# Αυτοματοποιήμενος Αναλυτής Δεδομένων


| | |
|---|---|
| Εκπόνηση | Ελένη Νησιώτη 7737 |
| Επίβλεψη | Επίκ. Καθ. Ανδρέας Συμεωνίδης |
| Συνεπίβλεψη | Δρ. Κυριάκος Χατζηδημητρίου |

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εργαστήριο Επεξεργασίας Πληροφορίας και Υπολογισμών

29 Μαρτίου 2017

issel

# Εισαγωγή

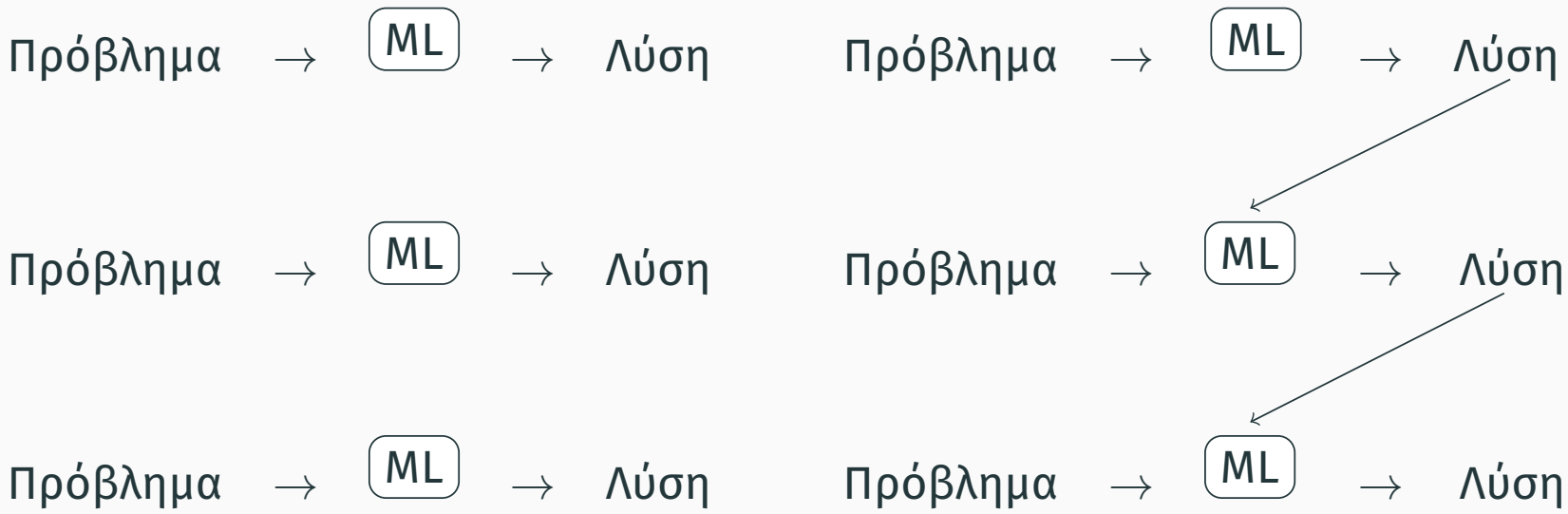
# Ο προβληματισμός

*Το* $75\%$ ενός πειράματος μηχανικής μάθησης αφιερώνεται στην προετοιμασία της εφαρμογής του αλγορίθμου και το $15\%$ στα βήματα που την ακολουθούν. Το μεγαλύτερο μέρος της έρευνας επικεντρώνεται στο ενδιάμεσο $10\%$ ...

— Rich Caruana, ICML 2015

Ανάγκη για επέκταση της τρέχουσας έρευνας σε χρονοβόρα στάδια της διαδικασίας εφαρμογής μηχανικής μάθησης που μέχρι τώρα γίνονταν χειροκίνητα.

# Η αναγκαιότητα της μεταφερσιμότητας



# Η εξέλιξη του ML ...

*Ο* προγραμματισμός στοχεύει στην αυτοματοποίηση, η μηχανική μάθηση στην αυτοματοποίηση της αυτοματοποίησης και η αυτοματοποιημένη μηχανική μάθηση στην αυτοματοποίηση του να αυτοματοποιείς την αυτοματοποίηση.

— Matthew Mayo, KDnuggets 2017

Το νέο στάδιο στην εξέλιξη της μηχανικής μάθησης στοχεύει στη δημιουργία μετα-γνώσης για την αυτοματοποίηση της ίδιας της διαδικασίας της μάθησης και όχι μεμονωμένων προβλημάτων.

