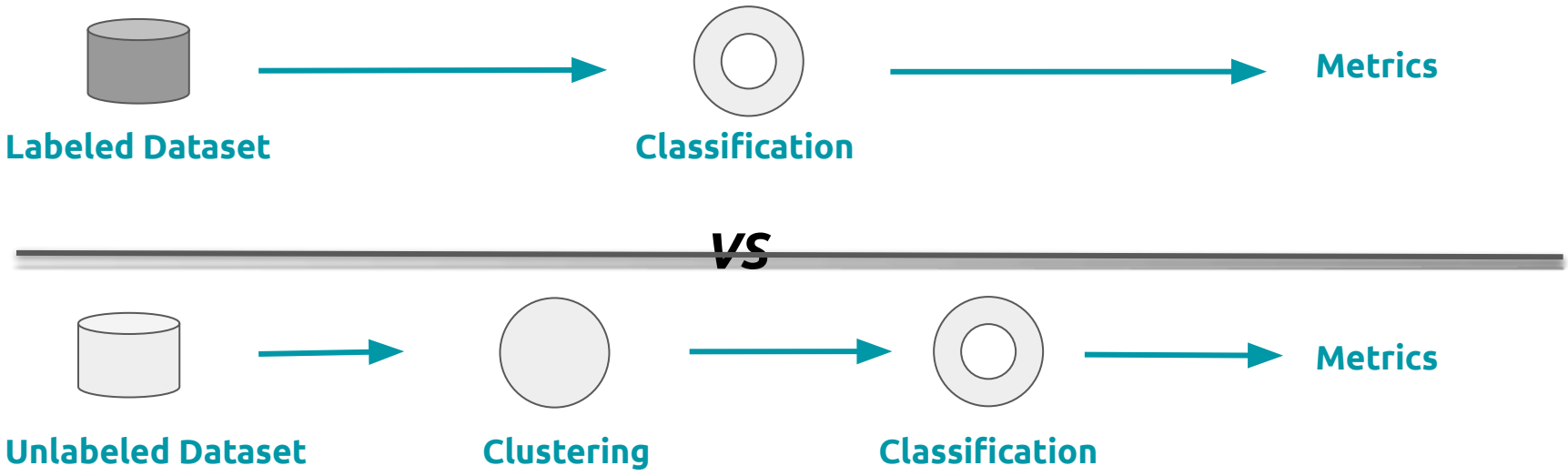
Υπολογιστική σύγκριση αλγορίθμων συσταδοποίησης

Διπλωματική εργασία του Μαντιού Ιωάννη
mai20034

Περιεχόμενα

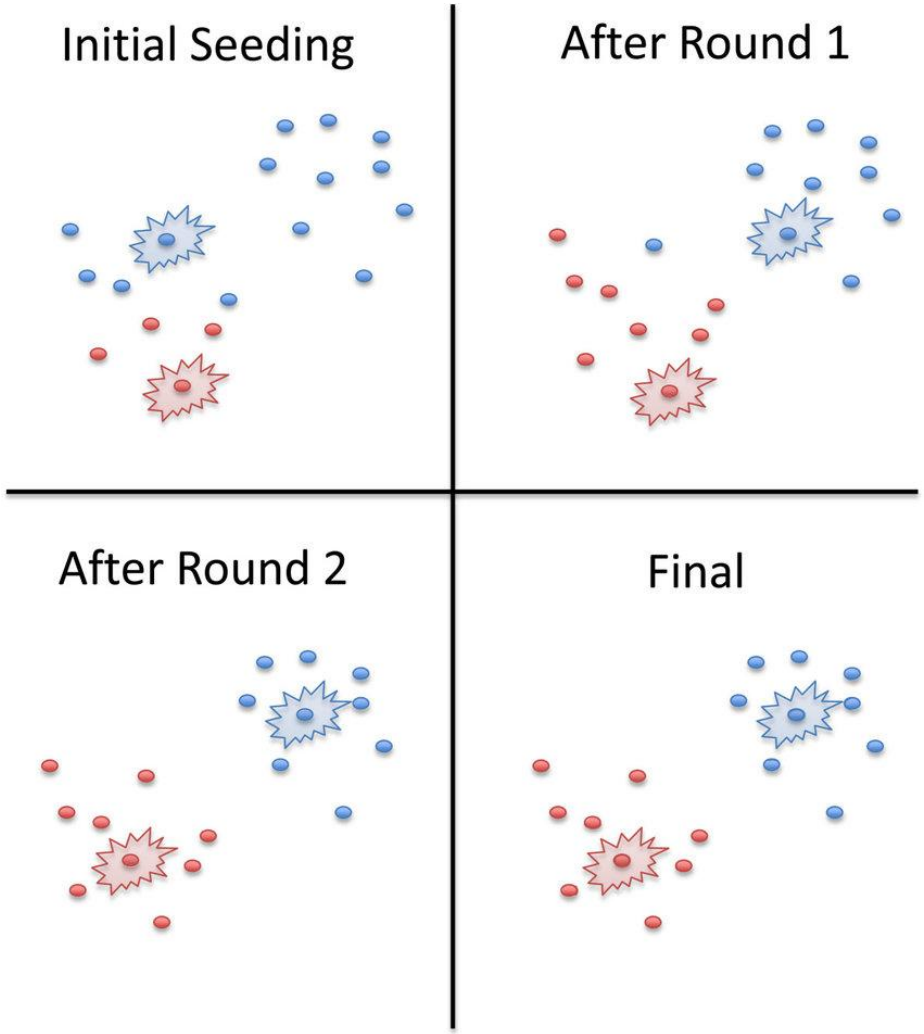
1. Case study : τί προσπαθούμε να πετύχουμε


Labeled Dataset
Classification
Metrics
VS
Unlabeled Dataset
Clustering
Classification
Metrics

