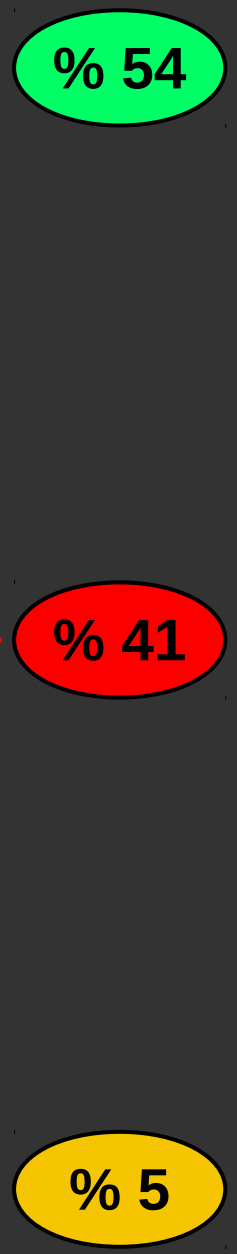
Data Set



https://archive.ics.uci.edu/ml/datasets/Arrhythmia

Data Set

Class code	Class	Number of instances
01	**Normal**	**245**
02	Ischemic changes (Coronary Artery Disease)	44
03	Old Anterior Myocardial Infarction	15
04	Old Inferior Myocardial Infarction	15
05	Sinus tachycardy	13
06	Sinus bradycardy	25
07	Ventricular Premature Contraction (PVC)	3
08	Supraventricular Premature Contraction	2
09	Left bundle branch block	9
10	Right bundle branch block	50
11	1. degree AtrioVentricular block	0
12	2. degree AV block	0
13	3. degree AV block	0
14	Left ventricule hypertrophy	4
15	Atrial Fibrillation or Flutter	5
16	Others	22