# ... στην επιστήμη του AutoML

| | |
|---|---|
| Απαρχές | Unica, MarketSwitch, KXEN |
| Πεδία Εφαρμογής | Προ-επεξεργασία, Επιλογή αλγορίθμου, Ρύθμιση μοντέλου |
| Σύγχρονα Εργαλεία | Auto-WEKA, Microsoft Azure, caret, HPOlib, Google AutoML |

# Το προτεινόμενο σύστημα

Ένας αυτοματοποιημένος αναλυτής δεδομένων για προβλήματα δυαδικής ταξινόμησης με εμπειρία παλαιότερων πειραμάτων και κατανοητή έξοδο.

- Αυτόματος σχηματισμός βέλτιστου ensemble
- Ενσωμάτωση μετα-μάθησης για ρύθμιση μοντέλου
- Παραγωγή επεξηγηματικού report για το χρήστη

# Μεθοδολογία

# No free lunch theorem

*Αν λάβουμε υπόψιν όλες τις πιθανές κατανομές δημιουργίας δεδομένων, τότε κάθε αλγόριθμος μηχανικής μάθησης επιδεικνύει κατά μέσο όρο το ίδιο σφάλμα στην πρόβλεψη άγνωστων παραδειγμάτων.*

— David Wolpert, 1996

Είναι αδύνατη η εύρεση ενός γενικά βέλτιστου αλγορίθμου. Ένα εργαλείο βελτιστοποίησης γενικής φύσης προβλημάτων οφείλει να εξερευνήσει το χώρο των πιθανών μηχανισμών δημιουργίας δεδομένων και να προσφέρει προσαρμοσμένες λύσεις.

# Ρύθμιση μοντέλου: σύγχρονες τεχνικές βελτιστοποίησης

| Πλεγματική | Τυχαία | Bayesian |
|---|---|---|
| χρονοβόρα | naive | χρονοβόρα |
| ισοπίθανη αντιμετώπιση | | intuitive |

# Ρύθμιση μοντέλου: επισκόπηση

Οι σύγχρονες τεχνικές:

- είναι χρονοβόρες
- δεν μπορεί να αποδειχθούν γενικά βέλτιστες
- είναι ad-hoc

# Μετα-μάθηση

| | |
|---|---|
| Σκοπός | Δημιουργία μετα-γνώσης από πειράματα μηχανικής μάθησης |
| Τρόπος | Εξαγωγή μετα-χαρακτηριστικών των σετ δεδομένων, τα οποία περιέχουν ουσιώδη πληροφορία |
| Εφαρμογή | Πρόβλεψη βέλτιστων υπερπαραμέτρων αλγορίθμου μηχανικής μάθησης |

# Μετα-χαρακτηριστικά

**Πίνακας 1:** Λίστα μετα-χαρακτηριστικών μετά από εφαρμογή φιλτραρίσματος στα 100 σετ δεδομένων εκπαίδευσης

| | |
|---|---|
| Άθροισμα αθροισμάτων | Τυπική απόκλιση επιπέδων |
| Άθροισμα μέγιστων τιμών | Κυρτότητα επιπέδων |
| Μέση τιμή τυπικών αποκλίσεων | Λοξότητα επιπέδων |
| Μέση τιμή ελαχίστων τιμών | Πλήθος χαρακτηριστικών |
| Μέση τιμή κυρτοτήτων | Λογάριθμος πλήθους χαρακτηριστικών |
| Μέση τιμή λοξοτήτων | Πλήθος παραδειγμάτων |
| Τυπική απόκλιση ελαχίστων τιμών | Λογάριθμος πλήθους παραδειγμάτων |
| Ελάχιστη τιμή μέσων τιμών | Ποσοστό αγνώστων τιμών |
| Ελάχιστη τιμή τυπικών αποκλίσεων | Πλήθος αριθμητικών χαρακτηριστικών |
| Ελάχιστη τιμή ελαχίστων τιμών | Πλήθος κατηγορικών χαρακτηριστικών |
| Ελάχιστη τιμή μεγίστων τιμών | Μέγιστη πιθανότητα κλάσης |
| Ελάχιστη τιμή λοξοτήτων | Μέση τιμή πιθανοτήτων κλάσης |
| Κυρτότητα ελαχίστων τιμών | Ποσοστό PC για $95\%$ διακύμανση |
| Κυρτότητα μεγίστων τιμών | Κυρτότητα πρώτης PC |
| Λοξότητα λοξοτήτων | Λοξότητα PC |
| Άθροισμα επιπέδων | |