2. Αλγόριθμοι clustering

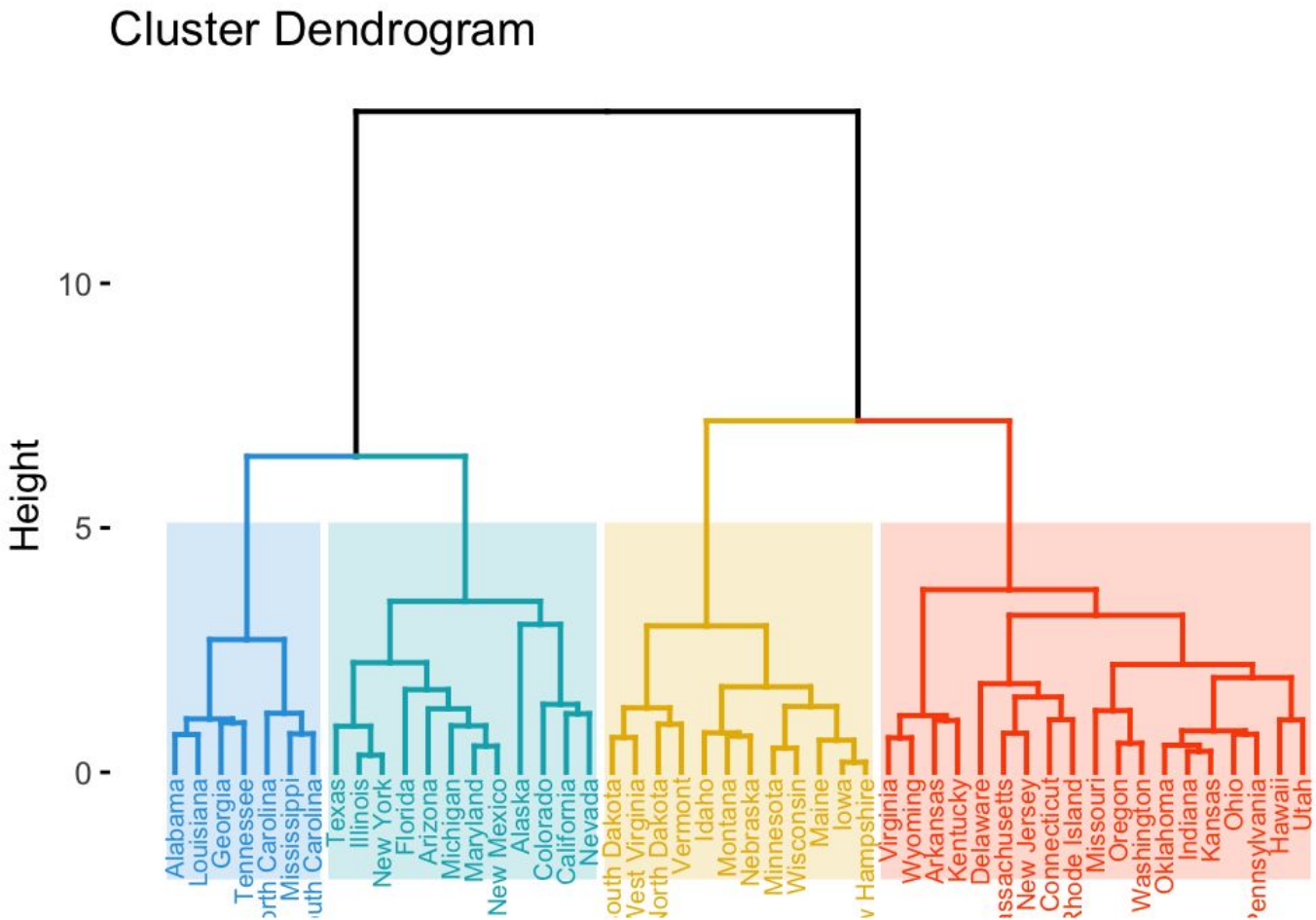
2.1 K-Means

- Κάθε συστάδα σχετίζεται με ένα κεντρικό σημείο (centroid)
- Κάθε σημείο ανατίθεται στην συστάδα με το κοντινότερο κεντρικό σημείο (π.χ. τετραγωνική ευκλείδεια απόσταση)
- Ξαναϋπολογίζονται τα κεντρικά σημεία
- Επιλέγουμε εκ των προτέρων τον αριθμό των ομάδων (K)


Initial Seeding
After Round 1
After Round 2
Final

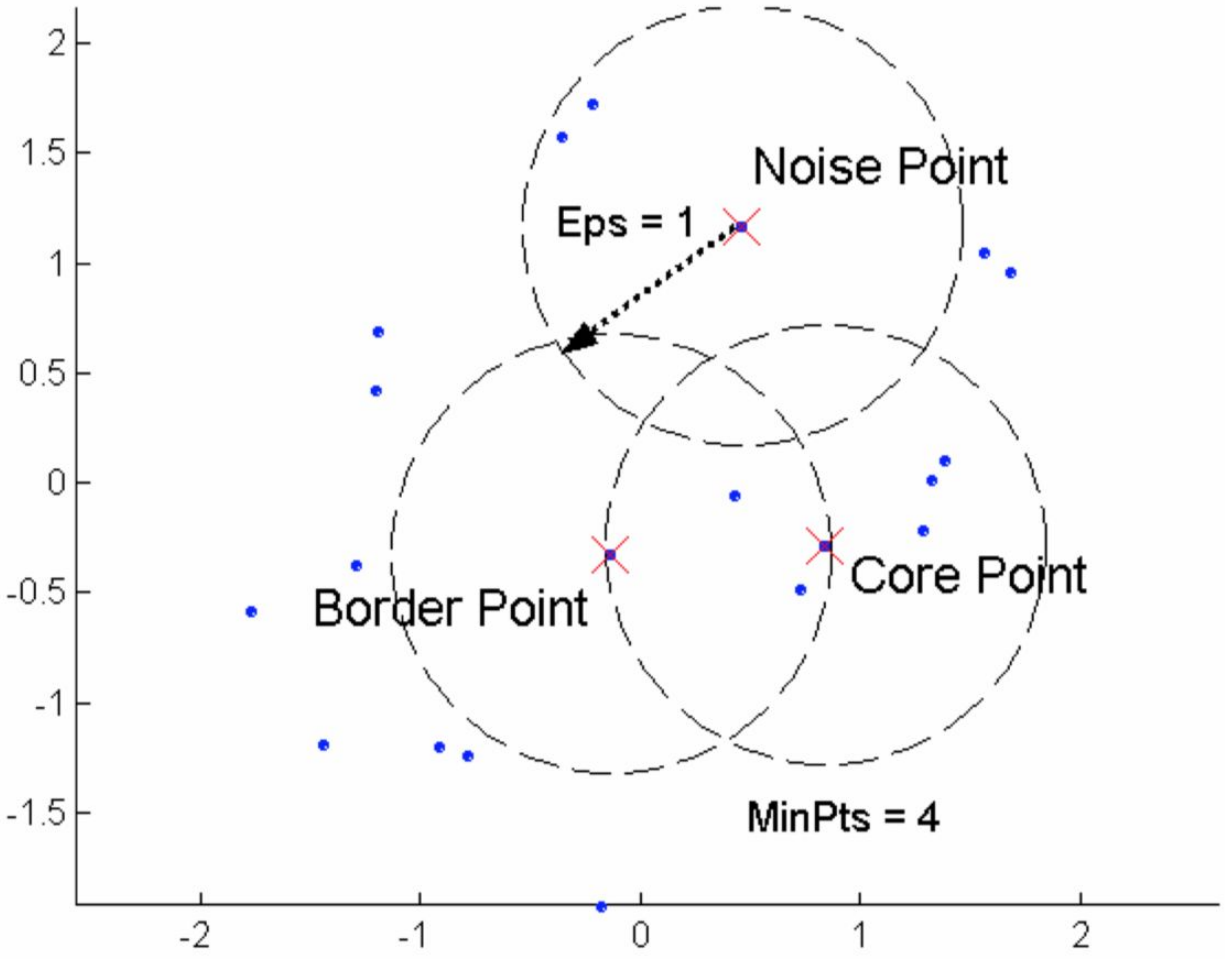
2.2 Agglomerative

- Δημιουργεί ομάδες με ιεραρχικό τρόπο από κάτω προς τα πάνω
- Στην αρχή όλες οι παρατηρήσεις αποτελούν ατομικές ομάδες οι οποίες συνενώνονται διαδοχικά σε μεγαλύτερες ομάδες.
- Επιλογή και ένωση ομάδων με βάση κάποιο κριτήριο ελάχιστης απόστασης και συγχώνευση των παρατηρήσεων.
- Δενδρόγραμμα


