BMEG 3103 (AI)

PROJECT 2

AU, Wai Tak (1155175068)
CHAN, Cheuk Ka (1155174356)

OUR TEAM

AU Wai Tak

7.

CHAN Cheuk Ka

9.

INTRODUCTION

This project's task is to create a machine learning algorithm that can classify the health of patients into 5 categories.

01

DATA PRE-PROCESSING

+ data sanitisation

IRRELEVANT COLUMNS

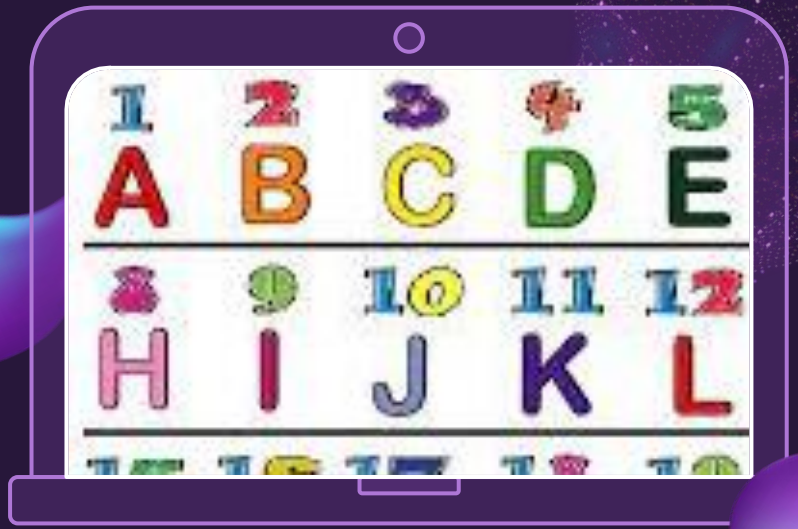
We dropped the first few useless columns pertaining username, timestamp, etc.

	A	B	C	D	E	
1		user_name	raw_timestamp_part_1	raw_timestamp_part_2	cvtd_timestamp	new
2	1	carlitos	1323084231	788290	5/12/201111:23	no
3	2	carlitos	1323084231	808298	5/12/201111:23	no
4	3	carlitos	1323084231	820366	5/12/201111:23	no
5	4	carlitos	1323084232	120339	5/12/201111:23	no
6	5	carlitos	1323084232	196328	5/12/201111:23	no
7	6	carlitos	1323084232	304277	5/12/201111:23	no
8	7	carlitos	1323084232	368296	5/12/201111:23	no
9	8	carlitos	1323084232	440390	5/12/201111:23	no
10	9	carlitos	1323084232	484323	5/12/201111:23	no
11	10	carlitos	1323084232	484434	5/12/201111:23	no
12	11	carlitos	1323084232	500302	5/12/201111:23	no
13	12	carlitos	1323084232	528316	5/12/201111:23	no
14	13	carlitos	1323084232	560359	5/12/201111:23	no
15	14	carlitos	1323084232	576390	5/12/201111:23	no
16	15	carlitos	1323084232	604281	5/12/201111:23	no

			NA	NA	
			NA	NA	
			NA	NA	
			NA	NA	
			NA	NA	
			NA	NA	
			NA	NA	
			NA	NA	
2.713152	#DIV/0!	#DIV/0!	-94.3	3	5.6
			NA	NA	
			NA	NA	
			NA	NA	

ERRONEOUS DATA

We trashed the columns containing empty cells and errors



DIGITISE CLASS

We changed class letters to numbers

SANITISED

After sanitisation, we acquired a 19622×55 CSV matrix (excluding labels).