# Ensembles

*Μεταξύ ανταγωνιζομένων υποθέσεων πρέπει να επιλέγεται η απλούστερη.*

— Το ξυράφι του Occam

*Ο συνδυασμός σωστών λύσεων σε ένα πρόβλημα, δε μπορεί παρά να λύνει το πρόβλημα τουλάχιστον εξίσου καλά.*

— Επίκουρος

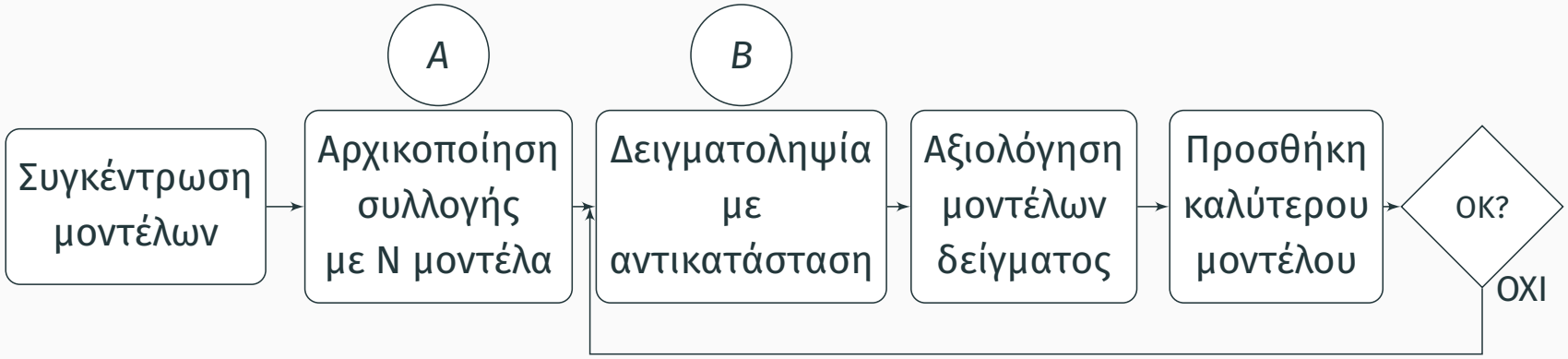
*Μία αναγκαία και ικανή συνθήκη για να είναι μία συλλογή μοντέλων πιο ακριβής από τα μοντέλα που την απαρτίζουν είναι αυτά να είναι ακριβή και ετερογενή.*

— Dietterich

# Ensembles από αποθήκες μοντέλων

| | |
|---|---|
| Πρόβλημα | Παρουσία πολλών, ετερογενών και ενδεχομένως κακής ποιότητας μοντέλων |
| Στόχος | Υπολογιστικά εφικτή τεχνική σχηματισμού συλλογής των αποδοτικότερων μοντέλων με αποφυγή υπερπροσαρμογής |