Cluster Dendrogram
Height
10
5
0
Alabama
Louisiana
Georgia
Tennessee
Mississippi
Texas
Illinois
New York
Florida
Arizona
Michigan
Maryland
New Mexico
Alaska
Colorado
California
Nevada
Idaho
Montana
Nebraska
Minnesota
Wisconsin
Maine
Iowa
Virginia
Wyoming
Arkansas
Kentucky
Delaware
New Jersey
Connecticut
Missouri
Oregon
Washington
Oklahoma
Indiana
Kansas
Ohio
Hawaii
Utah

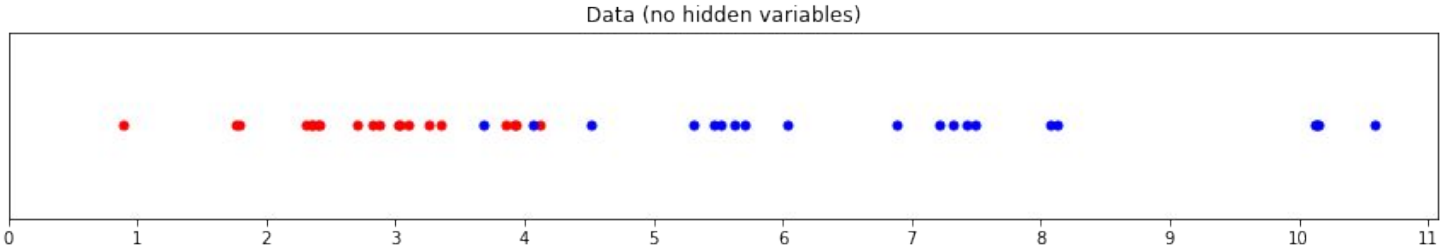
2.3 DBSCAN

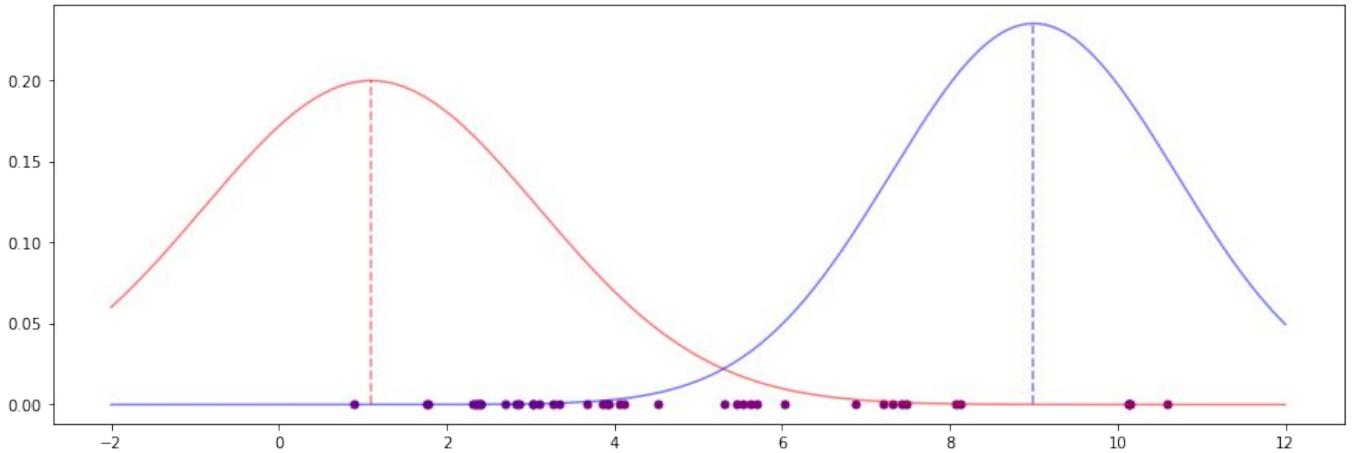
- Βασίζεται στην πυκνότητα των σημείων μέσα σε μια προκαθορισμένη ακτίνα
- Καθορισμός minPts και Eps
- Σημεία: Βασικά, Οριακά, Θορύβου

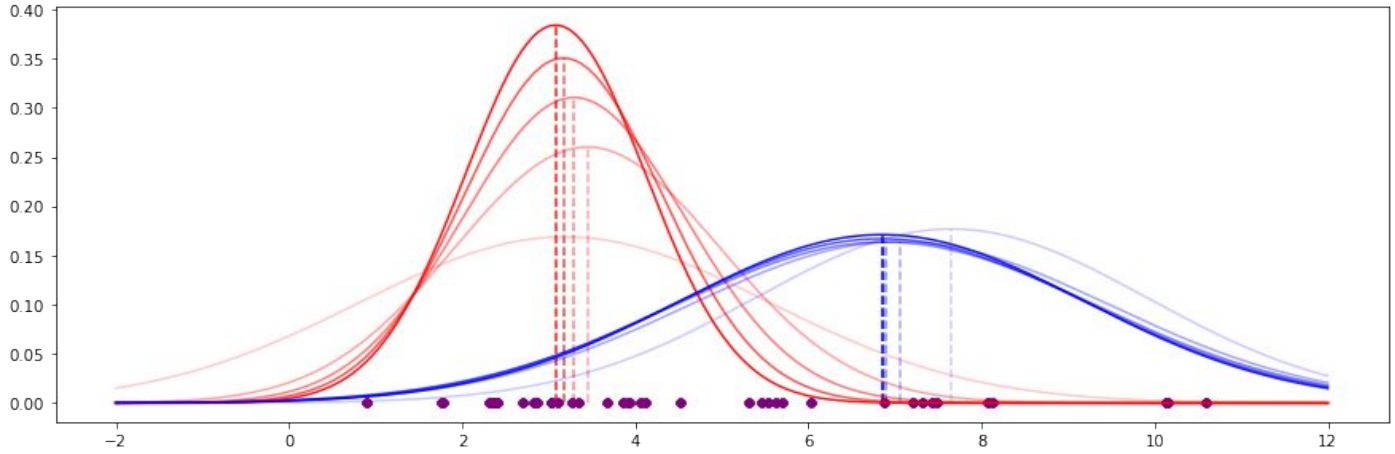

Noise Point
Eps = 1
Border Point
Core Point
MinPts = 4
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1
0
1
2

