Μηχανική Μάθηση

DETECTION OF PARKINSON'S DISEASE

Σκοπός

- Ανίχνευση της νόσου Parkinson από ηχογραφημένες φωνές
- Κατασκευή μοντέλου ανίχνευσης με τη δυνατότερη ακρίβεια
  1. Χρήση και σύγκριση 5 διαφορετικών αλγορίθμων ταξινόμησης
  2. Χρήση Ensemble τεχνικών
     - Boosting
     - Bagging

Δεδομένα

- Προέρχονται από UCI Machine Learning Repository

- Συνολικά :
  - 195 δείγματα
  - 24 χαρακτηριστικά

- Οι μετρήσεις έγιναν μεταξύ 31 ανθρώπων εκ των οποίων οι 23 είχαν τη νόσο Parkinson.

Δεδομένα

- 24 χαρακτηριστικά ειδικών μετρήσεων ως προς τη συχνότητα
- Status :
  - 0 για μη ασθενής
  - 1 για ασθενής

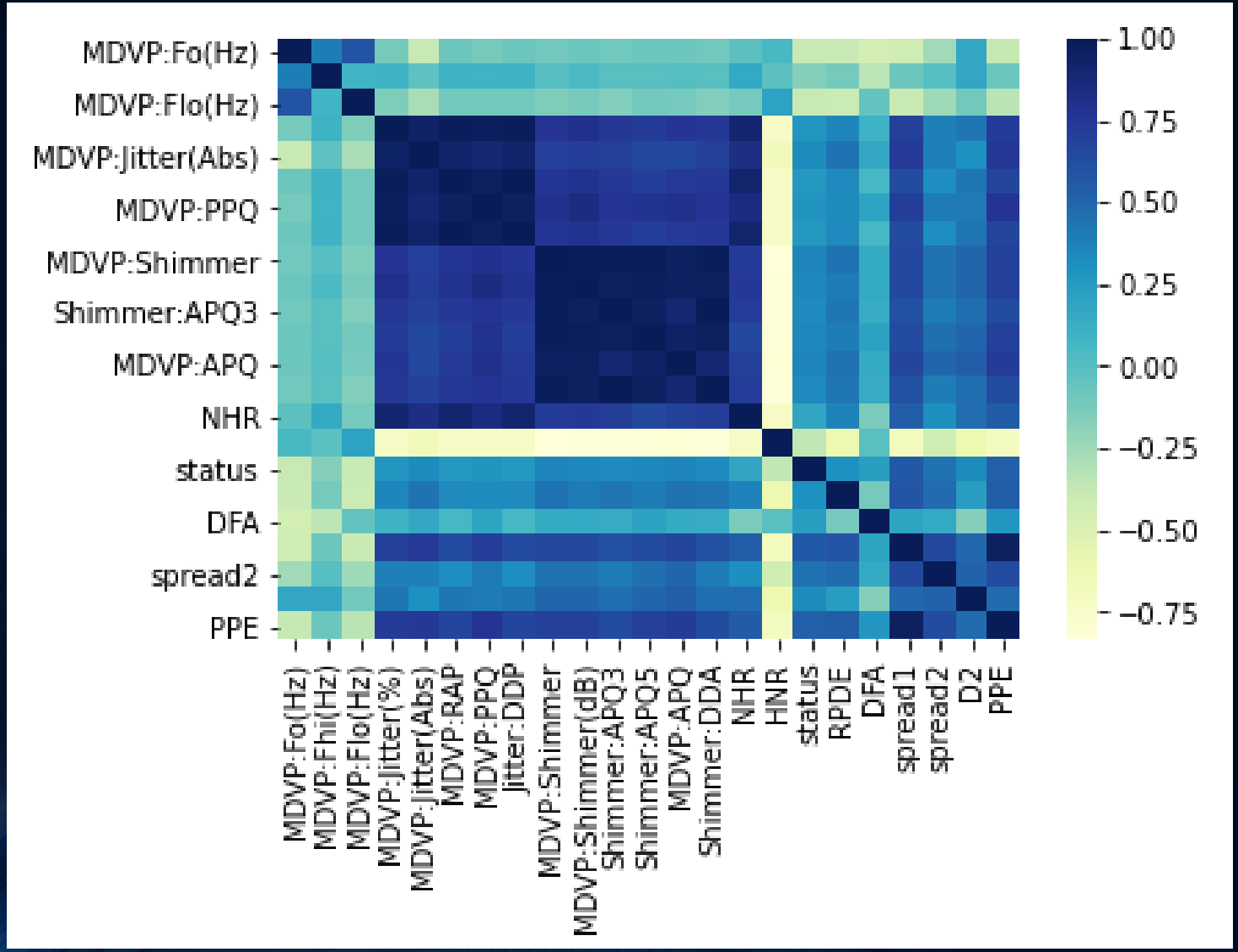
```
name - ASCII subject name and recording number
MDVP:Fo(Hz) - Average vocal fundamental frequency
MDVP:Fhi(Hz) - Maximum vocal fundamental frequency
MDVP:Flo(Hz) - Minimum vocal fundamental frequency

MDVP:Jitter(%),
MDVP:Jitter(Abs),
MDVP:RAP,           Several measures of variation in fundamental frequency
MDVP:PPQ,
Jitter:DDP

MDVP:Shimmer,
MDVP:Shimmer(dB),
Shimmer:APQ3,       Several measures of variation in amplitude
Shimmer:APQ5,
MDVP:APQ,
Shimmer:DDA

NHR,
HNR       Two measures of ratio of noise to tonal components in the voice

status - Health status of the subject , (one) - Parkinson's, (zero) - healthy

RPDE,
D2        Two nonlinear dynamical complexity measures

DFA       Signal fractal scaling exponent

spread1,
spread2,   Three nonlinear measures of fundamental frequency variation
PPE
```