# Ensembles με προς τα εμπρός επιλογή μοντέλων



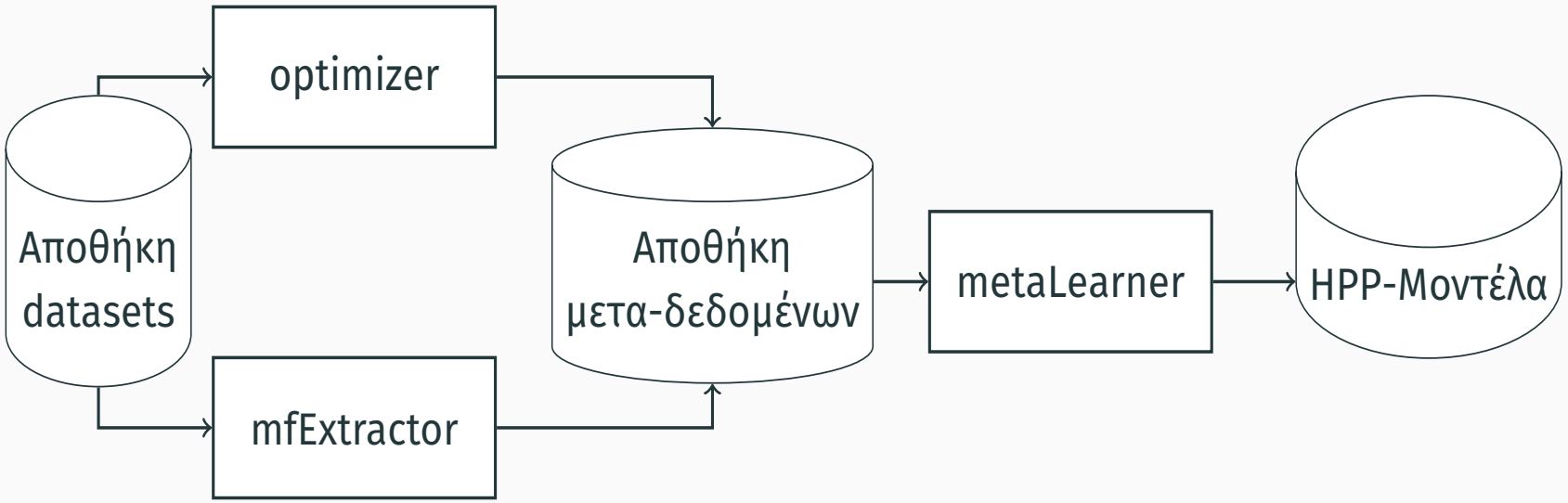

Α: αποφυγή υπερ-προσαρμογής σε μικρές αποθήκες

Β: αποφυγή υπερ-προσαρμογής σε μεγάλες αποθήκες και αναγκαστικής συμπερίληψης κακών μοντέλων

# Αρχιτεκτονική Συστήματος

# Υποσύστημα εκπαίδευσης

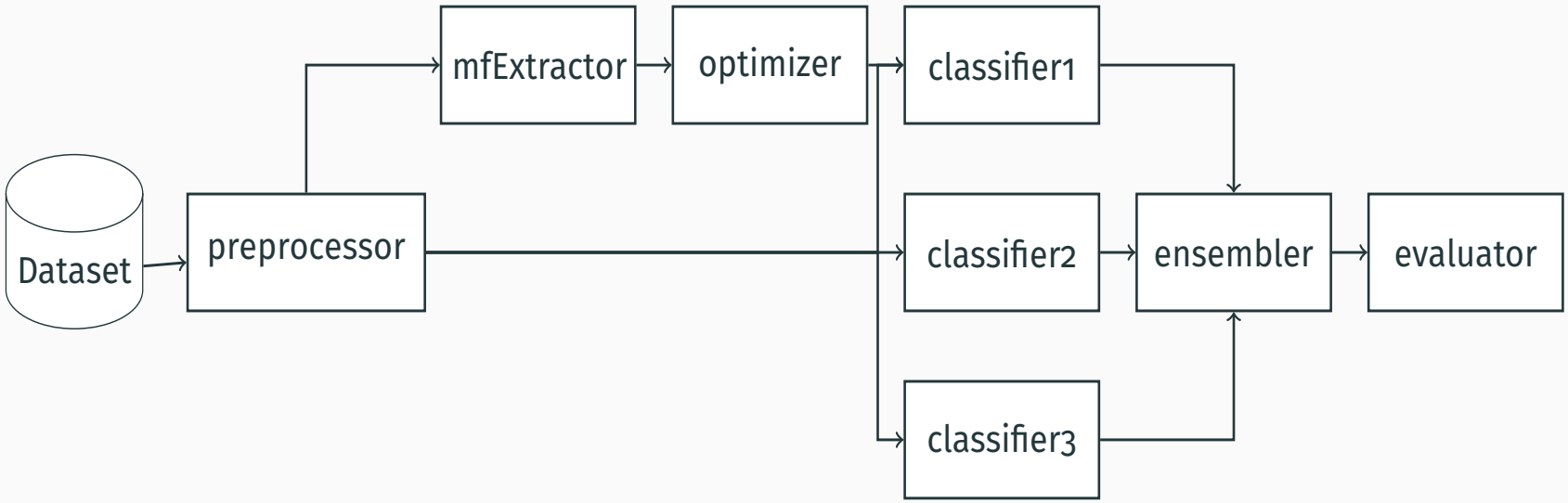
Σκοπός: Εκπαίδευση HyperParameterPrediction (HPP) μοντέλων




1. optimizer: Βελτιστοποίηση υπερ-παραμέτρων
2. mfExtractor: Εξαγωγή μετα-χαρακτηριστικών
3. metaLearner: Εκπαίδευση HPP μοντέλων