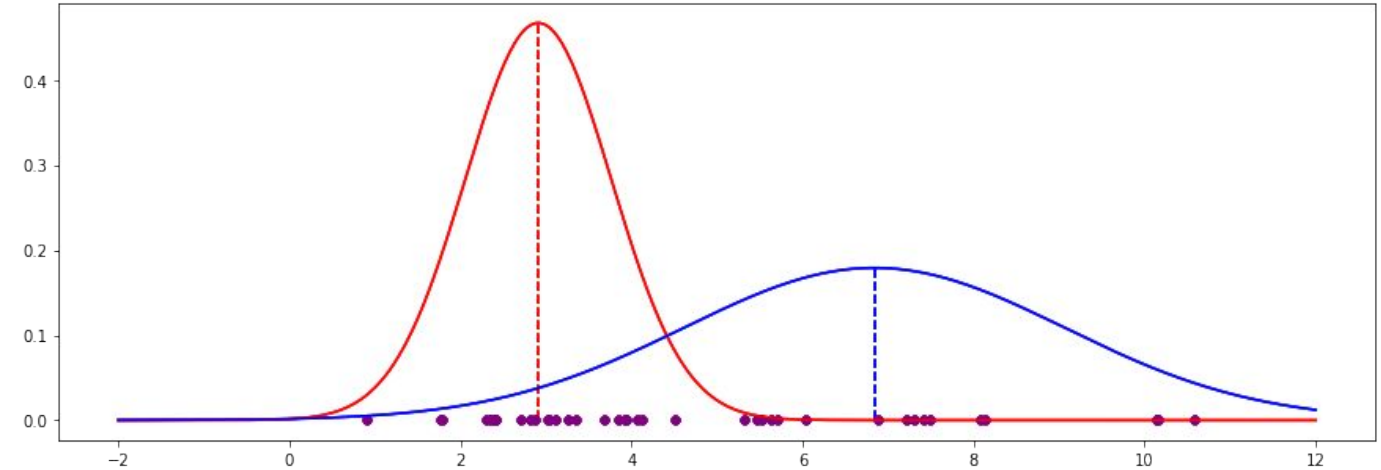
2.4 Expectation-Maximization

- Βασίζεται σε ένα μοντέλο πιθανοτήτων που συνδυάζει διαφορετικές Gaussian κατανομές
- Επιλέγει τυχαίες αρχικές τιμές για το σύνολο Θ των παραμέτρων που προσδιορίζουν την κάθε κατανομή/ ομάδα (π.χ. μέση τιμή, τυπική απόκλιση κτλ.)
- Όσο αλλάζουν οι τιμές των παραμέτρων του Θ : υπολογίζει για κάθε σημείο αν ανήκει σε μια κατανομή (expectation) και για τις πιθανότητες αυτές υπολογίζει το νέο Θ (maximization)