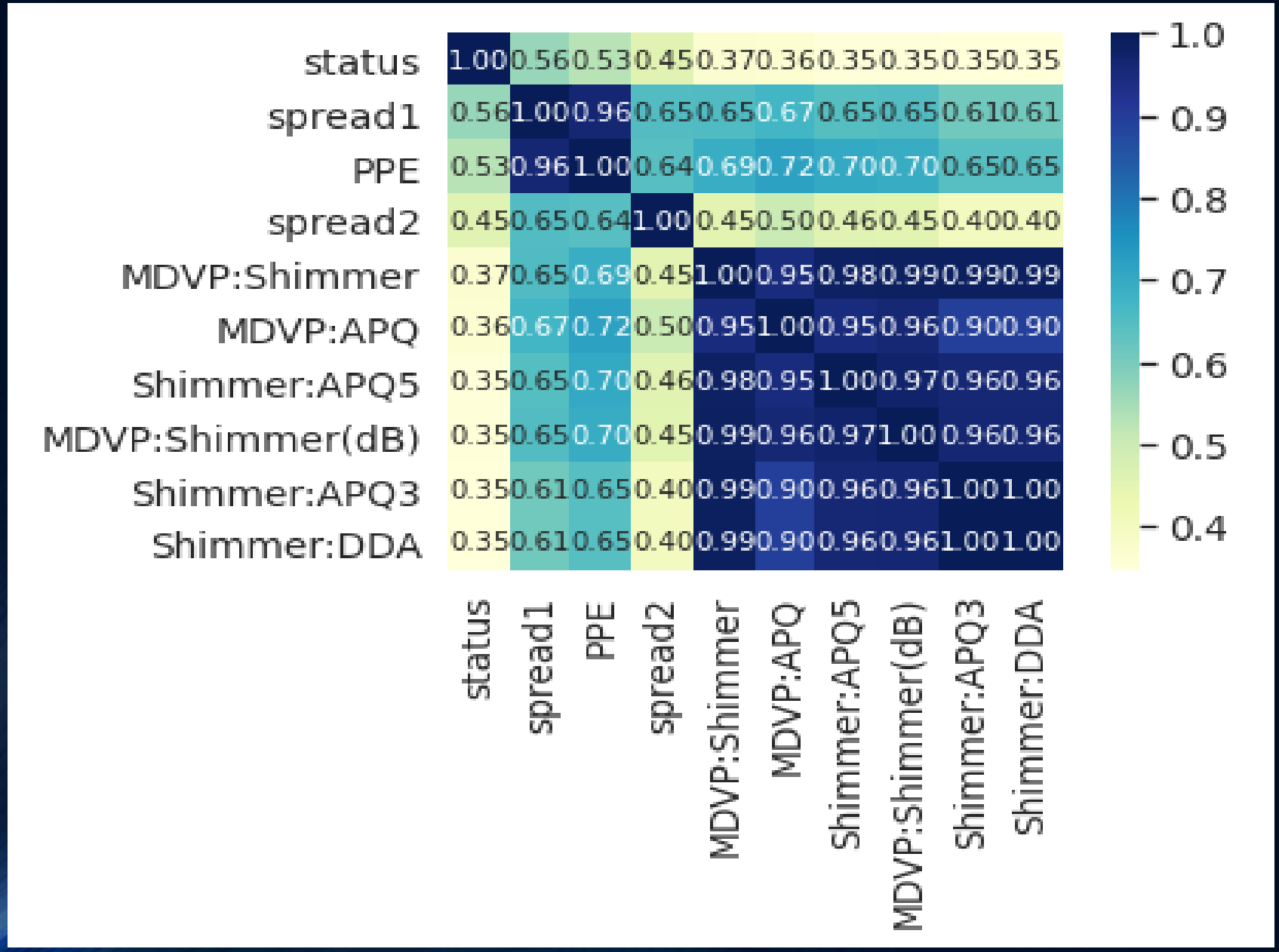
Pre-processing δεδομένων

- Descriptive Statistics

	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP
count	195.000000	195.000000	195.000000	195.000000	195.000000	195.000000
mean	154.228641	197.104918	116.324631	0.006220	0.000044	0.003306
std	41.390065	91.491548	43.521413	0.004848	0.000035	0.002968
min	88.333000	102.145000	65.476000	0.001680	0.000007	0.000680
25%	117.572000	134.862500	84.291000	0.003460	0.000020	0.001660
50%	148.790000	175.829000	104.315000	0.004940	0.000030	0.002500
75%	182.769000	224.205500	140.018500	0.007365	0.000060	0.003835
max	260.105000	592.030000	239.170000	0.033160	0.000260	0.021440

Συντελεστής συσχέτισης μεταβλητών




status 1.00 0.56 0.53 0.45 0.37 0.36 0.35 0.35 0.35 0.35
spread1 0.56 1.00 0.96 0.65 0.65 0.67 0.65 0.65 0.61 0.61