# Υποσύστημα πειράματος

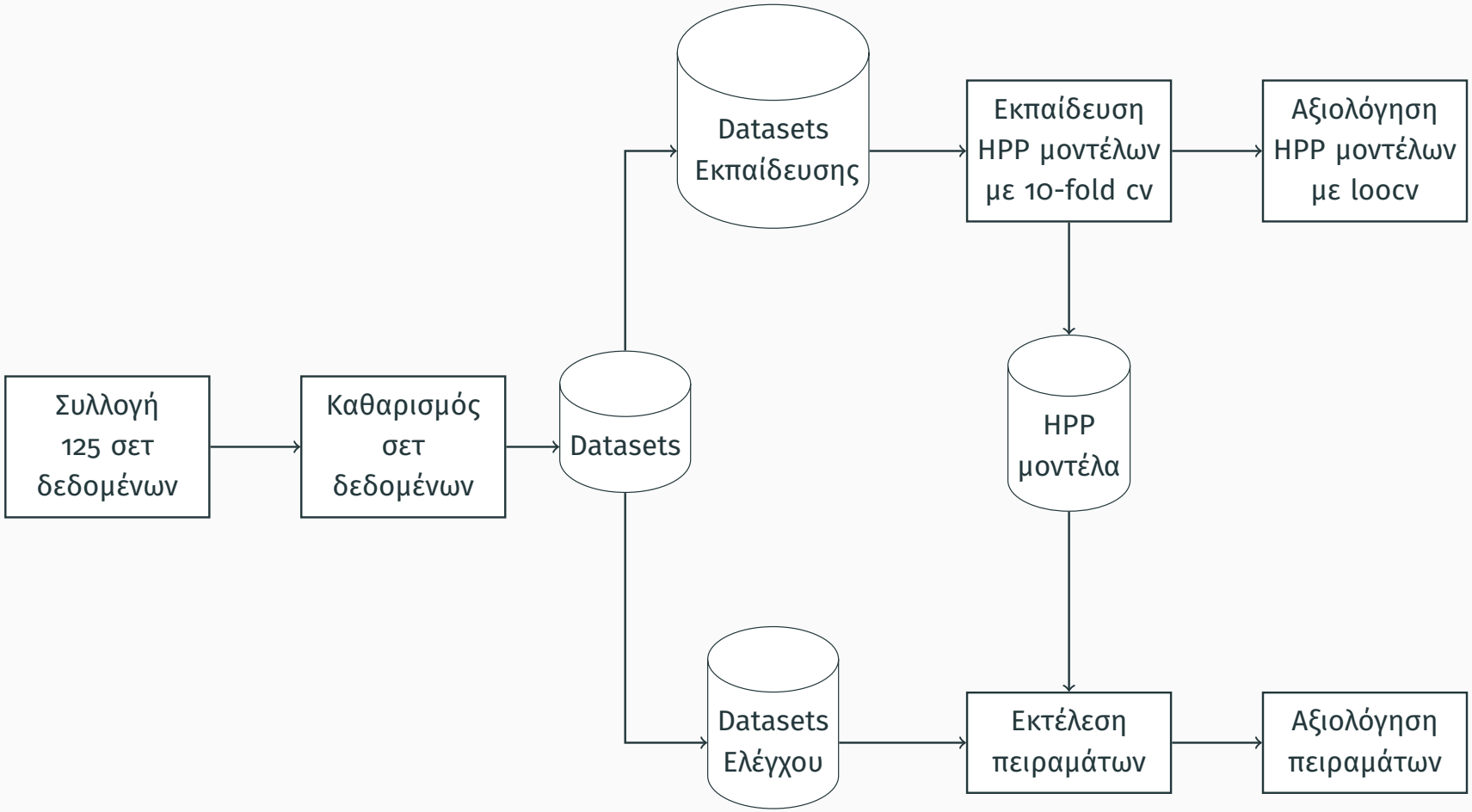
Σκοπός: Παραγωγή βέλτιστου ensemble για δεδομένο πρόβλημα




1. preprocessor: Προεπεξεργασία
2. mfExtractor: Εξαγωγή μετα-χαρακτηριστικών
3. optimizer: Ρύθμιση μοντέλων
5. $\text{classifier}_i$: Εκπαίδευση μοντέλου
6. ensembler: Σχηματισμός ensemble
7. evaluator: Αξιολόγηση