Data (no hidden variables)
0
1
2
3
4
5
6
7
8
9
10
11


0.20
0.15
0.10
0.05
0.00
−2
0
2
4
6
8
10
12

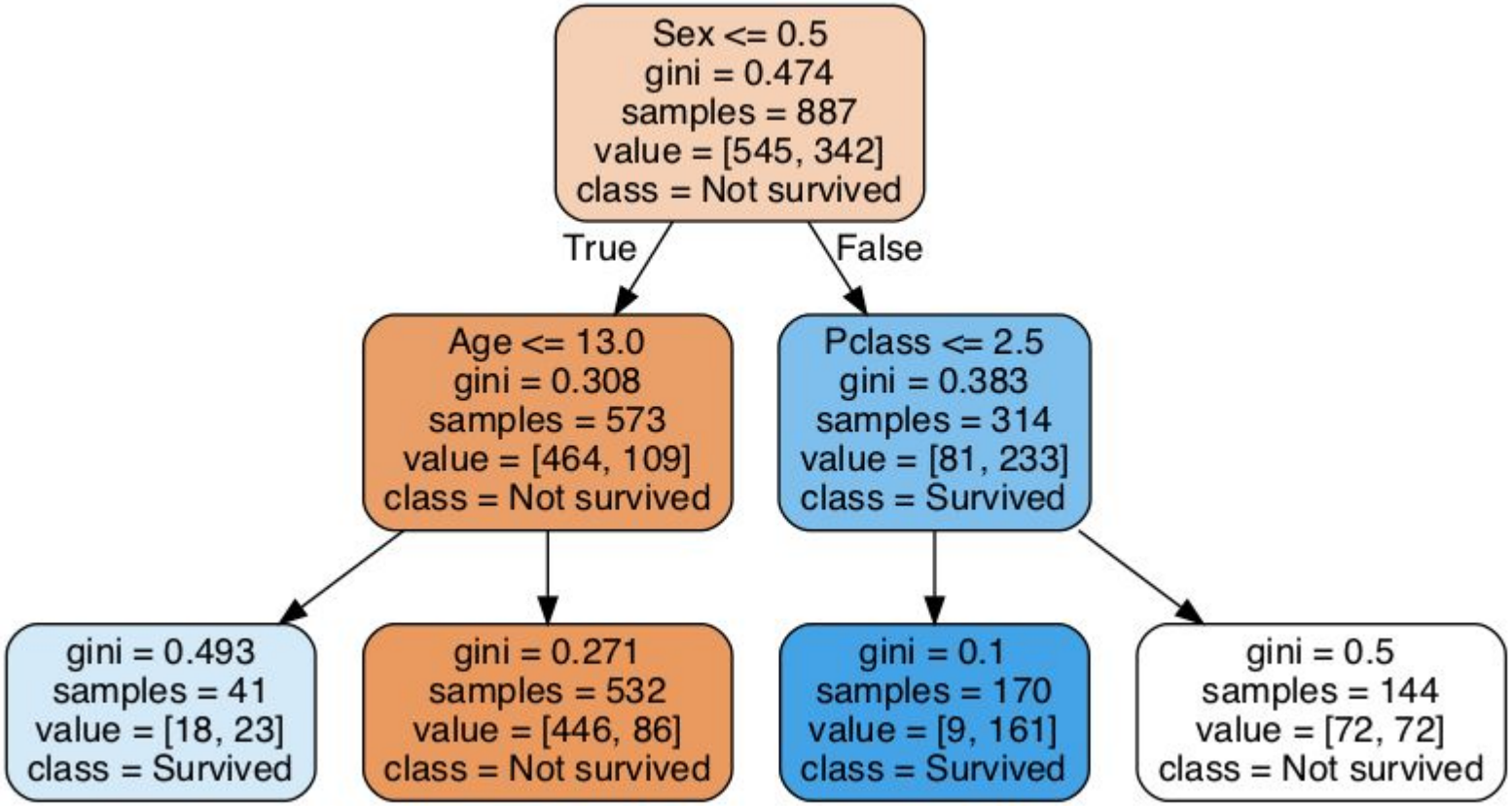

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−2
0
2
4
6
8
10
12


0.4
0.3
0.2
0.1
0.0
−2
0
2
4
6
8
10
12

3. Αλγόριθμοι classification

3.1 Decision Trees (CART)

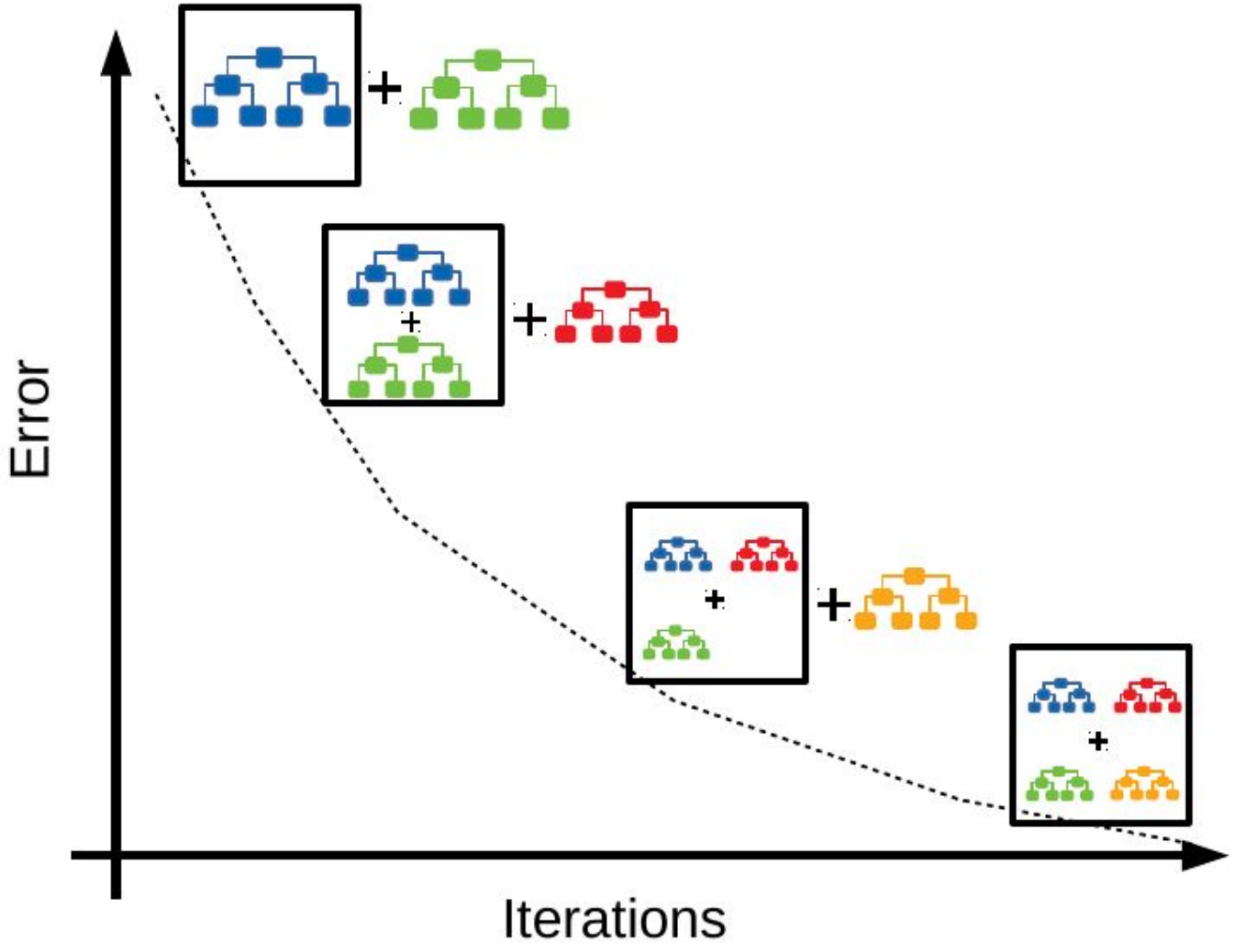
- Επιλέγει το χαρακτηριστικό που πετυχαίνει τον καλύτερο διαχωρισμό μεταξύ των κατηγοριών.
- Χωρίζει τα δεδομένα σε υποσύνολα με βάση της τιμές του χαρακτηριστικού αυτού.
- Για κάθε υποσύνολο που περιέχει περισσότερες από μία κατηγορίες, επαναλαμβάνει τη διαδικασία.
- Σταματάει εφόσον δεν υπάρχουν υποσύνολα που περιέχουν περισσότερες από μία κατηγορίες ή έχουν χρησιμοποιηθεί όλα τα χαρακτηριστικά.


Sex <= 0.5
gini = 0.474
samples = 887
value = [545, 342]
class = Not survived
True
False
Age <= 13.0
gini = 0.308
samples = 573
value = [464, 109]
class = Not survived
Pclass <= 2.5
gini = 0.383
samples = 314
value = [81, 233]
class = Survived
gini = 0.493
samples = 41
value = [18, 23]
class = Survived
gini = 0.271
samples = 532
value = [446, 86]
class = Not survived
gini = 0.1
samples = 170
value = [9, 161]
class = Survived
gini = 0.5
samples = 144
value = [72, 72]
class = Not survived

3.2 Gradient Boosting

- Η βασική ιδέα του Boosting είναι ο συνδυασμός αλγορίθμων των οποίων το σφάλμα είναι ελαφρώς καλύτερο από την τυχαία επιλογή.
- Ο αλγόριθμος Gradient Boosting προσθέτει νέα μοντέλα με στόχο την διόρθωση σφαλμάτων που έγιναν από τα υπάρχοντα μοντέλα. Τα μοντέλα προστίθενται διαδοχικά μέχρις ότου δεν μπορούν να γίνουν περαιτέρω βελτιώσεις.

- Ένα ενισχυμένο δέντρο χρησιμοποιεί μικρότερα δέντρα που εξηγούν μόνο ένα κομμάτι των χαρακτηριστικών ανά δέντρο.
- Ο αλγόριθμος δημιουργεί ένα νέο δέντρο που αναλύει τα προηγούμενα δένδρα. Στη συνέχεια δημιουργεί ένα νέο δέντρο που επιχειρεί να διορθώσει τα σφάλματα στα προηγούμενα κ.ο.κ.


Error
Iterations

3.3 Extreme Gradient Boosting

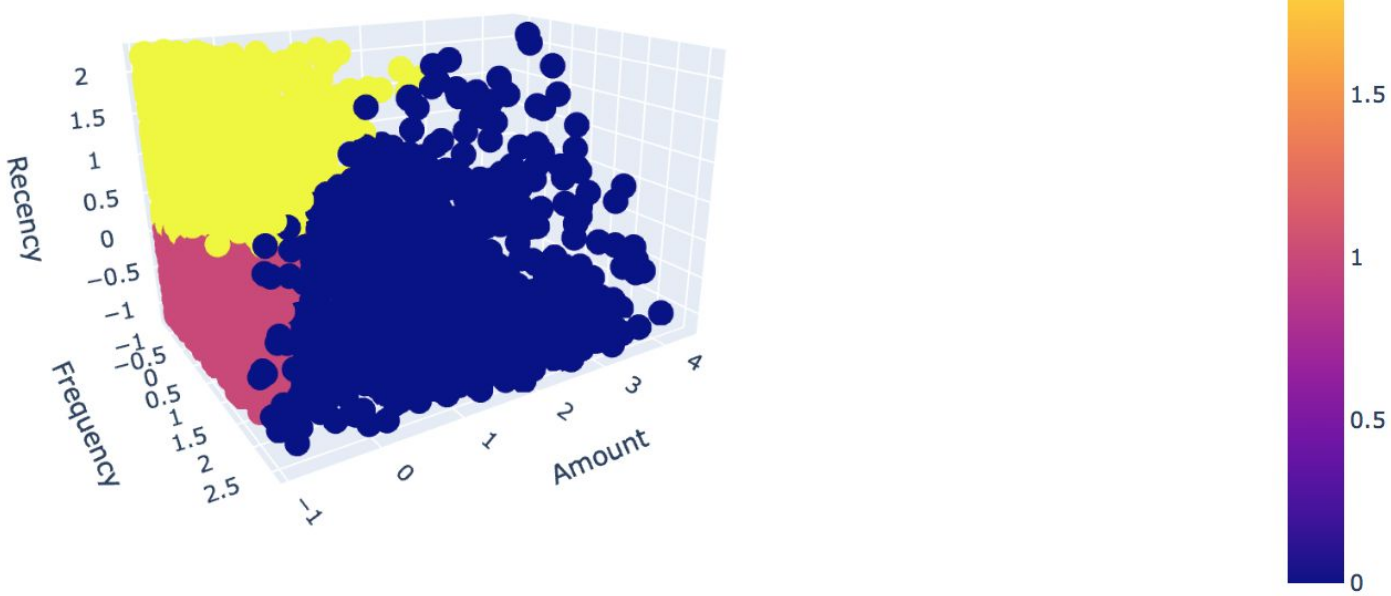
- Η state of the art έκδοση του gradient boosting.
- Ο XGBoost επικεντρώνεται στην υπολογιστική ταχύτητα και την απόδοση του μοντέλου.
- Καινοτομίες του XGBoost: αυτόματος χειρισμός των δεδομένων που λείπουν και δομή block για τον παραλληλισμό της δομής των δέντρων

4. Unlabeled Datasets

4.1 Online Retail II Dataset

- Συναλλαγές 2 ετών ενός online καταστήματος
- Χαρακτηριστικά του δείγματος: Αριθμός τιμολογίου, κωδικός προϊόντος, περιγραφή, ποσότητα, ημερομηνία αγοράς, τιμή μονάδας, κωδικός πελάτη, χώρα πελάτη.
- Ανάλυση RFM (Recency, Frequency, Monetary) για την αξιολόγηση κάθε πελάτη.