project2_trainingset_sanitised.csv

dataset > project2_trainingset_sanitised.csv > data

Edit csv

```
belt_x,magnet_belt_y,magnet_belt_z,roll_arm,pitch_arm,yaw_arm,total_accel_arm,gyros_arm_x,gyros_arm_y,gyros_arm_z,accel_arm_x,
771,1,-0.27,0.58,0.26,1,-3,7,-571,242,-10,-147,7.17,-47.1,49,-0.98,-0.75,-0.92,241,-292,-302,411,-618,-251,5
237,5,-0.37,0.88,0.33,-1,46,25,-564,255,5,-148,8.74,-46.2,49,-0.42,0.51,-0.15,245,-310,-281,413,-639,-286,5
12,7,-0.45,1.25,0.61,7,61,35,-552,284,9,-147,10.1,-48.5,44,0.32,1.27,0.64,211,-270,-264,389,-653,-266,5
,-0.5,1.43,0.64,6,68,46,-540,309,-1,-146,10.6,-50.1,38,0.5,1.28,0.89,139,-233,-255,379,-654,-259,5
-0.72,2.18,0.46,-8,100,50,-496,374,7,-144,12.8,-55.5,34,0.4,1.8,0.61,72,-208,-256,306,-666,-177,5
,14,-0.87,2.5,0.13,-14,120,54,-427,440,38,-144,15.5,-59.9,34,0.16,2.12,0.51,57,-210,-246,243,-665,-137,5
-0.93,2.63,0.25,-20,129,48,-407,453,38,-144,17.3,-62.4,33,0.08,2.33,0.49,50,-210,-245,224,-672,-132,5
,14,-0.95,2.71,0.44,-12,125,43,-359,497,44,-145,19,-64.9,33,0.08,2.62,0.51,37,-217,-237,185,-673,-117,5
2,-7,116,42,-342,502,49,-145,21.1,-68,33,0.08,2.83,0.64,33,-224,-234,157,-672,-108,5
972912,25,-0.5,1.04,0.23,53,231,-66,-78,583,134,0,0,0,39,-0.61,4.56,1.12,-237,298,33,-703,382,803,5
838,24,-0.32,0.79,0.13,46,222,-62,-44,591,134,0,0,0,34,-0.67,4.61,1.28,-107,309,70,-693,517,890,5
25287,28,0.43,0.37,-0.28,84,253,-74,10,596,123,0,0,0,52,0.19,4.5,2.24,505,7,-463,839,905,5
469457,28,0.43,0.39,-0.38,82,252,-70,10,601,117,0,0,0,51,0.22,4.06,1.84,38,496,4,-408,852,878,5
8,27,0.58,0.02,-0.34,88,236,-75,15,590,135,0,0,0,52,0.51,3.15,1.36,64,501,0,-204,821,777,5
136895,27,0.66,-0.27,-0.44,78,241,-79,-8,589,163,0,0,0,48,0.4,2.79,1.1,81,461,-14,-92,765,700,5
608714,26,0.71,-0.43,-0.49,73,237,-77,-17,589,146,0,0,0,47,0.61,2.76,1.07,89,450,-9,-68,752,684,5
899,26,0.8,-0.83,-0.69,64,243,-60,-54,583,166,0,0,0,44,0.71,2.1,1.26,66,425,-14,2,712,630,5
39,23,1.04,-1.38,-0.59,43,214,-73,-141,553,180,0,0,0,41,1.09,1.35,1.64,53,393,-24,67,653,537,5
06,23,1.11,-1.33,-0.71,30,215,-73,-177,540,172,0,0,0,40,1.2,1.3,1.76,38,391,-35,89,633,512,5
3196,23,1.17,-1.27,-0.79,17,213,-75,-189,535,189,0,0,0,41,1.27,1.28,1.82,40,394,-44,91,629,488,5
58974,22,1.32,-1.32,-0.89,11,200,-80,-219,523,185,0,0,0,42,1.33,1.14,1.84,52,401,-53,109,613,460,5
90316,20,0.51,-0.5,-0.72,-63,99,-154,-568,257,36,0,0,0,39,1.4,-3.1,-0.59,-201,323,5,-706,273,620,5
3207178,20,0.48,-0.43,-0.79,-58,99,-164,-567,241,44,0,0,0,37,1.38,-2.94,-0.52,-193,307,8,-699,229,596,5
.7372696,20,0.56,-0.42,-0.79,-49,92,-165,-574,240,29,0,0,0,36,1.38,-2.94,-0.48,-193,291,3,-683,152,555,5
.5303382,20,0.66,-0.39,-0.72,-52,87,-167,-576,232,42,0,0,0,34,1.46,-3.08,-0.49,-203,266,-4,-672,119,540,5
.3970337,18,0.61,-0.32,-0.49,-55,77,-152,-586,220,26,0,0,0,33,1.38,-2.86,-0.56,-227,236,-1,-622,-11,480,5
.0376856,19,0.55,-0.37,-0.34,-53,67,-160,-595,211,24,0,0,0,32,1.36,-2.83,-0.61,-227,221,0,-592,-85,431,5
3.9202735,19,0.58,-0.48,-0.39,-53,57,-169,-591,210,12,0,0,0,31,1.33,-2.89,-0.71,-230,200,-8,-571,-126,404,5
16.9489792,19,0.43,-0.4,-0.97,-45,58,-173,-609,165,1,0,0,0,25,1.51,-2.18,-1.02,-241,33,-17,-377,-420,173,5
130785,19,0.34,-0.31,-0.51,-42,70,-167,-624,127,8,0,0,0,27,1.75,-1.91,-0.38,-255,-50,-30,-226,-570,27,5
.4998335,19,0.32,-0.26,-0.36,-42,66,-168,-618,134,0,0,0,0,29,1.73,-1.75,-0.25,-271,-68,-37,-205,-587,6,5
14.5256424,18,0.24,-0.24,0.05,-41,62,-164,-618,116,7,0,0,0,29,1.59,-1.36,0,-271,-91,-43,-151,-635,-36,5
18037,19,0.22,-0.27,0.21,-38,54,-170,-621,113,-9,0,0,0,29,1.54,-1.2,0.05,-263,-99,-45,-116,-651,-70,5
458728,19,0.13,-0.14,0.34,-40,42,-176,-628,116,0,0,0,0,32,1.48,-0.9,0.05,-270,-141,-98,5
,0.36,-36,38,-176,-627,119,2,0,0,0,33,1.38,-0.64,0.08,-278,-159,-52,-60,-686,-110,5
```