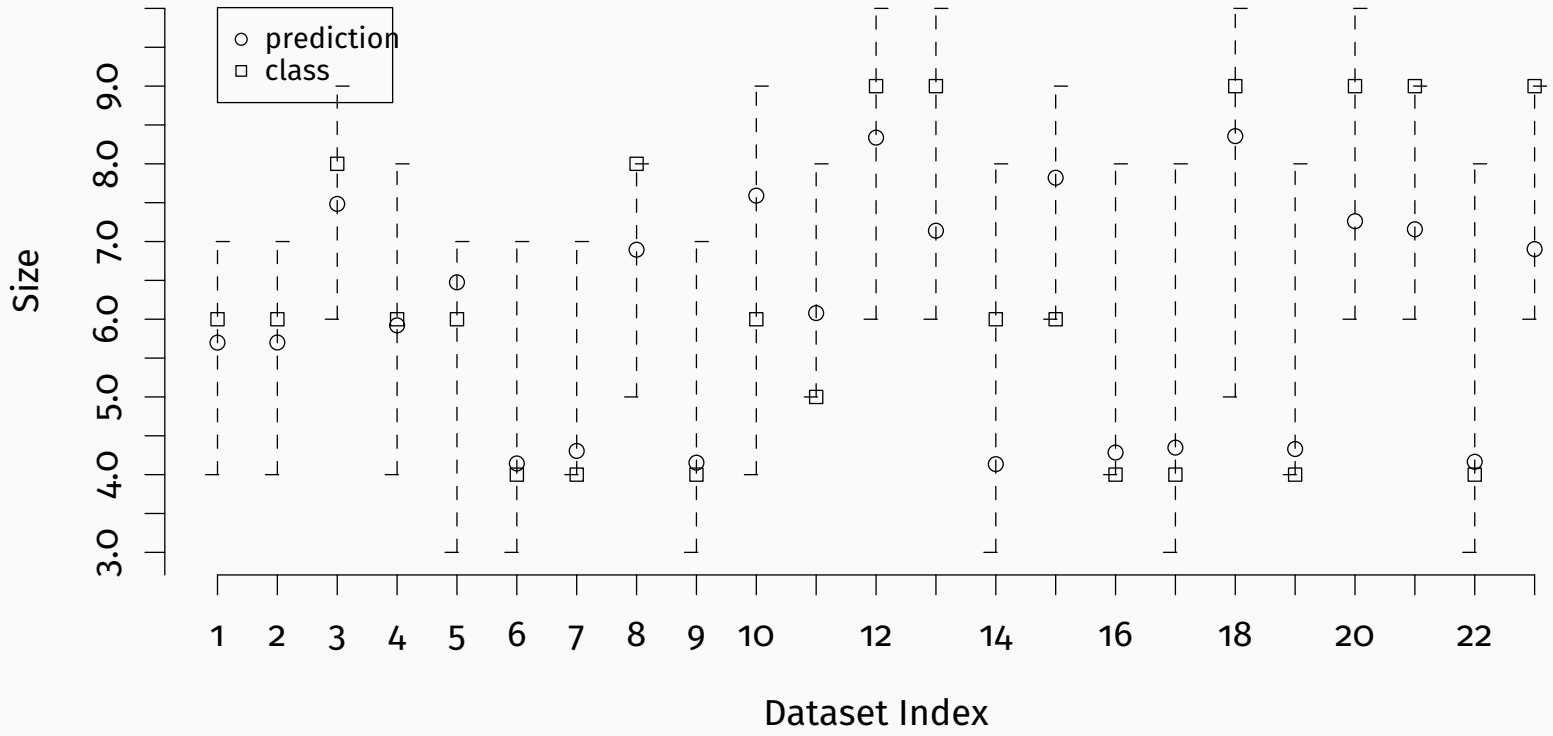
# Πειραματικά Αποτελέσματα

# Περιγραφή πειραμάτων



# Αξιολόγηση HPP μοντέλων

Αξιολόγηση HPP μοντέλου για πρόβλεψη υπερ-παραμέτρου size για το ANN μοντέλο

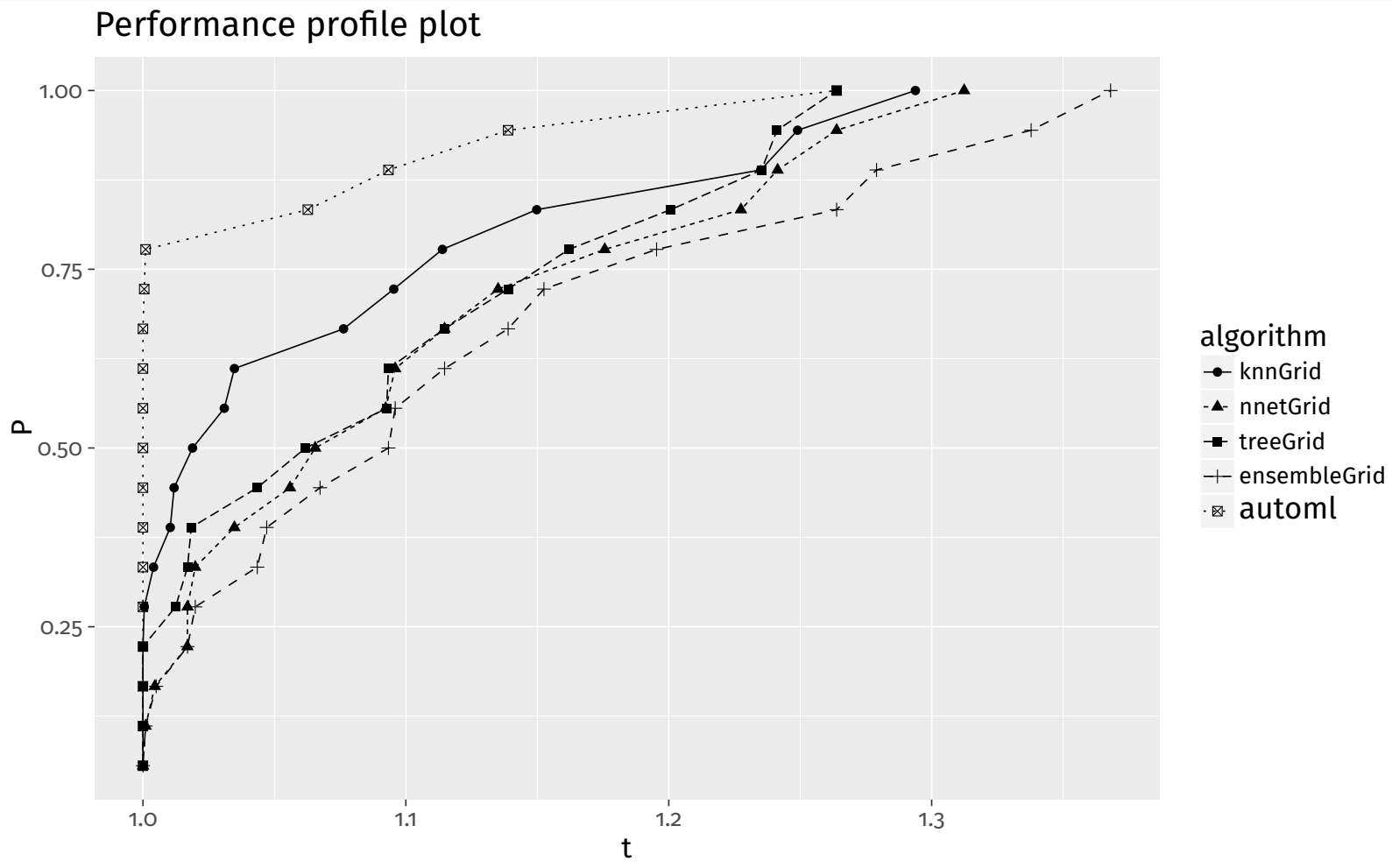


# Αξιολόγηση HPP μοντέλων: συμπεράσματα

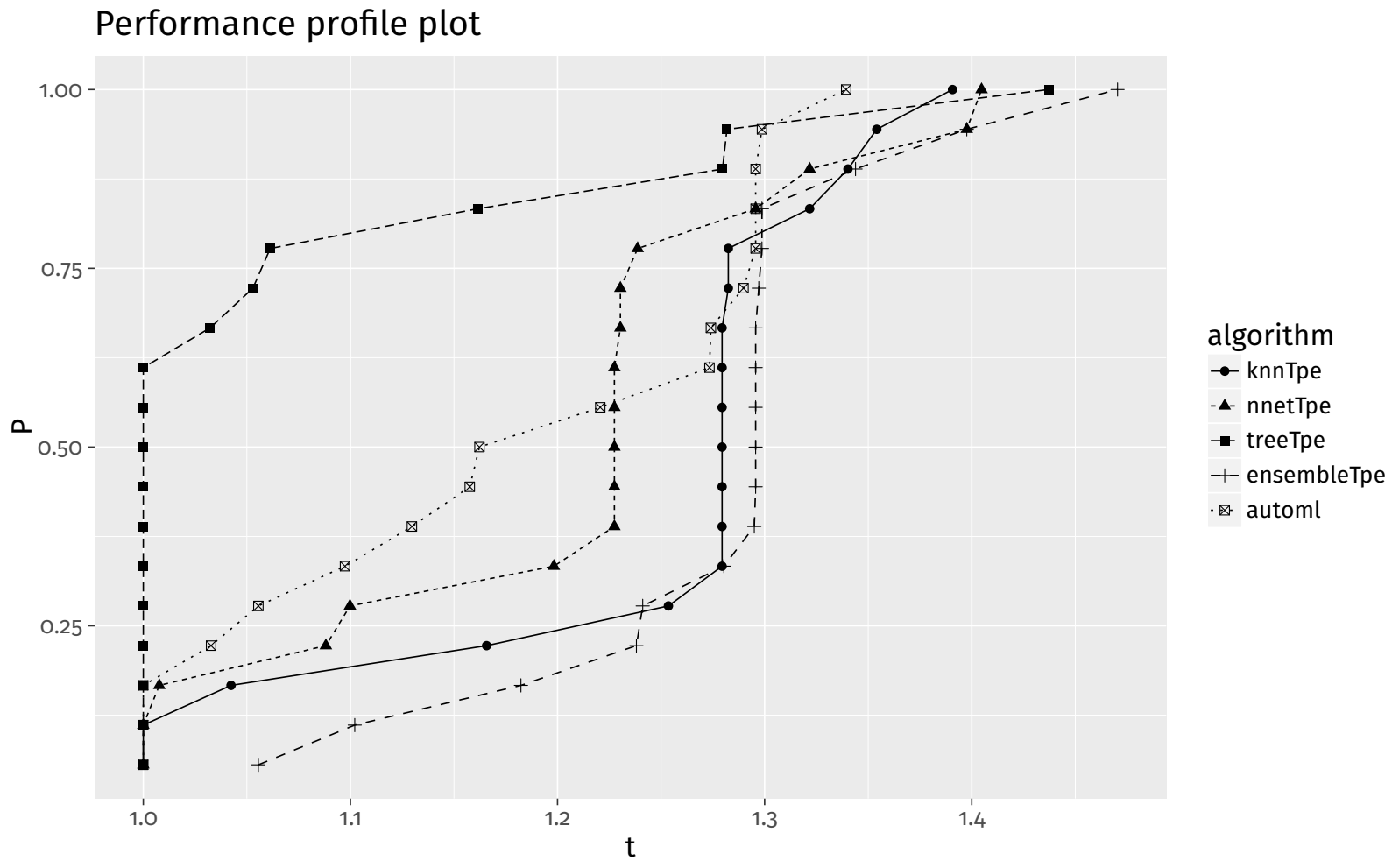
Από τα πειράματά μας προκύπτει ότι:

- τα HPP μοντέλα δεν προβλέπουν επακριβώς τις βέλτιστες υπερ-παραμέτρους.