Data Set

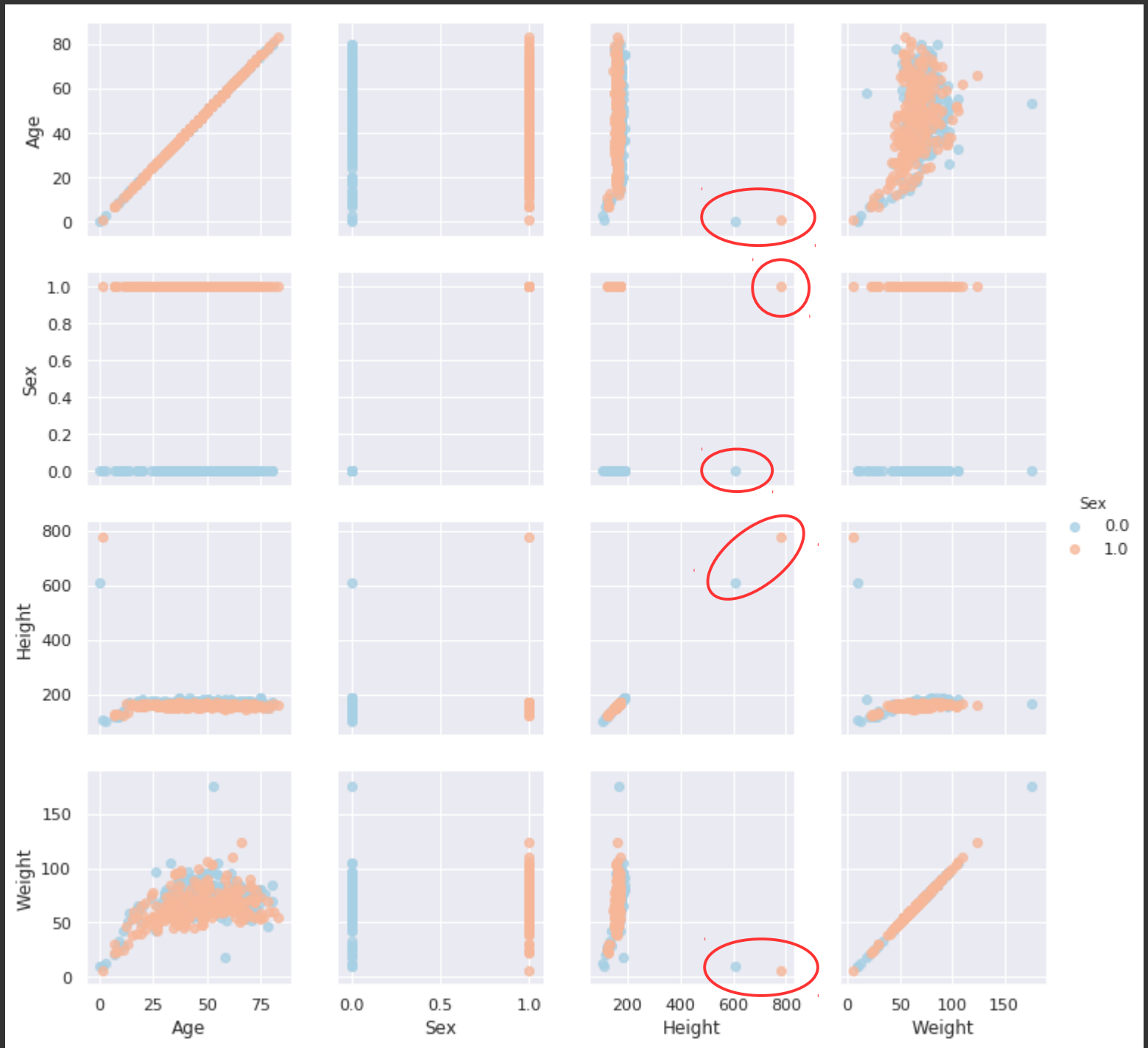
Class code	Class	Number of instances
01	**Normal**	**245**
02	Ischemic changes (Coronary Artery Disease)	44
03	Old Anterior Myocardial Infarction	15
04	Old Inferior Myocardial Infarction	15
05	Sinus tachycardy	13
06	Sinus bradycardy	25
07	Ventricular Premature Contraction (PVC)	3
08	Supraventricular Premature Contraction	2
09	Left bundle branch block	9
10	Right bundle branch block	50
11	1. degree AtrioVentricular block	0
12	2. degree AV block	0
13	3. degree AV block	0
14	Left ventricule hypertrophy	4
15	Atrial Fibrillation or Flutter	5
16	Others	22