Col 55: class

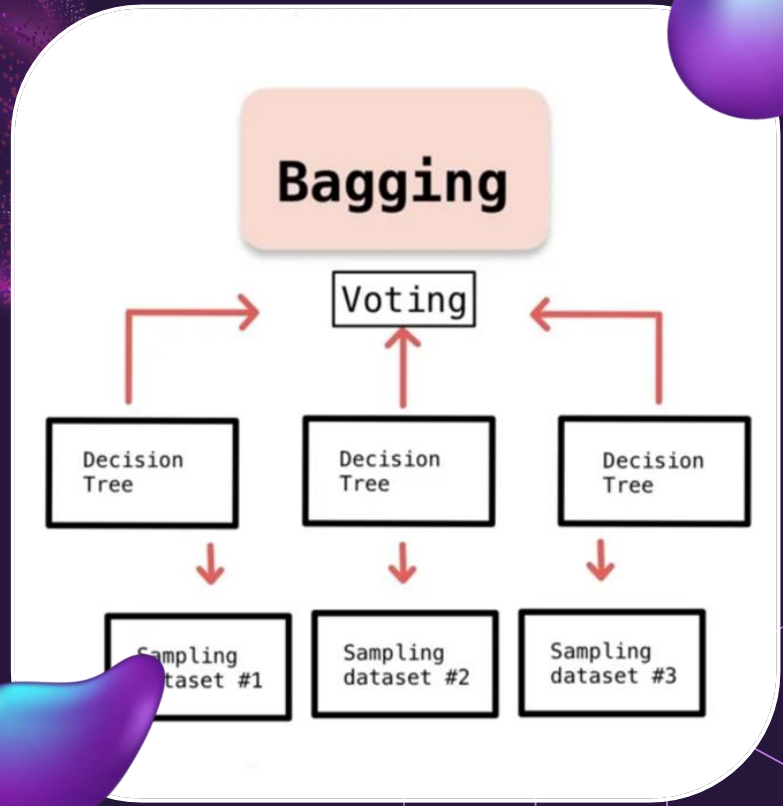
02

TRAINING AI MODEL

+ hyperparameters

