ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2022

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Η κατηγοριοποίηση (classification) είναι ίσως η πιο γνωστή και πιο δημοφιλής τεχνική εξόρυξης γνώσης (data mining).
- Όλες οι προσεγγίσεις στην εκτέλεση της κατηγοριοποίησης προϋποθέτουν γνώση των δεδομένων. Συνήθως χρησιμοποιούμε ένα σύνολο εκπαίδευσης για να καθορίσει τις συγκεκριμένες παραμέτρους που απαιτούνται από την τεχνική. Τα δεδομένα εκπαίδευσης (training data) αποτελούνται από ένα δείγμα δεδομένων εισόδου καθώς επίσης και από την κατηγοριοποίηση που έχει δοθεί σε αυτά τα δεδομένα.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Η διαδικασία της κατηγοριοποίησης, ή αλλιώς ταξινόμησης (classification) περιλαμβάνει την οργάνωση ενός συνόλου αντικειμένων (objects) που περιγράφονται από ένα σύνολο χαρακτηριστικών (attributes), σε μια σειρά από προκαθορισμένες κλάσεις (classes), χρησιμοποιώντας μεθόδους μάθησης με επίβλεψη (supervised learning methods).
- Οι τεχνικές της ταξινόμησης ή αλλιώς κατηγοριοποίησης χρησιμοποιούν κατά κανόνα ένα σύνολο εκπαίδευσης (training set), όπου όλα τα αντικείμενα είναι ήδη συνδεδεμένα με γνωστές κλάσεις. Ο αλγόριθμος ταξινόμησης μαθαίνει από αυτό το σύνολο, χρησιμοποιώντας την μάθηση αυτή για την κατασκευή ενός μοντέλου και το μοντέλο αυτό στην συνέχεια ταξινομεί νέα αντικείμενα στις κατάλληλες κλάσεις

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Παράδειγμα : Οι δάσκαλοι κατηγοριοποιούν τους μαθητές ως A,B,C,D ή F με βάση τους βαθμούς τους. Χρησιμοποιώντας απλά όρια (60, 70, 80, 90) μπορούμε να έχουμε τον παρακάτω διαχωρισμό των μαθητών σε κλάσεις:

$$
\begin{aligned}
&90 \leq \text{βαθμός} \rightarrow \text{A},\\
&80 \leq \text{βαθμός} < 90 \rightarrow \text{B},\\
&70 \leq \text{βαθμός} < 80 \rightarrow \text{C},\\
&60 \leq \text{βαθμός} < 70 \rightarrow \text{D},\\
&\text{Βαθμός} < 60 \rightarrow \text{F}
\end{aligned}
$$

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένο) το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων (κατηγοριών).

- **Μοντελοποίηση συνεχών συναρτήσεων, π.χ., πρόβλεψη άγνωστων ή απώντων τιμών**

ΟΡΙΣΜΟΙ

Ορισμός 1 : Η κατηγοριοποίηση (classification) είναι η διαδικασία η οποία απεικονίζει ένα σύνολο δεδομένων σε προκαθορισμένες ομάδες. Τις ομάδες αυτές συχνά τις καλούμε κατηγορίες ή κλάσεις.

Ορισμός 2 : Έστω μια Βάση Δεδομένων $DB = \{t_1, t_2, \ldots, t_n\}$ πλειάδων (στοιχείων, εγγραφών) και ένα σύνολο από κατηγορίες $C = \{C_1, C_2, \ldots, C_m\}$. Το πρόβλημα της κατηγοριοποίησης είναι ο ορισμός μιας απεικόνισης

$$f: DB \rightarrow C$$

όπου κάθε t_i τοποθετείται σε μια κατηγορία. Μια κατηγορία ή κλάση C_j, περιέχει ακριβώς αυτές τις πλειάδες όπου έχουν απεικονιστεί σε αυτή, δηλαδή

$$C_j = \{t_i \mid f(t_i) = C_j, 1 \leq i \leq n, \text{ και } t_i \in DB\}.$$

ΟΡΙΣΜΟΙ

- Ουσιαστικά, η κατηγοριοποίηση διαμερίζει τη D σε κλάσεις ισοδυναμίας.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Πρέπει να υπογραμμιστεί ότι οι κατηγορίες είναι προκαθορισμένες, δεν επικαλύπτονται και διαμερίζουν ολόκληρη την Βάση Δεδομένων. Κάθε στοιχείο της Βάσης Δεδομένων τοποθετείται σε ακριβώς μια κατηγορία. Οι κατηγορίες που υπάρχουν σε ένα πρόβλημα κατηγοριοποίησης είναι στην πραγματικότητα κλάσεις ισοδυναμίας (equivalence classes).