PPE 0.53 0.96 1.00 0.64 0.69 0.72 0.70 0.70 0.65 0.65
spread2 0.45 0.65 0.64 1.00 0.45 0.50 0.46 0.45 0.40 0.40
MDVP:Shimmer 0.37 0.65 0.69 0.45 1.00 0.95 0.98 0.99 0.99 0.99
MDVP:APQ 0.36 0.67 0.72 0.50 0.95 1.00 0.95 0.96 0.90 0.90
Shimmer:APQ5 0.35 0.65 0.70 0.46 0.98 0.95 1.00 0.97 0.96 0.96
MDVP:Shimmer(dB) 0.35 0.65 0.70 0.45 0.99 0.96 0.97 1.00 0.96 0.96
Shimmer:APQ3 0.35 0.61 0.65 0.40 0.99 0.90 0.96 0.96 1.00 1.00
Shimmer:DDA 0.35 0.61 0.65 0.40 0.99 0.90 0.96 0.96 1.00 1.00
status
spread1
PPE
spread2
MDVP:Shimmer
MDVP:APQ
Shimmer:APQ5
MDVP:Shimmer(dB)
Shimmer:APQ3
Shimmer:DDA
1.0
0.9
0.8
0.7
0.6
0.5
0.4

Feature Split

- Διαχωρισμός των δεδομένων σε χαρακτηριστικά εισόδου και εξόδου

- Διαχωρισμός των δεδομένων σε training και test σύνολα
  - Train: 80%
  - Test: 20%