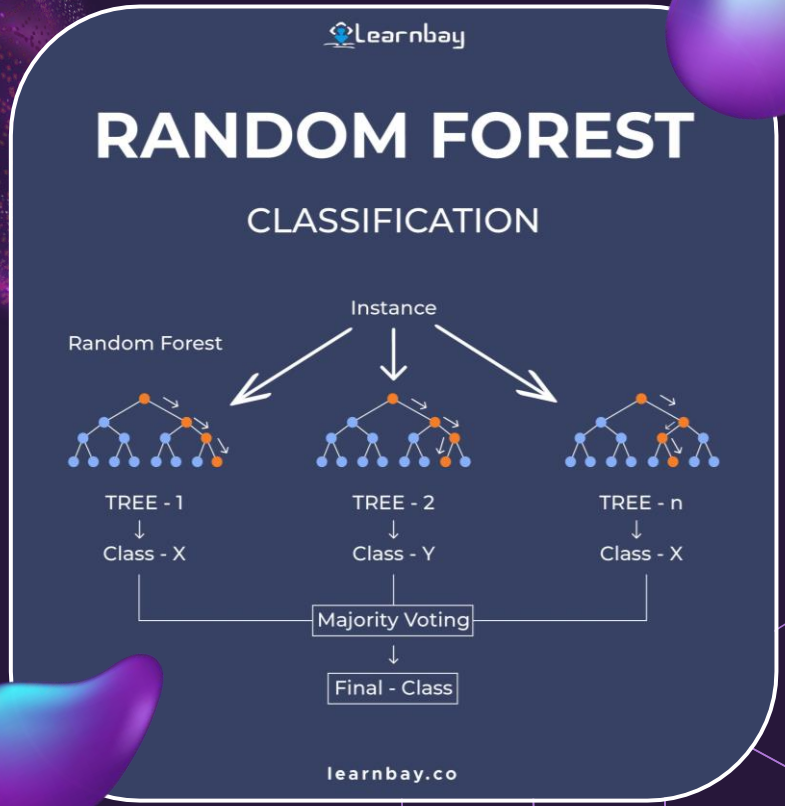
MODEL FAMILY

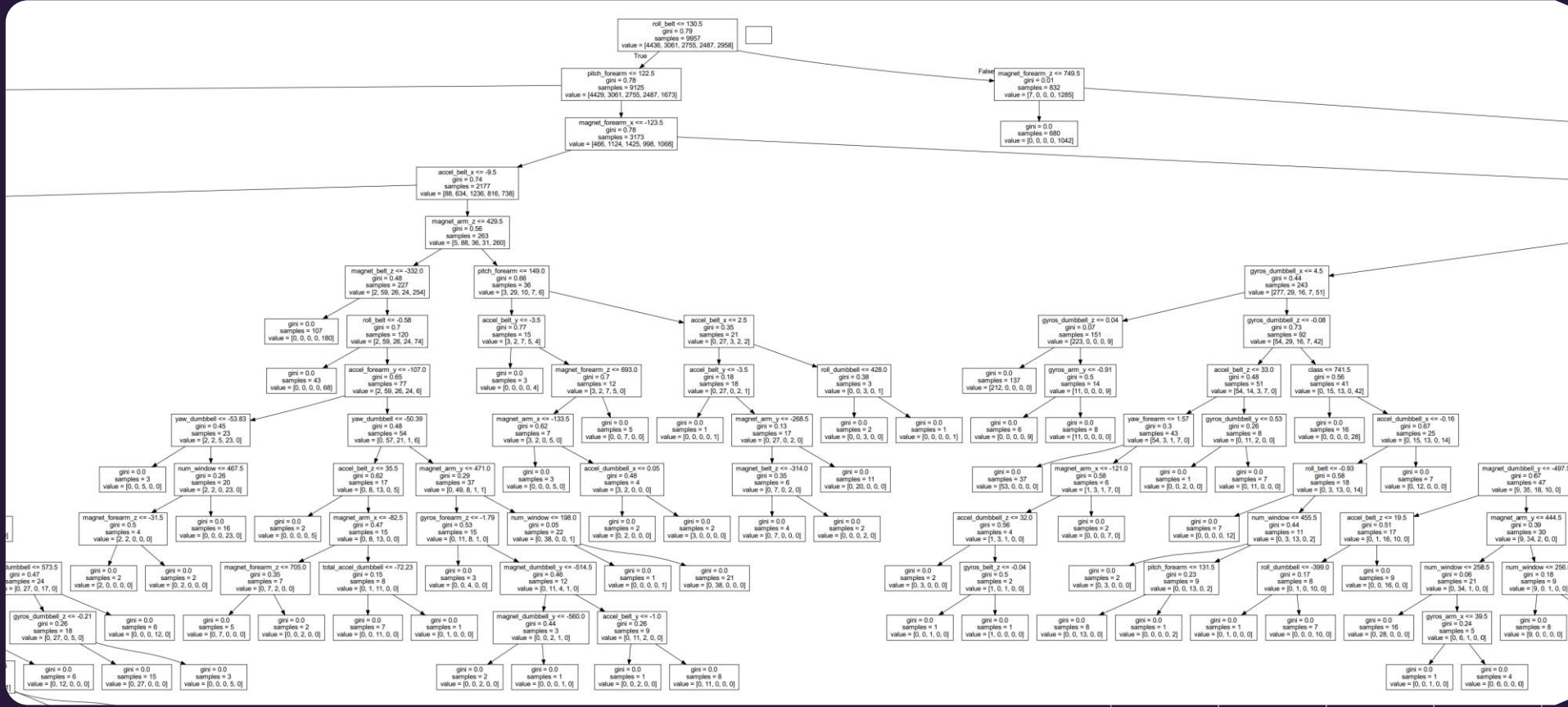
Ensemble learning (bagging)