- η χρήση διαστημάτων πρόβλεψης οδηγεί σχεδόν πάντα στη συμπερίληψη της σωστής τιμής.

# Αξιολόγηση συστήματος: σύγκριση με πλεγματική αναζήτηση



# Αξιολόγηση συστήματος: σύγκριση με Bayesian βελτιστοποίηση



# Αξιολόγηση συστήματος: συμπεράσματα

Με βάση τα διαγράμματα προφίλ απόδοσης και τα στατιστικά τεστ που εφαρμόσαμε μπορούμε να συμπεράνουμε ότι:

- το σύστημά μας είναι αποδοτικότερο από όλα τα μοντέλα που ρυθμίστηκαν με πλεγματική αναζήτηση
- το σύστημά μας είναι αποδοτικότερο από όλα τα μοντέλα που ρυθμίστηκαν με bayesian βελτιστοποίηση, εκτός από τα CART δέντρα.

# Μελλοντικές επεκτάσεις

**Βελτίωση μοντέλων μετα-μάθησης:**

- εύρεση νέων μετα-χαρακτηριστικών
- πειραματισμός με μεταβλητά διαστήματα εμπιστοσύνης

**Περαιτέρω παραλληλοποίηση των embarassingly parallel διαδικασιών**:

- βελτιστοποίηση υπερ-παραμέτρων (διαστήματα πρόβλεψης)

**Ενσωμάτωση διεπαφών αυτοματοποίησης για**:

- συλλογή σετ δεδομένων
- εκπαίδευση μετα-μοντέλων
- χρήση ευριστικών κανόνων

?