Σύγκριση 5 διαφορετικών αλγορίθμων χωρίς feature scaling

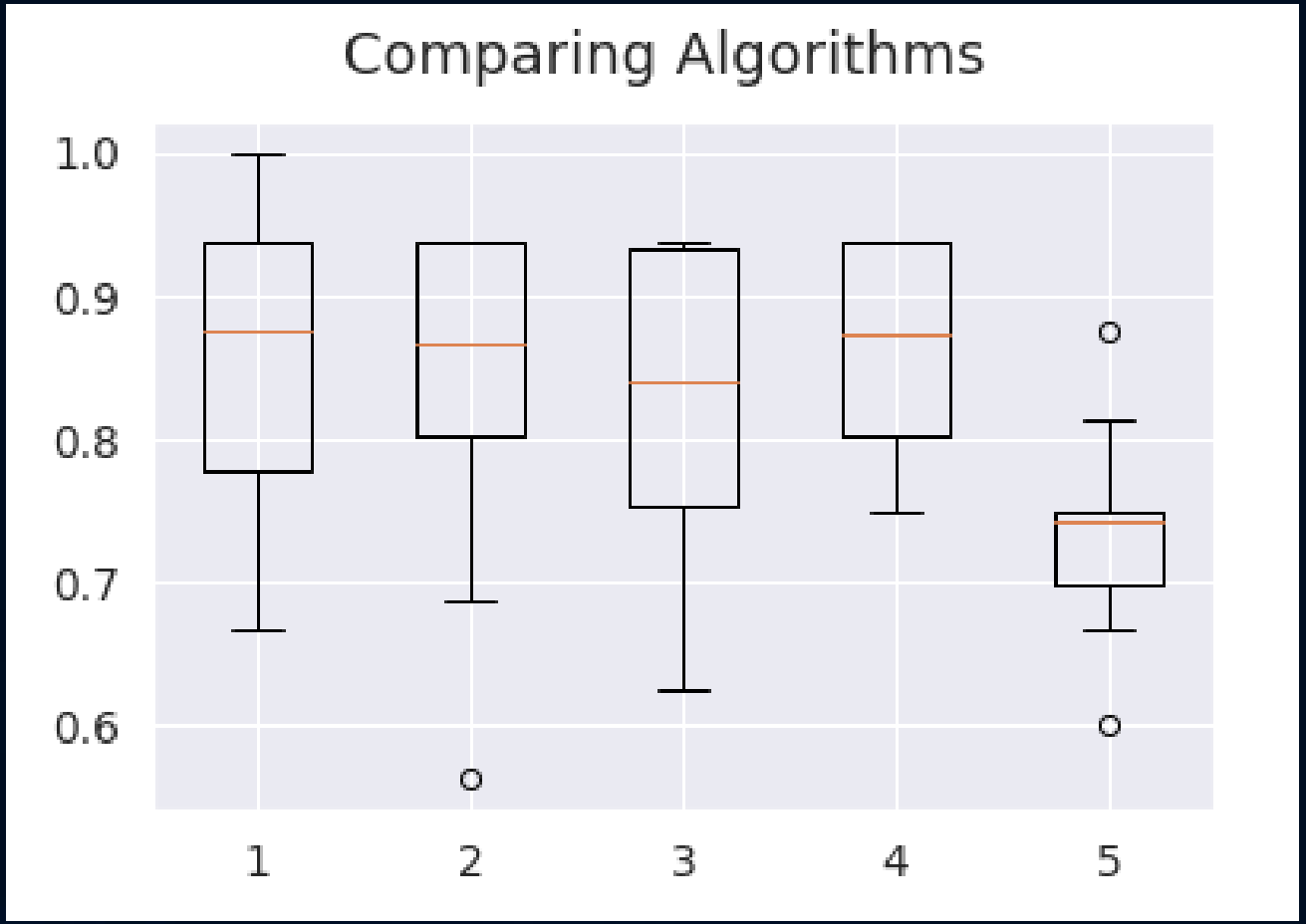
1. Logistic Regression : 0.859583 (0.114429)

2. knn : 0.834167 (0.118714)

3. SVC : 0.821667 (0.117951)

4. decision tree: 0.865417 (0.072314)