RANDOM FOREST CLASSIFIER

From sklearn, Random Forest consists of fully grown trees built on bootstrapped data and the majority vote rule to make predictions.




roll_belt <= 130.5
gini = 0.79
samples = 9957
value = [4436, 3061, 2755, 2487, 2958]
True
False
pitch_forearm <= 122.5
gini = 0.78
samples = 9125
value = [4429, 3061, 2755, 2487, 1673]
magnet_forearm_z <= 749.5
gini = 0.01
samples = 832
value = [7, 0, 0, 0, 1285]
magnet_forearm_x <= -123.5
gini = 0.78
samples = 3173
value = [466, 1124, 1425, 998, 1068]
gini = 0.0
samples = 680
value = [0, 0, 0, 0, 1942]
accel_belt_x <= -9.5
gini = 0.74
samples = 2177
value = [88, 634, 1236, 816, 738]
magnet_arm_z <= 429.5
gini = 0.56
samples = 263
value = [5, 88, 36, 31, 260]
magnet_belt_z <= -332.0
gini = 0.48
samples = 227
value = [2, 59, 26, 24, 254]
pitch_forearm <= 149.0
gini = 0.66
samples = 36
value = [3, 29, 10, 7, 6]
gyros_dumbbell_x <= 4.5
gini = 0.44
samples = 243
value = [277, 29, 16, 7, 51]
gini = 0.0
samples = 107
value = [0, 0, 0, 0, 180]
roll_belt <= -0.58
gini = 0.7
samples = 120
value = [2, 59, 26, 24, 74]
accel_belt_y <= -3.5
gini = 0.77
samples = 15
value = [3, 2, 7, 5, 4]
accel_belt_x <= 2.5
gini = 0.35
samples = 21
value = [0, 27, 3, 2, 2]
gyros_dumbbell_z <= 0.04
gini = 0.07
samples = 151
value = [223, 0, 0, 0, 9]
gyros_dumbbell_z <= -0.08
gini = 0.73
samples = 92
value = [54, 29, 16, 7, 42]
gini = 0.0
samples = 43
value = [0, 0, 0, 0, 68]
accel_forearm_y <= -107.0
gini = 0.65
samples = 77
value = [2, 59, 26, 24, 6]
gini = 0.0
samples = 3
value = [0, 0, 0, 0, 4]
magnet_forearm_z <= 693.0
gini = 0.7
samples = 12
value = [3, 2, 7, 5, 0]
accel_belt_y <= -3.5
gini = 0.18
samples = 18
value = [0, 27, 0, 2, 1]
roll_dumbbell <= 428.0
gini = 0.38
samples = 3
value = [0, 0, 3, 0, 1]
gini = 0.0
samples = 137
value = [212, 0, 0, 0, 0]
gyros_arm_y <= -0.91
gini = 0.5
samples = 14
value = [11, 0, 0, 0, 9]
accel_belt_z <= 33.0
gini = 0.48
samples = 51
value = [54, 14, 3, 7, 0]
class <= 741.5
gini = 0.56
samples = 41
value = [0, 15, 13, 0, 42]
yaw_dumbbell <= -53.83
gini = 0.45
samples = 23
value = [2, 2, 5, 23, 0]
yaw_dumbbell <= -50.39
gini = 0.48
samples = 54
value = [0, 57, 21, 1, 6]
magnet_arm_x <= -133.5
gini = 0.62
samples = 7
value = [3, 2, 0, 5, 0]
gini = 0.0
samples = 5
value = [0, 0, 7, 0, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 1]
magnet_arm_y <= -268.5
gini = 0.13
samples = 17
value = [0, 27, 0, 2, 0]
gini = 0.0
samples = 2
value = [0, 0, 3, 0, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 1]
gini = 0.0
samples = 6
value = [0, 0, 0, 0, 6]
gyros_dumbbell_z <= 0.53
gini = 0.26
samples = 8
value = [0, 11, 2, 0, 0]
gini = 0.0
samples = 16
value = [0, 0, 0, 0, 28]
accel_dumbbell_x <= -0.16
gini = 0.67
samples = 25
value = [0, 15, 13, 0, 14]
gini = 0.0
samples = 1
value = [0, 0, 2, 0, 0]
num_window <= 467.5
gini = 0.26
samples = 20
value = [2, 2, 0, 23, 0]
accel_belt_z <= 35.5
gini = 0.62
samples = 17
value = [0, 8, 13, 0, 5]
magnet_arm_y <= 471.0
gini = 0.29
samples = 37
value = [0, 49, 8, 1, 1]
gini = 0.0
samples = 3
value = [0, 0, 0, 5, 0]
accel_dumbbell_x <= 0.05
gini = 0.48
samples = 4
value = [3, 2, 0, 0, 0]
magnet_belt_z <= -314.0
gini = 0.35
samples = 6
value = [0, 7, 0, 2, 0]
gini = 0.0
samples = 11
value = [0, 20, 0, 0, 0]
gini = 0.0
samples = 37
value = [53, 0, 0, 0, 0]
magnet_arm_x <= -121.0
gini = 0.58
samples = 6
value = [1, 3, 1, 7, 0]
gini = 0.0
samples = 1
value = [0, 0, 2, 0, 0]
gini = 0.0
samples = 7
value = [0, 11, 0, 0, 0]