% 54

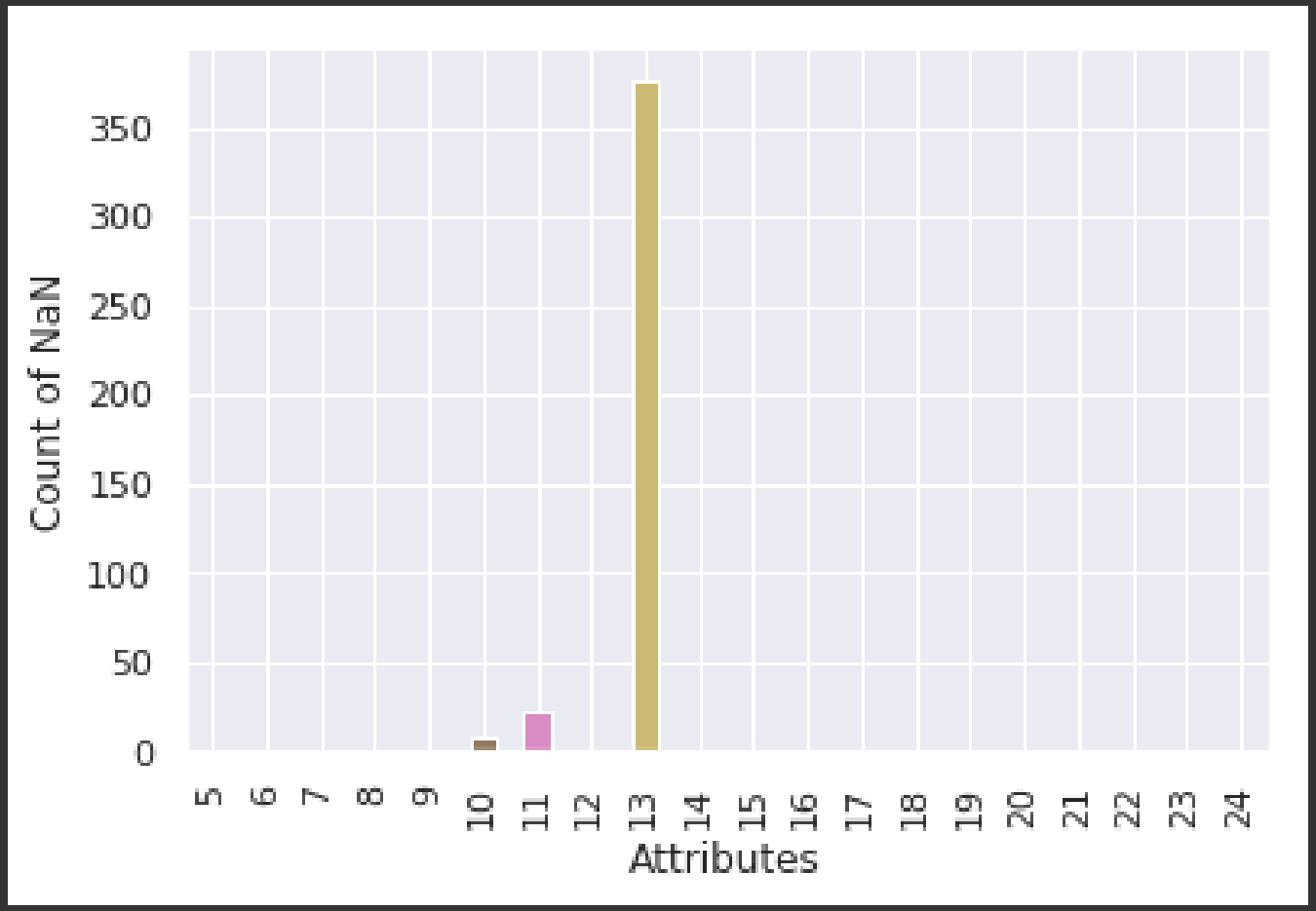
% 41

% 5

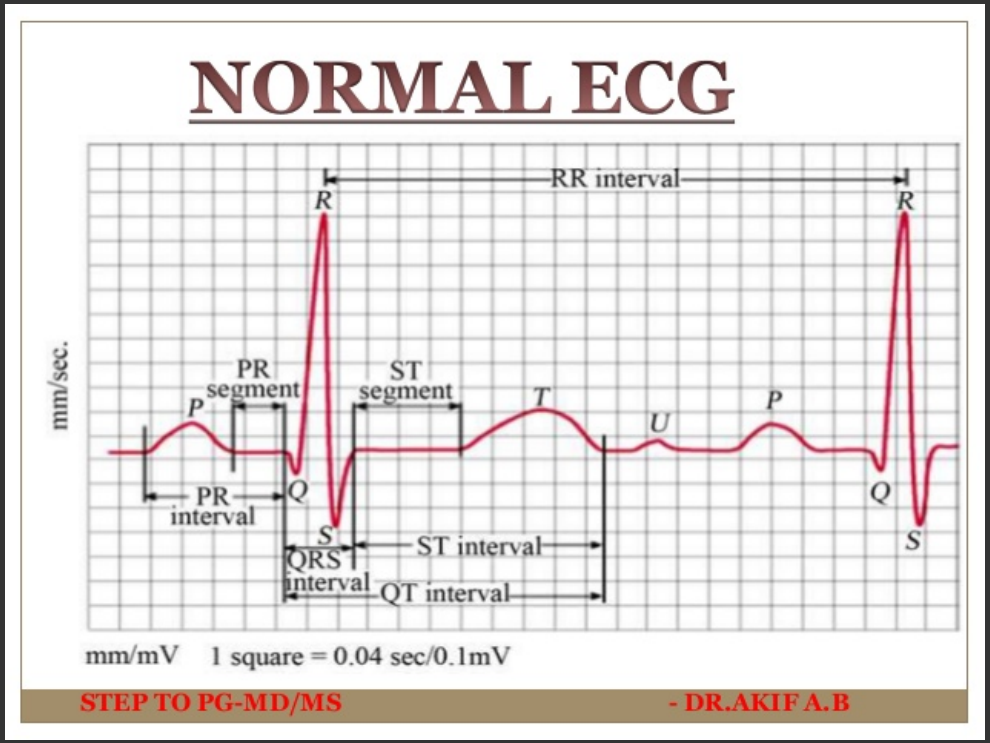
Exploratory Data Analysis (EDA)