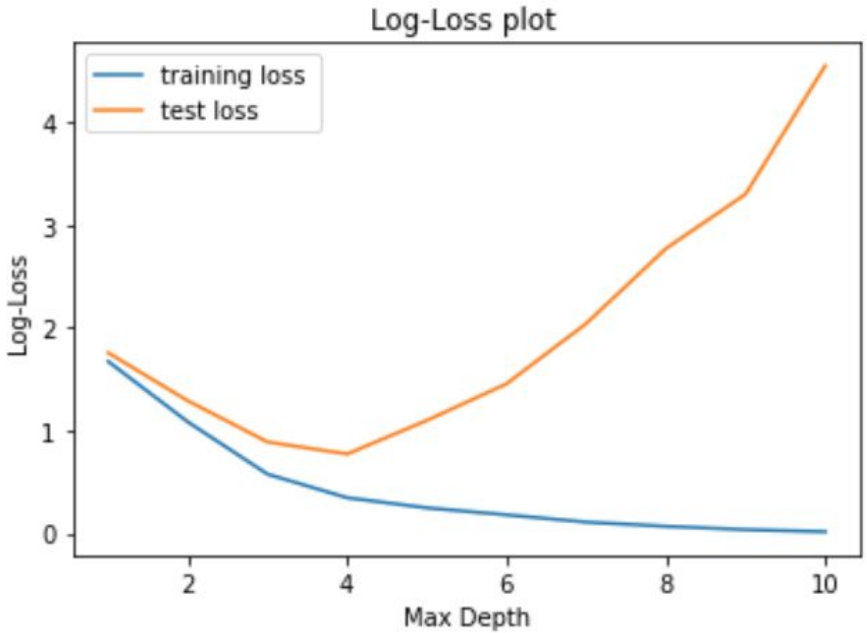
- Μετά από καθαρισμό των κακών δεδομένων και ομαδοποίηση κατά το μοντέλο RFM: πίνακας διαστάσεων 1.000.000 X 8 >>> 4.075 X 3
- Πριν το clustering: διαγραφή outliers και standardization


Recency
2
1.5
1
0.5
0
−0.5
−1
Frequency
−1
−0.5
0
0.5
1
1.5
2
2.5
Amount
−1
0
1
2
3
4
kmeans
2
1.5
1
0.5
0

Μετά την τεχνητή παραγωγή κλάσεων μέσω clustering:

- Oversampling με αλγόριθμο SMOTE
- Stratified Train/test split: 80%/20%
- Αντιμετώπιση overfitting βάσει log-loss μεταξύ training και test set

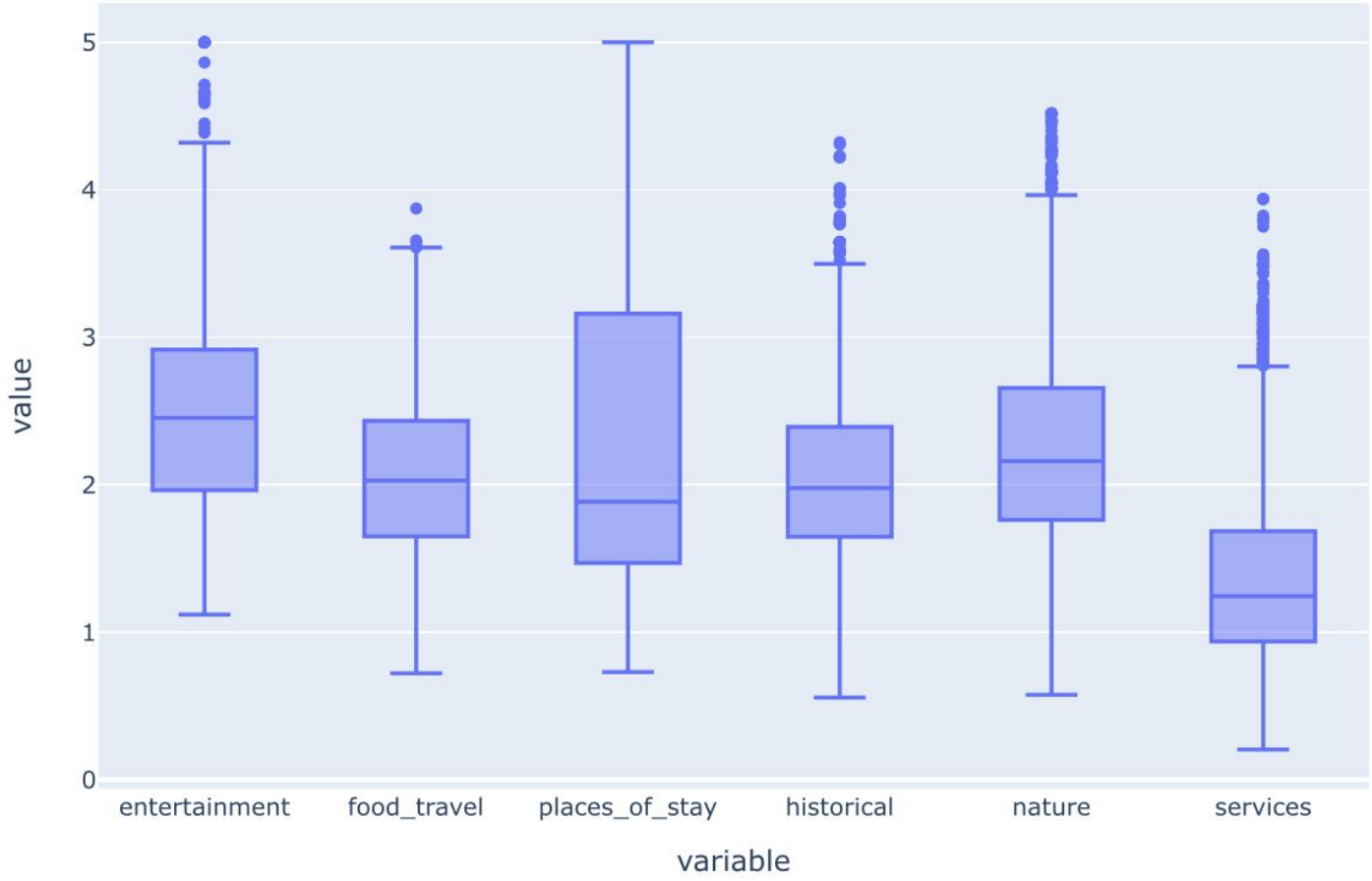
Παράδειγμα log-loss plot



4.2 Travel Review Ratings Dataset

- Κριτικές google σε 24 κατηγορίες χώρων στην Ευρώπη
- Αξιολόγηση με αστέρια από 1 εώς 5