roll_belt <= -0.93
gini = 0.58
samples = 18
value = [0, 3, 13, 0, 14]
gini = 0.0
samples = 7
value = [0, 12, 0, 0, 0]
magnet_dumbbell_y <= -497.5
gini = 0.67
samples = 47
value = [0, 35, 16, 10, 0]
magnet_forearm_z <= -31.5
gini = 0.5
samples = 4
value = [2, 2, 0, 0, 0]
gini = 0.0
samples = 16
value = [0, 0, 0, 0, 23]
gini = 0.0
samples = 2
value = [0, 0, 0, 0, 5]
magnet_arm_x <= -82.5
gini = 0.47
samples = 15
value = [0, 8, 13, 0, 0]
gyros_forearm_z <= -1.79
gini = 0.53
samples = 15
value = [0, 11, 8, 1, 0]
num_window <= 198.0
gini = 0.05
samples = 22
value = [0, 38, 0, 0, 1]
gini = 0.0
samples = 2
value = [0, 2, 0, 0, 0]
gini = 0.0
samples = 2
value = [3, 0, 0, 0, 0]
gini = 0.0
samples = 4
value = [0, 7, 0, 0, 0]
gini = 0.0
samples = 2
value = [0, 0, 0, 2, 0]
accel_dumbbell_z <= 32.0
gini = 0.56
samples = 4
value = [1, 3, 1, 0, 0]
gini = 0.0
samples = 2
value = [0, 0, 0, 0, 7]
gini = 0.0
samples = 7
value = [0, 0, 0, 0, 12]
num_window <= 455.5
gini = 0.44
samples = 11
value = [0, 3, 13, 0, 2]
accel_belt_z <= 19.5
gini = 0.51
samples = 17
value = [0, 1, 16, 10, 0]
magnet_arm_y <= 444.5
gini = 0.39
samples = 30
value = [9, 34, 2, 0, 0]
gini = 0.0
samples = 2
value = [2, 0, 0, 0, 0]
gini = 0.0
samples = 2
value = [0, 2, 0, 0, 0]
magnet_forearm_z <= 705.0
gini = 0.35
samples = 7
value = [0, 7, 2, 0, 0]
total_accel_dumbbell <= -72.23
gini = 0.15
samples = 8
value = [0, 1, 11, 0, 0]
gini = 0.0
samples = 3
value = [0, 0, 4, 0, 0]
magnet_dumbbell_y <= -514.5
gini = 0.46
samples = 12
value = [0, 11, 4, 1, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 1]
gini = 0.0
samples = 21
value = [0, 38, 0, 0, 0]
gini = 0.0
samples = 3
value = [0, 3, 0, 0, 0]
gyros_belt_z <= -0.04
gini = 0.5
samples = 2
value = [1, 0, 1, 0, 0]
gini = 0.0
samples = 2
value = [0, 3, 0, 0, 0]
pitch_forearm <= 131.5
gini = 0.23
samples = 9
value = [0, 0, 13, 0, 2]
roll_dumbbell <= -399.0
gini = 0.17
samples = 8
value = [0, 1, 0, 10, 0]
gini = 0.0
samples = 9
value = [0, 0, 16, 0, 0]
num_window <= 258.5
gini = 0.06
samples = 21
value = [0, 34, 1, 0, 0]
num_window <= 256.
gini = 0.18
samples = 9
value = [9, 0, 1, 0, 0]
gyros_dumbbell_z <= -0.21
gini = 0.26
samples = 18
value = [0, 27, 0, 5, 0]
gini = 0.0
samples = 6
value = [0, 0, 0, 12, 0]
gini = 0.0
samples = 5
value = [0, 7, 0, 0, 0]
gini = 0.0
samples = 2
value = [0, 0, 2, 0, 0]
gini = 0.0
samples = 7
value = [0, 0, 11, 0, 0]
gini = 0.0
samples = 1
value = [0, 1, 0, 0, 0]
magnet_dumbbell_y <= -560.0
gini = 0.44
samples = 3
value = [0, 0, 2, 1, 0]
accel_belt_y <= -1.0
gini = 0.26
samples = 9
value = [0, 11, 2, 0, 0]
gini = 0.0
samples = 1
value = [0, 1, 0, 0, 0]
gini = 0.0
samples = 1
value = [1, 0, 0, 0, 0]
gini = 0.0
samples = 8
value = [0, 0, 13, 0, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 0, 2]
gini = 0.0
samples = 1
value = [0, 1, 0, 0, 0]
gini = 0.0
samples = 7
value = [0, 0, 0, 10, 0]
gini = 0.0
samples = 16
value = [0, 28, 0, 0, 0]
gyros_arm_x <= 39.5
gini = 0.24
samples = 5
value = [0, 6, 1, 0, 0]
gini = 0.0
samples = 8
value = [9, 0, 0, 0, 0]
gini = 0.0
samples = 6
value = [0, 12, 0, 0, 0]
gini = 0.0
samples = 15
value = [0, 27, 0, 0, 0]
gini = 0.0
samples = 3
value = [0, 0, 0, 5, 0]
gini = 0.0
samples = 2
value = [0, 0, 2, 0, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 1, 0]
gini = 0.0
samples = 1
value = [0, 0, 2, 0, 0]
gini = 0.0
samples = 8
value = [0, 11, 0, 0, 0]
gini = 0.0
samples = 1
value = [0, 0, 0, 1, 0]
gini = 0.0
samples = 4
value = [0, 6, 0, 0, 0]