5. Naive Bayes : 0.735833 (0.071715)



Σύγκριση 5 διαφορετικών αλγορίθμων με feature scaling

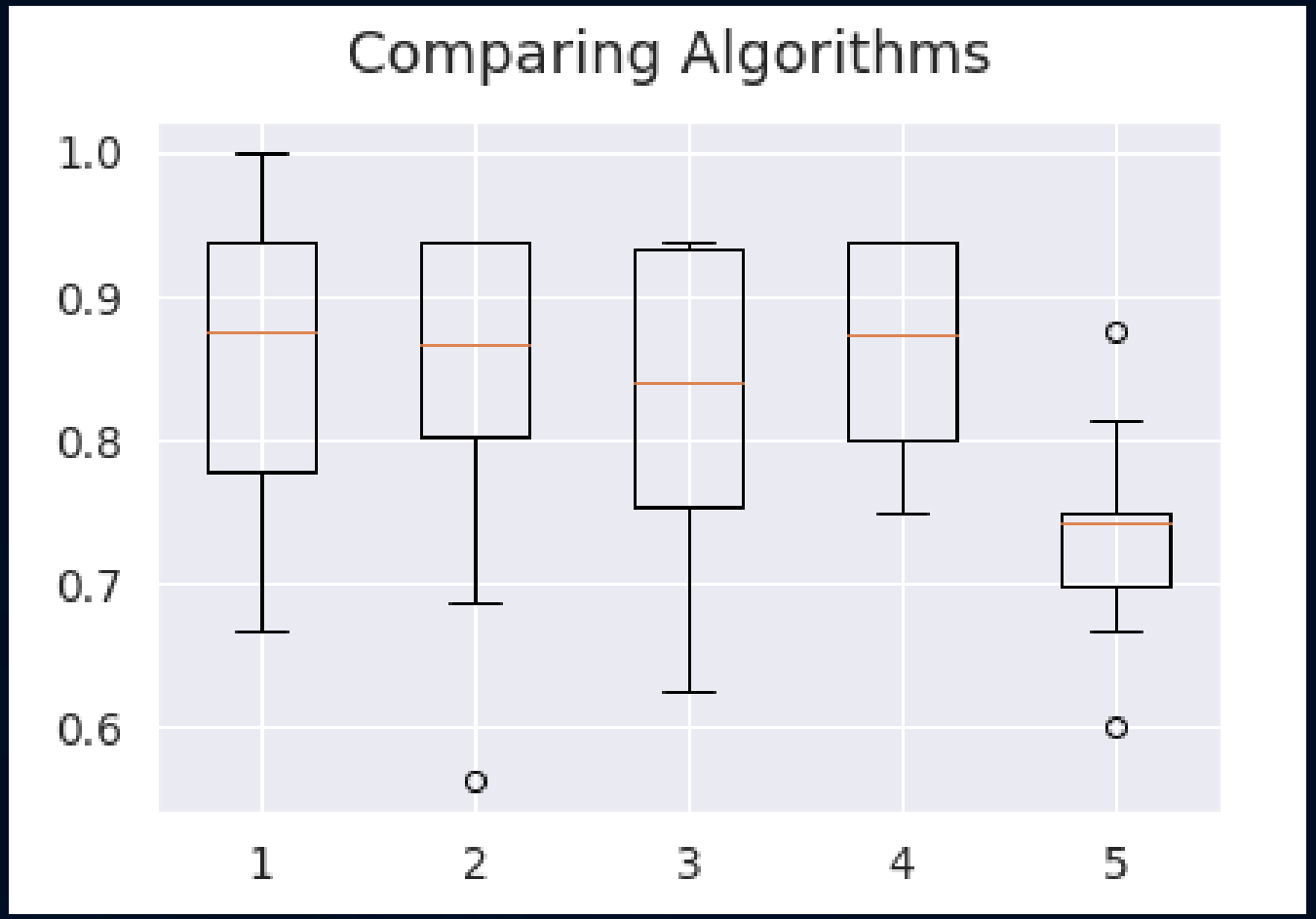
1. Logistic Regression : 0.859583 (0.114429)

2. knn : 0.834167 (0.118714)

3. SVC : 0.821667 (0.117951)

4. decision tree : 0.859167 (0.079009)

5. Naive Bayes : 0.735833 (0.071715)



Regularization Tuning για τους 2 καλύτερους αλγορίθμους

- Decision Tree Tunning αλγόριθμος
  - Best: 0.878750 using {}
- Logistic Regression Tuning αλγόριθμος
  - Best: 0.859167 using {'C': 0.7}