- Γκρουπάρισμα ομοειδών τοποθεσιών σε κατηγορίες για καλύτερο clustering (μείωση του dimensionality)
- Δεν χρειάζεται standardization, ίδιο εύρος τιμών μεταξύ των κατηγοριών


5
4
3
2
1
0
value
entertainment
food_travel
places_of_stay
historical
nature
services
variable

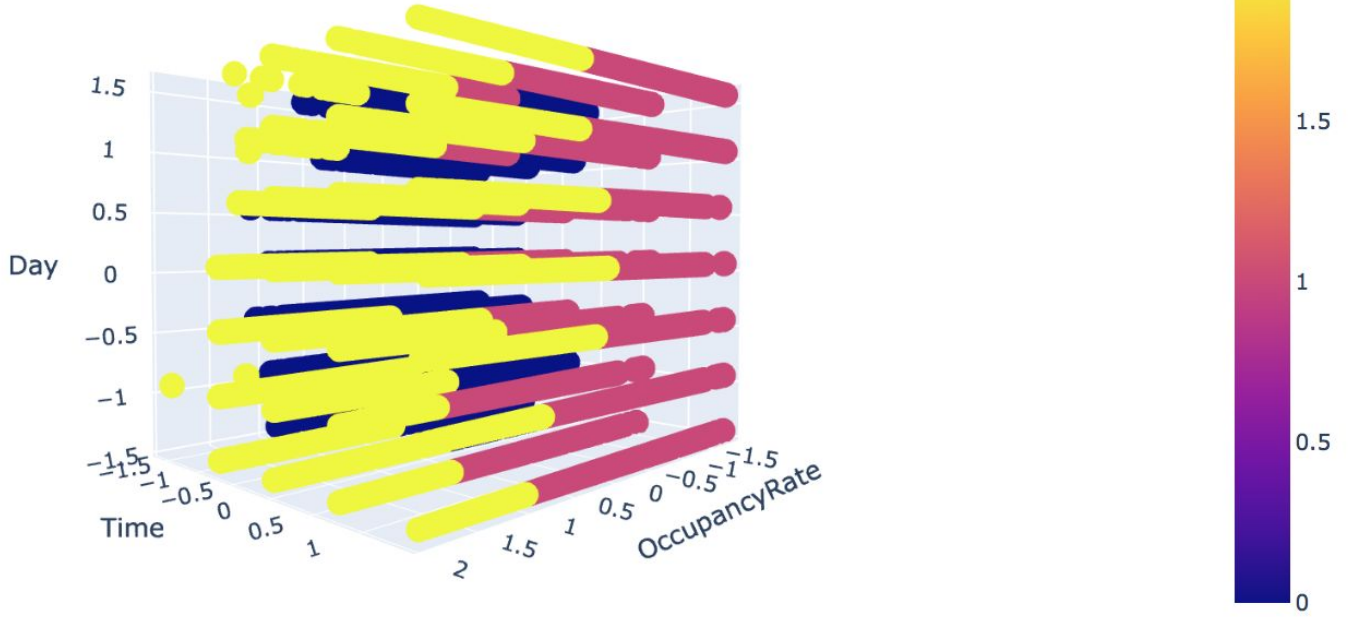
Μετά την τεχνητή παραγωγή κλάσεων μέσω clustering:

- Oversampling με αλγόριθμο SMOTE
- Stratified Train/test split: 80%/20%
- Αντιμετώπιση overfitting βάσει log-loss μεταξύ training και test set

4.3 Birmingham Parking Dataset

- Ημερήσια δεδομένα 2 μηνών από δημοτικά πάρκινγκ στο Birmingham
- Χαρακτηριστικά του δείγματος: κωδικός πάρκινγκ, χωρητικότητα, πληρότητα, στιγμή καταγραφής

- Ομαδοποίηση του χρόνου σε δίωρα και ημέρες
- Δημιουργία στήλης ‘Ποσοστό Πληρότητας’
- Τελικές στήλες για clustering : Χωρητικότητα, Ποσοστό Πληρότητας, Ώρα, Ημέρα


Day
1.5
1
0.5
0
−0.5
−1
−1.5
Time
−1.5
−1
−0.5
0
0.5
1
2
1.5
1
0.5
0
−0.5
−1
−1.5
OccupancyRate
Agg_Labels
2
1.5
1
0.5
0

Μετά την τεχνητή παραγωγή κλάσεων μέσω clustering:

- Oversampling με αλγόριθμο SMOTE
- Stratified Train/test split: 80%/20%
- Αντιμετώπιση overfitting βάσει log-loss μεταξύ training και test set

5. Labeled Datasets

5.1 Teaching Assistant Evaluation Dataset

- Αξιολογήσεις 151 καθηγητών σε περίοδο 2 χρόνων στο πανεπιστήμιο του Wisconsin
- Βαθμοί αξιολόγησης: low, medium, high
- Χαρακτηριστικά του δείγματος: γλώσσα ομιλίας του διδάσκοντα, διδάσκοντας, μάθημα, εξάμηνο, μέγεθος τάξης, αξιολόγηση

- Δεν χρειάζεται προεπεξεργασία των δεδομένων
- Oversampling με αλγόριθμο SMOTE
- Stratified Train/test split: 80%/20%
- Αντιμετώπιση overfitting βάσει log-loss μεταξύ training και test set

5.2 Car Evaluation Dataset

- Αξιολογήσεις 1728 αυτοκινήτων βάσει 6 χαρακτηριστικών
- Χαρακτηριστικά δείγματος: τιμή αγοράς, κόστος συντήρησης, αριθμός θυρών, χωρητικότητα ατόμων, μέγεθος πορτ μπαγκάζ, εκτιμώμενη ασφάλεια
- Κλάσεις: ακατάλληλο, κατάλληλο, καλό, πολύ καλό

- Μετατροπή κατηγορικών μεταβλητών σε σειριακούς ακεραίους
- Oversampling με αλγόριθμο SMOTE
- Stratified Train/test split: 80%/20%
- Αντιμετώπιση overfitting βάσει log-loss μεταξύ training και test set