HYPERPARAMETERS

SPLIT

We split 20% of the training data for testing

ESTIMATORS

We started from *n_estimators = 100*, then incremented by *50* until it reached >95% accuracy at *200*

METRICS

on the 20% testing data, after hyperparameters tuning

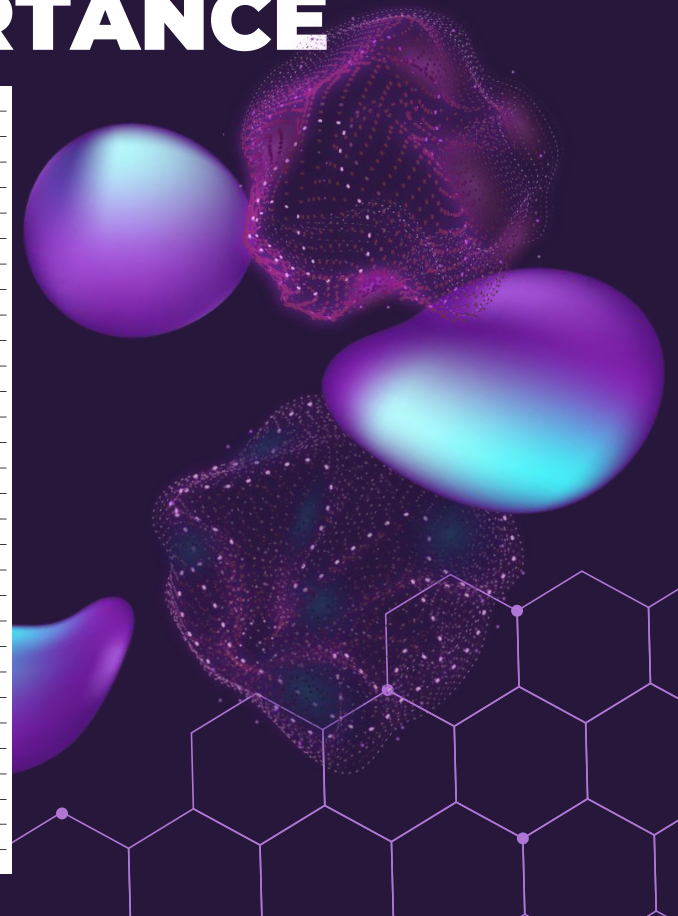
ACCURACY	0.9982165605095541
PRECISION	0.9982171198022268
RECALL	0.9982165605095541
CONFUSION MATRIX	1097 0 0 0 0 2 731 0 0 0 0 1 685 0 0 0 0 2 668 1 0 0 0 1 737

The metrics are already very good, so we didn't further tune the hyperparameters. ¯_(ツ)_/¯