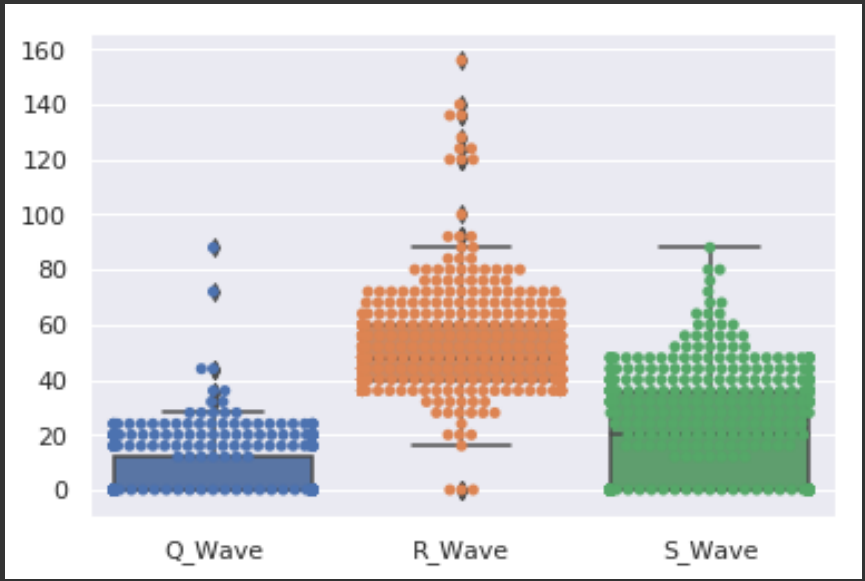


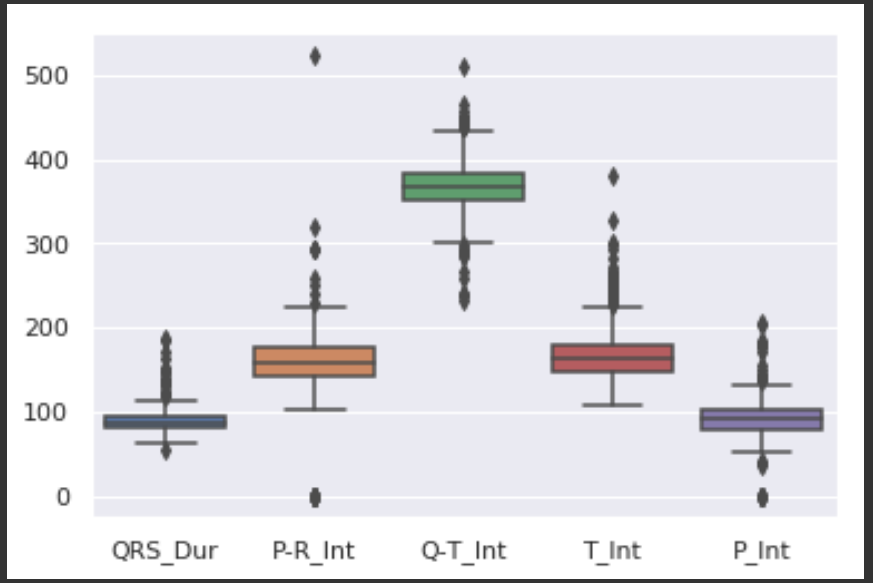
Exploratory Data Analysis (EDA)