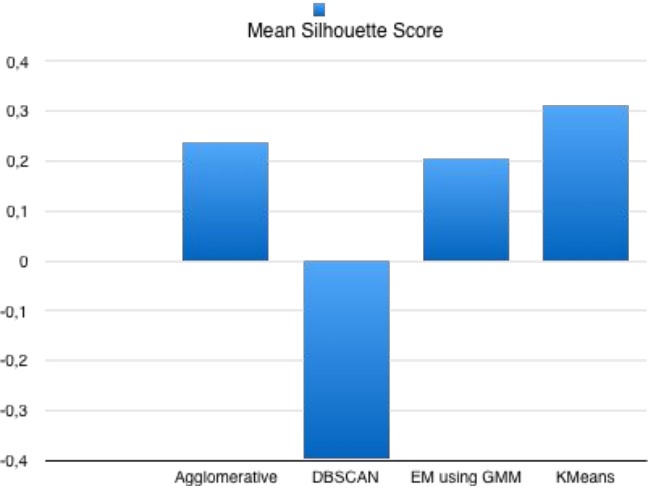
5.3 Cardiotocography Dataset

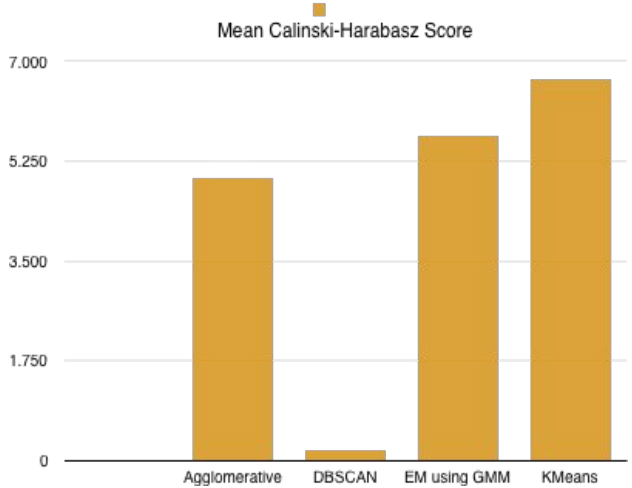
- Μετρήσεις εμβρυικού καρδιακού ρυθμού σε 2126 βρέφη (καρδιοτοκογράφημα)
- 22 χαρακτηριστικά όπως χτύποι το λεπτό, εμβρυικές κινήσεις το δευτερόλεπτο κ.ο.κ
- 10 κλάσεις που αντιπροσωπεύουν 10 διαφορετικά μοτίβα καρδιακών ρυθμών

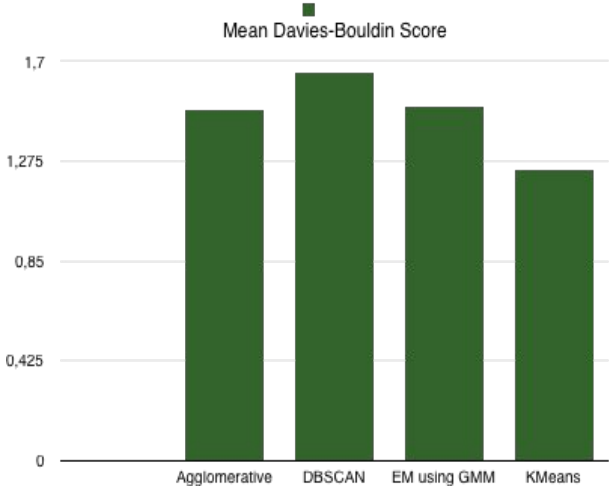
- Oversampling με αλγόριθμο SMOTE
- Stratified Train/test split: 80%/20%
- Αντιμετώπιση overfitting βάσει log-loss μεταξύ training και test set

6. Συμπεράσματα

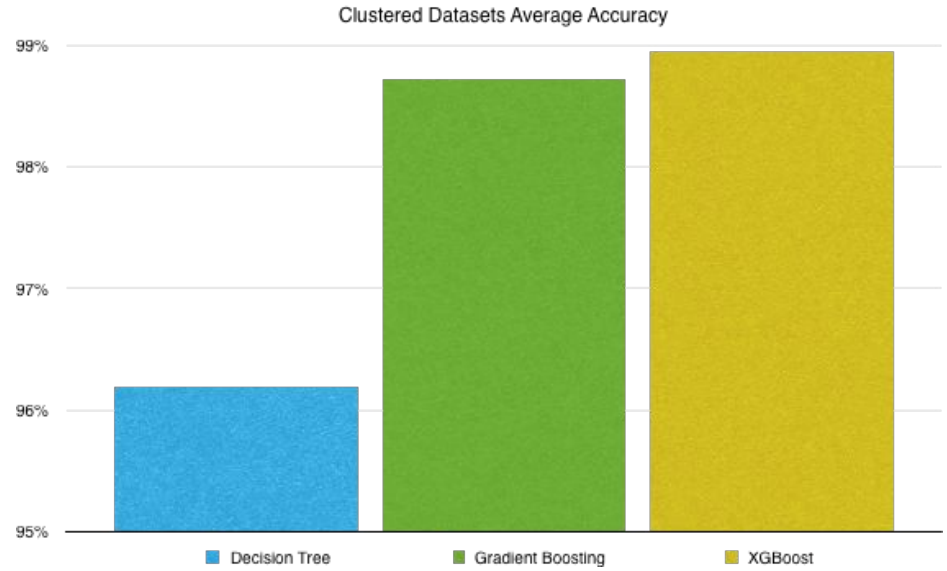
Μέσοι Όροι Clustering Metrics

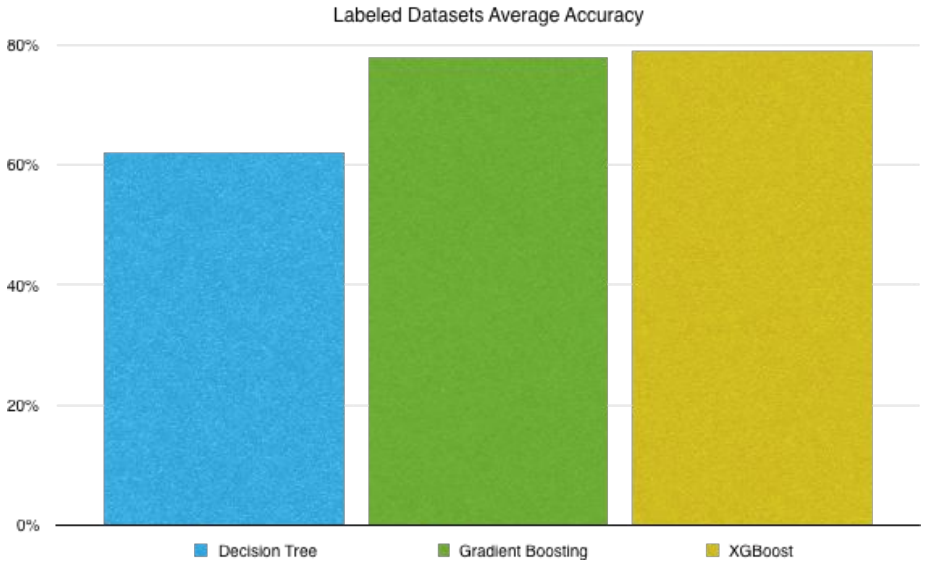






Μέσοι Όροι Accuracy





- Οι αλγόριθμοι κατηγοριοποίησης έχουν πολύ υψηλή απόδοση όταν οι κλάσεις έχουν παραχθεί με clustering

- Ένας απλός αλγόριθμος classification σε συνδυασμό με clustering έχει αρκετά υψηλότερη ακρίβεια πρόβλεψης σε σχέση με έναν πολύπλοκο αλγόριθμο classification χωρίς να έχει προηγηθεί clustering.

- Ο αριθμός των παραγόμενων clusters εξαρτάται σε τεράστιο βαθμό από το hyperparameter tuning

- *Η σύγκριση μοντέλων πάνω σε διαφορετικά σετ δεδομένων χρειάζεται επιφύλαξη λόγω της διαφορετικότητας των δεδομένων από dataset σε dataset

Ευχαριστώ