FEATURE IMPORTANCE

Feature label	Importance
num_window	0.092442
roll_belt	0.074710
yaw_belt	0.051930
magnet_dumbbell_z	0.045757
pitch_forearm	0.045721
magnet_dumbbell_y	0.043380
pitch_belt	0.041529
roll_forearm	0.035012
magnet_dumbbell_x	0.030000
roll_dumbbell	0.026281
magnet_belt_y	0.025142
accel_dumbbell_y	0.024722
accel_belt_z	0.024186
magnet_belt_z	0.022496
accel_dumbbell_z	0.020494
accel_forearm_x	0.019814
roll_arm	0.017512
magnet_belt_x	0.016070
yaw_dumbbell	0.015871
magnet_arm_x	0.015695
total_accel_dumbbell	0.015565
gyros_belt_z	0.015538
magnet_forearm_z	0.015358
accel_dumbbell_x	0.014625

Feature label	Importance
gyros_dumbbell_y	0.014411
accel_forearm_z	0.014018
accel_arm_x	0.013997
total_accel_belt	0.013754
yaw_arm	0.012924
magnet_arm_y	0.012513
magnet_forearm_x	0.011880
magnet_forearm_y	0.011529
pitch_dumbbell	0.010522
pitch_arm	0.010240
yaw_forearm	0.009970
magnet_arm_z	0.009346
accel_arm_y	0.008292
accel_belt_y	0.008053
accel_forearm_y	0.007447
gyros_arm_y	0.007316
gyros_arm_x	0.007274
gyros_belt_y	0.007097
accel_belt_x	0.006923
gyros_dumbbell_x	0.006735
accel_arm_z	0.006690
gyros_forearm_y	0.006485
total_accel_arm	0.005695
total_accel_forearm	0.005571
gyros_belt_x	0.005471
gyros_dumbbell_z	0.004686
gyros_forearm_z	0.004218
gyros_forearm_x	0.003909
gyros_arm_z	0.003128
new_window	0.000053

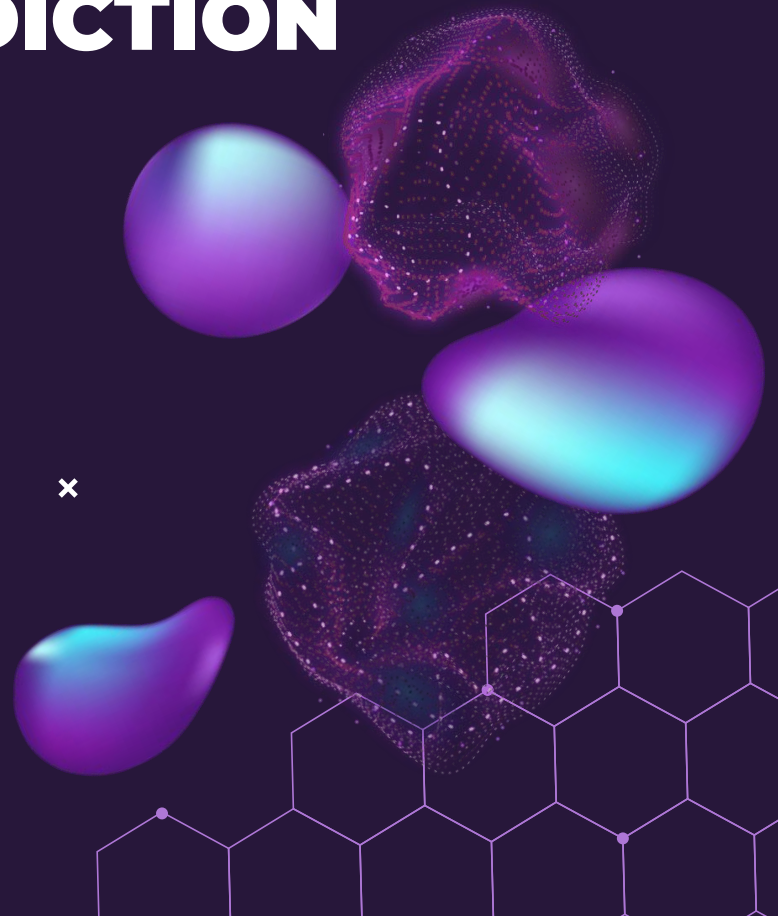


CLASS PREDICTION

The unseen data was pre-processed the same manner as described before.

Prediction results:
[A B D B B E D E A B A A C A D B C E D B]

* We had to change the 1/2/3/4/5 back into A/B/C/D/E due to the pre-processing steps



FIN.