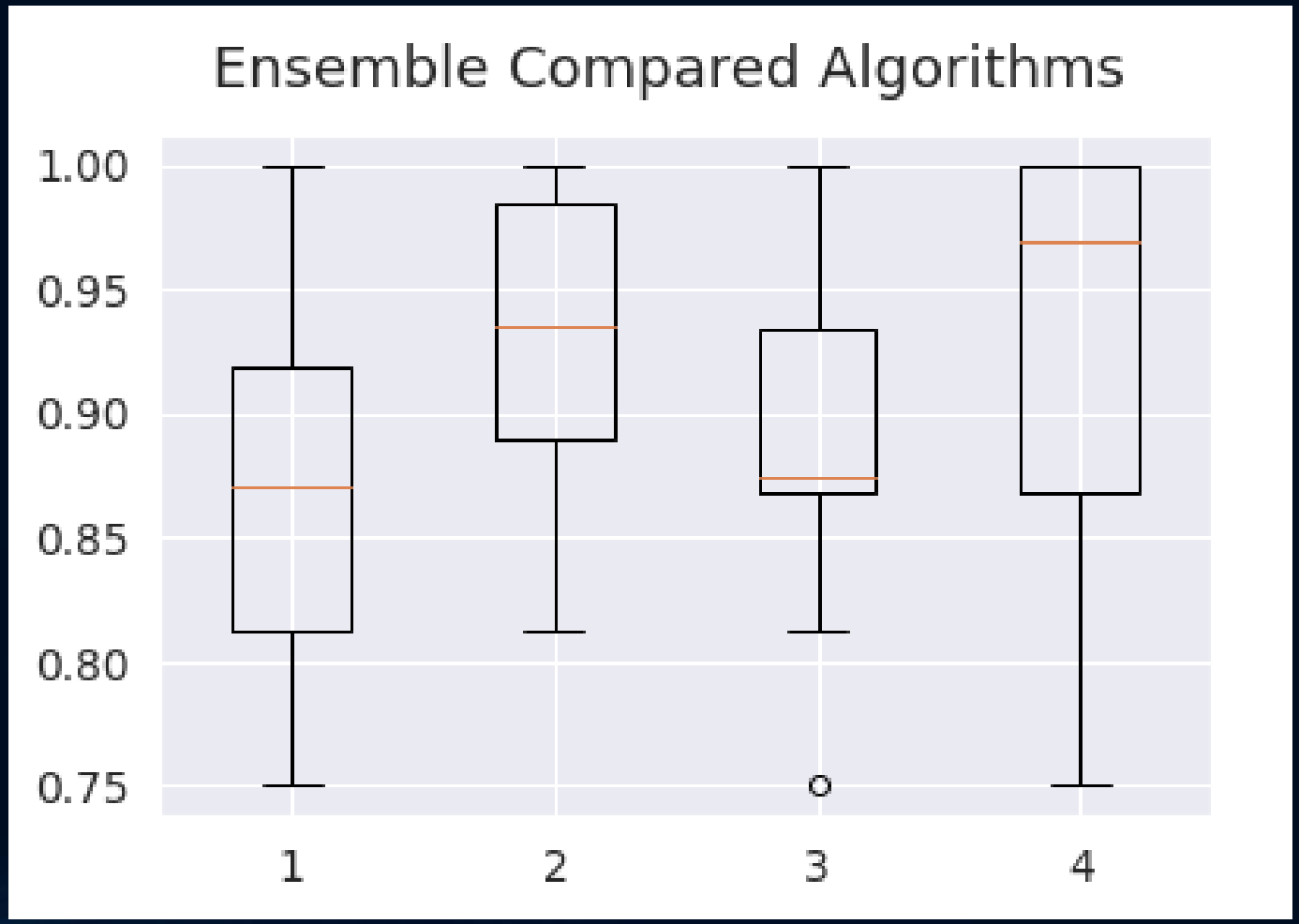
Ensemble Μάθηση
Boosting και Bagging αλγόριθμοι

1. Scaled AB : 0.867083 (0.070155)

2. Scaled GBC : 0.929583 (0.059729)

3. Scaled RFC : 0.885833 (0.066906)

4. Scaled ETC : 0.924167 (0.088010)



Regularization Tuning για τους 2 καλύτερους αλγορίθμους

- Gradient Boosting Classifier Tuning
  - Best: 0.942500 using {'learning rate’: 1.0, ‘n estimators': 100}
- Extra Trees Classifier Tuning
  - Best: 0.930833 using {‘n estimators': 30}

Αποτελέσματα

- Επιλογή του καλύτερου αλγορίθμου για ανίχνευση της νόσου Parkinson
- train set ακρίβεια 1.0 → train set matrix

38	0
0	118

- test set ακρίβεια 0.94 → test set matrix

8	2
0	29