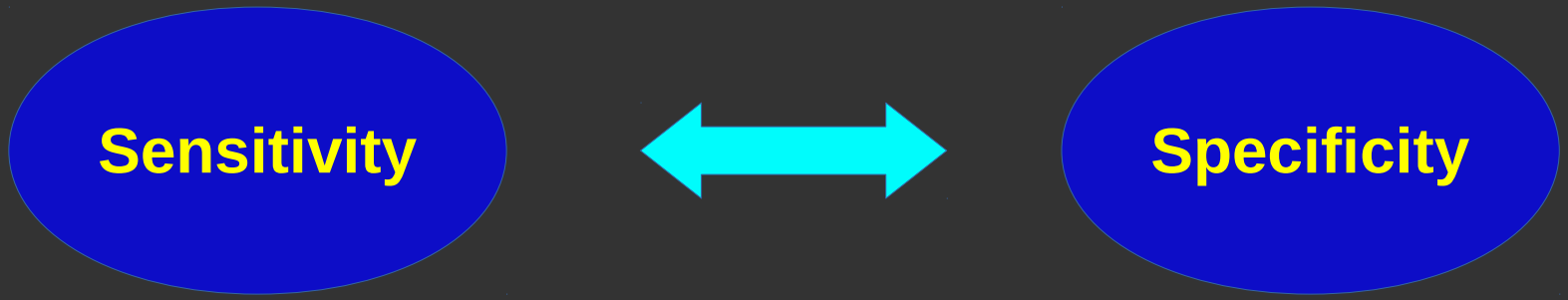
Exploratory Data Analysis (EDA)







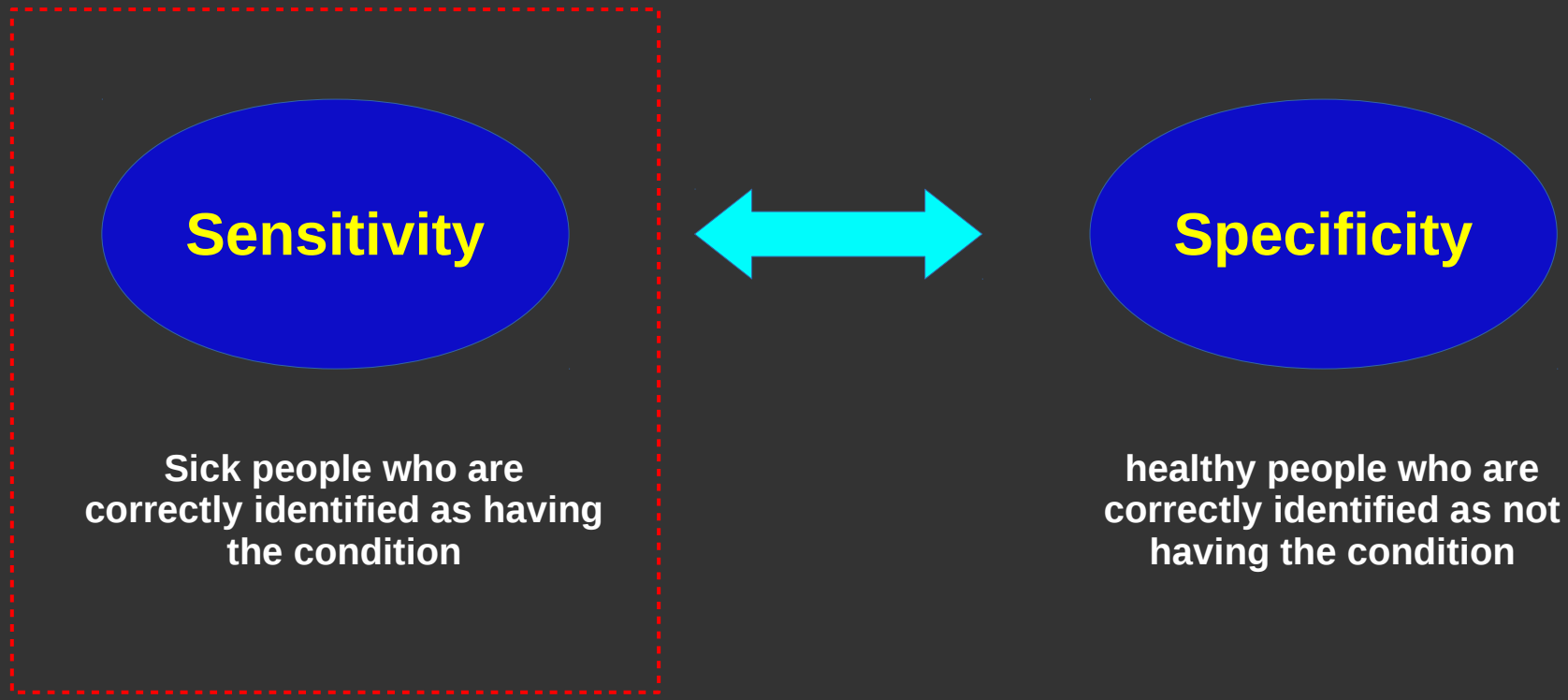
Machine Learning



Sick people who are correctly identified as having the condition

healthy people who are correctly identified as not having the condition

Machine Learning



$$\text{Recall} = \frac{tp}{tp + fn}$$

* Applied **"WEIGHTED RECALL"** as evaluation strategy due to importance of our predictions.

Machine Learning

Best Parameters

	Train Recall Score	Test Recall Score
KNN Clasification	0.669271	0.647059
Logistic Regression	0.841146	0.676471
Linear SVM	0.783854	0.720588
Kernelized SVM	0.976562	0.676471
Naive Bayes	0.760417	0.632353
Decision Tree	0.750000	0.661765
Random Forest	0.940104	0.750000
KNN Classification with PCA	0.677083	0.647059
Logistic Regression with PCA	0.825521	0.676471
Linear SVM with PCA	0.776042	0.735294
Kernalised SVM with PCA	0.968750	0.676471
Decision Trees with PCA	0.674479	0.573529
Random Forest with PCA	0.966146	0.632353

Conclusion

* Bagging and boosting methods generally raised the average training accuracy for the models but the test accuracy got reduced.

* PCA provided better results.

Conclusion

* Bagging and boosting methods generally raised the average training accuracy for the models but the test accuracy got reduced.

* PCA provided better results.

Conclusion

* Heavily biased data set:

  - 3 of 16 classes has no samples, some has pretty low instances (2,3,4,5)

  - preventing models to make predictions more accurate.

* Data set should include more samples for ML models to learn and make predictions.

Appendix

Project guidelines of David Yakobovitch

https://github.com/davidyakobovitch/project_guidelines

Thank you!

Muzaffer Estelik

Email : estelik.muzaffer@gmail.com
Linkedin : https://www.linkedin.com/in/muzaffer-estelik-42822b157/
Github : https://github.com/MuzafferEstelik