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Υπάρχουν τρεις βασικές μέθοδοι που χρησιμοποιούνται για να λύσουν το πρόβλημα της κατηγοριοποίησης:

Καθορισμός των ορίων: Η κατηγοριοποίηση εκτελείται με διαίρεση του χώρου της εισόδου των εν δυνάμει πλειάδων της Βάσης Δεδομένων σε περιοχές όπου κάθε περιοχή συνδέεται με μια κατηγορία

Χρήση κατανομών πιθανότητας: Για κάθε κατηγορία που δίνεται C_j $P(t_i | C_j)$ είναι η συνάρτηση κατανομής πιθανότητας (probability distribution function) για την κατηγορία υπολογισμένη σε ένα σημείο, t_i. Αν η πιθανότητα εμφάνισης κάθε πιθανότητας $P(C_j)$, είναι γνωστή τότε $P(C_j)$ $P(t_i | C_j)$ είναι η εκτίμηση της πιθανότητας ότι η t_i ανήκει στην κατηγορία C_j

.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Χρήση εκ των υστέρων πιθανοτήτων: Με δεδομένη μια τιμή δεδομένων t_i, θέλουμε να καθορίσουμε την πιθανότητα για την οποία η t_i, ανήκει στην κατηγορία C_j. Αυτό υποδηλώνεται με το $P(C_j|t_i)$ που ονομάζεται εκ των υστέρων πιθανότητα (posterior probability). Μια προσέγγιση κατηγοριοποίησης είναι ο καθορισμός της εκ των υστέρων πιθανότητας για κάθε κατηγορία και στη συνέχεια η τοποθέτηση των πλειάδων στην κατηγορία με τη μεγαλύτερη πιθανότητα.

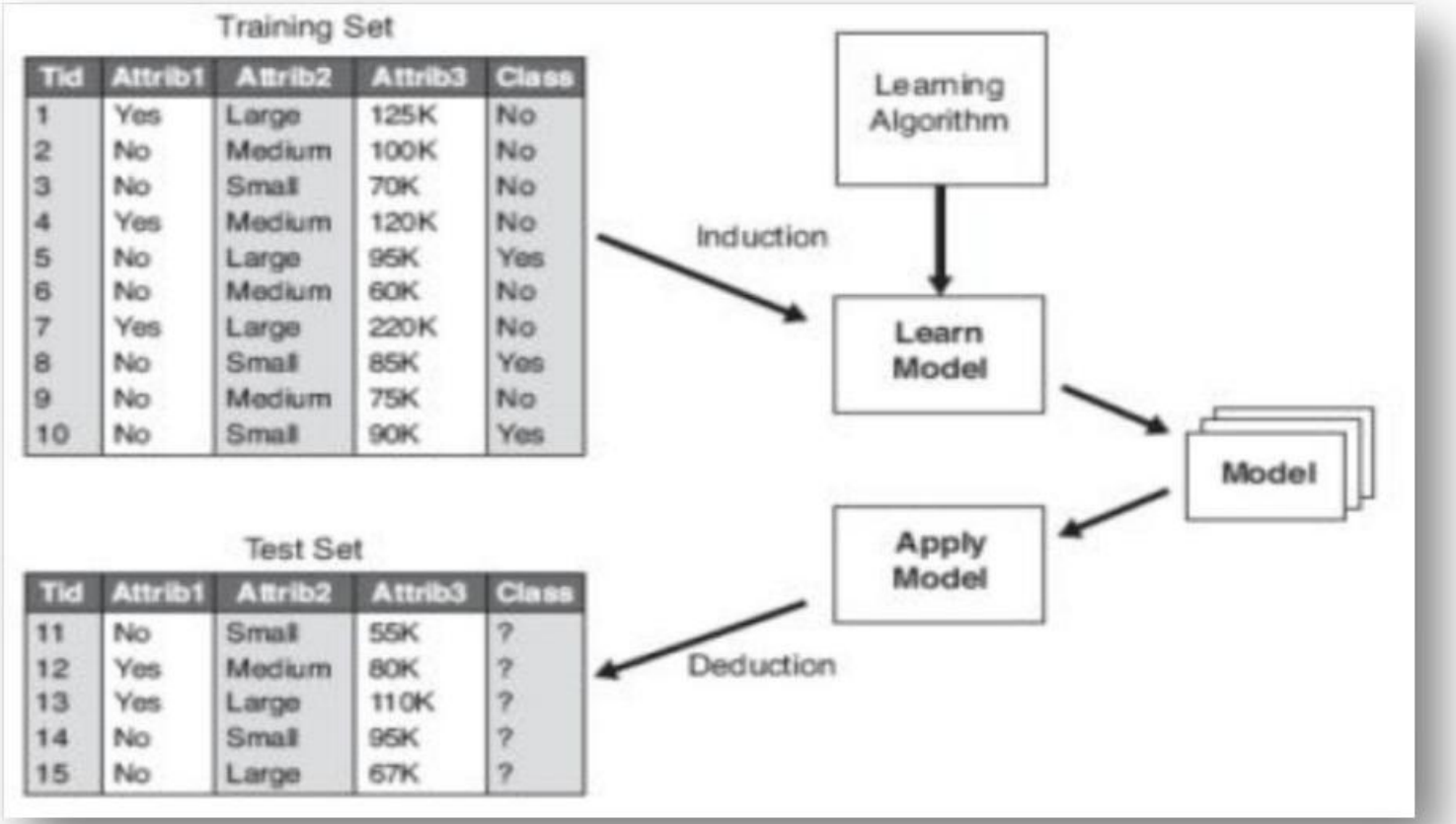
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Στην πράξη μια διαδικασία κατηγοριοποίησης μπορεί να οριστεί ως η εκτέλεση δύο συγκεκριμένων βημάτων:

1. Δημιουργία μοντέλου βασιζόμενου σε δεδομένα εκπαίδευσης
2. Εφαρμογή του μοντέλου στο σύνολο των δεδομένων

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ κατηγοριοποιημένα παραδείγματα.

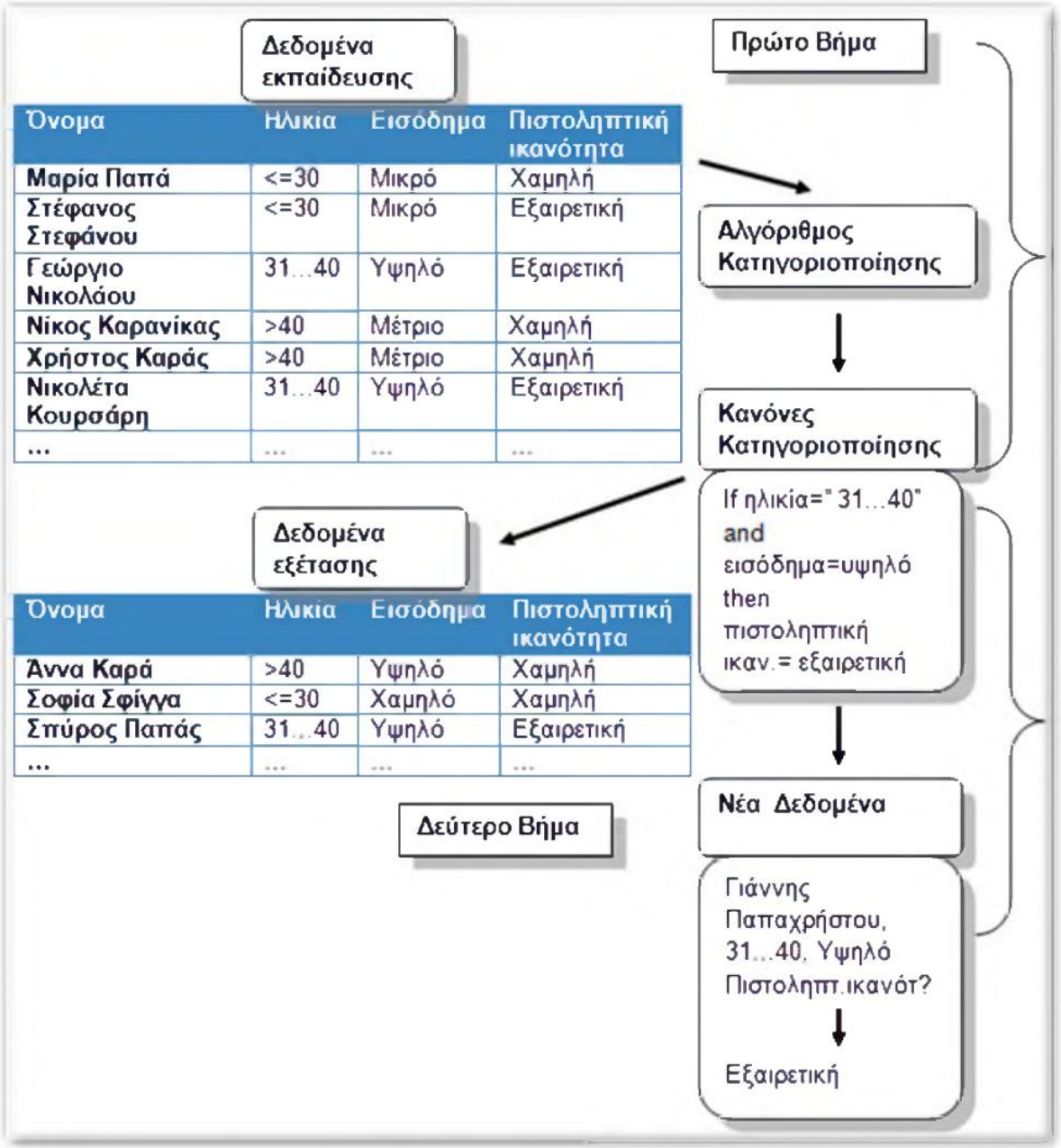
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ



ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Αρχικά, πρέπει να δοθεί ένα training set το οποίο περιέχει εγγραφές των οποίων οι ετικέτες κατηγορίας είναι σωστές. Το training set χρησιμοποιείται για να φτιάξει το μοντέλο ταξινόμησης, το οποίο μετέπειτα εφαρμόζεται στο test set, όπου περιέχει εγγραφές των οποίων οι ετικέτες κατηγορίας είναι άγνωστες. Η διαδικασία που ακολουθείται δηλαδή έχει να κάνει με την παραγωγή της test set με άγνωστες ετικέτες από το αρχικό training set οι οποίες πρέπει να προβλεφθούν από κάποιον αλγόριθμο με όσο το δυνατό μεγαλύτερη επιτυχία.

διαδικασία κατηγοριοποίησης



ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Η αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης βασίζεται στον αριθμό των εγγραφών του test set που προβλέφθηκαν σωστά ή λάθος από τον ταξινομητή. Για να είναι ευκολότερη η σύγκριση των αποδόσεων διαφορετικών μοντέλων χρησιμοποιούνται δύο δείκτες επίδοσης, η ακρίβεια (accuracy) και η αποτίμηση του σφάλματος (error rate)

$$\text{Ακρίβεια} = \frac{\text{Σωστές προβλέψεις}}{\text{Σύνολο προβλέψεων}}$$

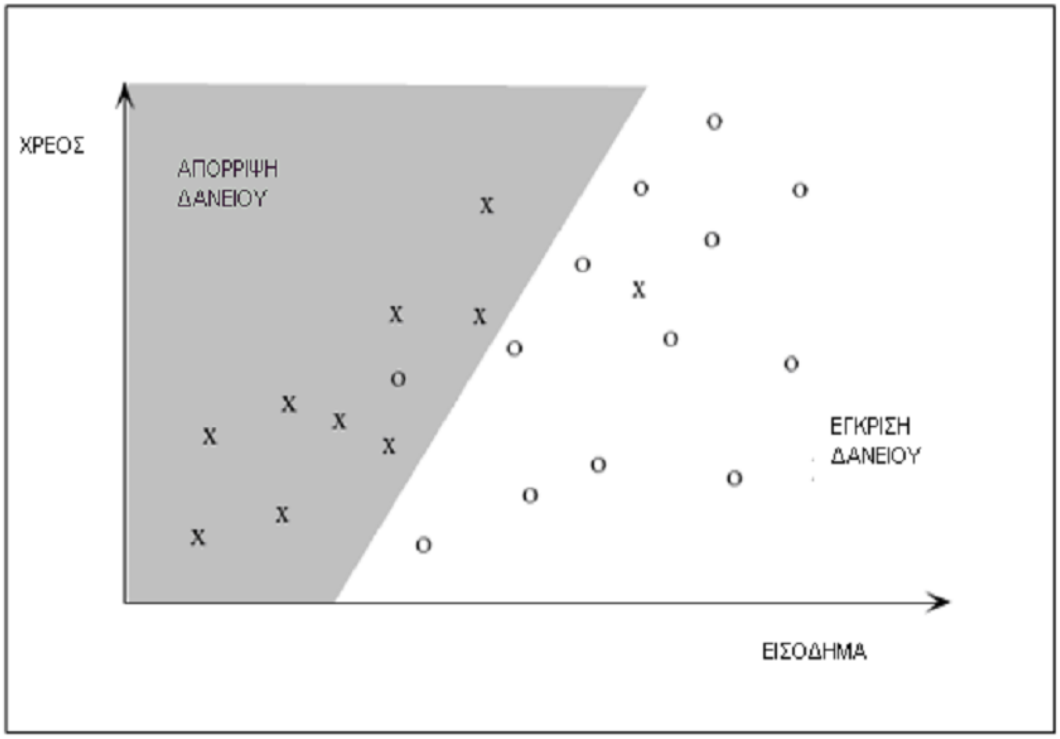
$$\text{Αποτίμηση σφάλματος} = \frac{\text{Λάθος προβλέψεις}}{\text{Σύνολο προβλέψεων}}$$

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Έτσι τελικά ο ταξινομητής με τη μεγαλύτερη ακρίβεια και το μικρότερη αποτίμηση σφάλματος είναι ορθότερος και πιο αποτελεσματικός, δηλαδή μπορεί και κάνει καλύτερες προβλέψεις.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

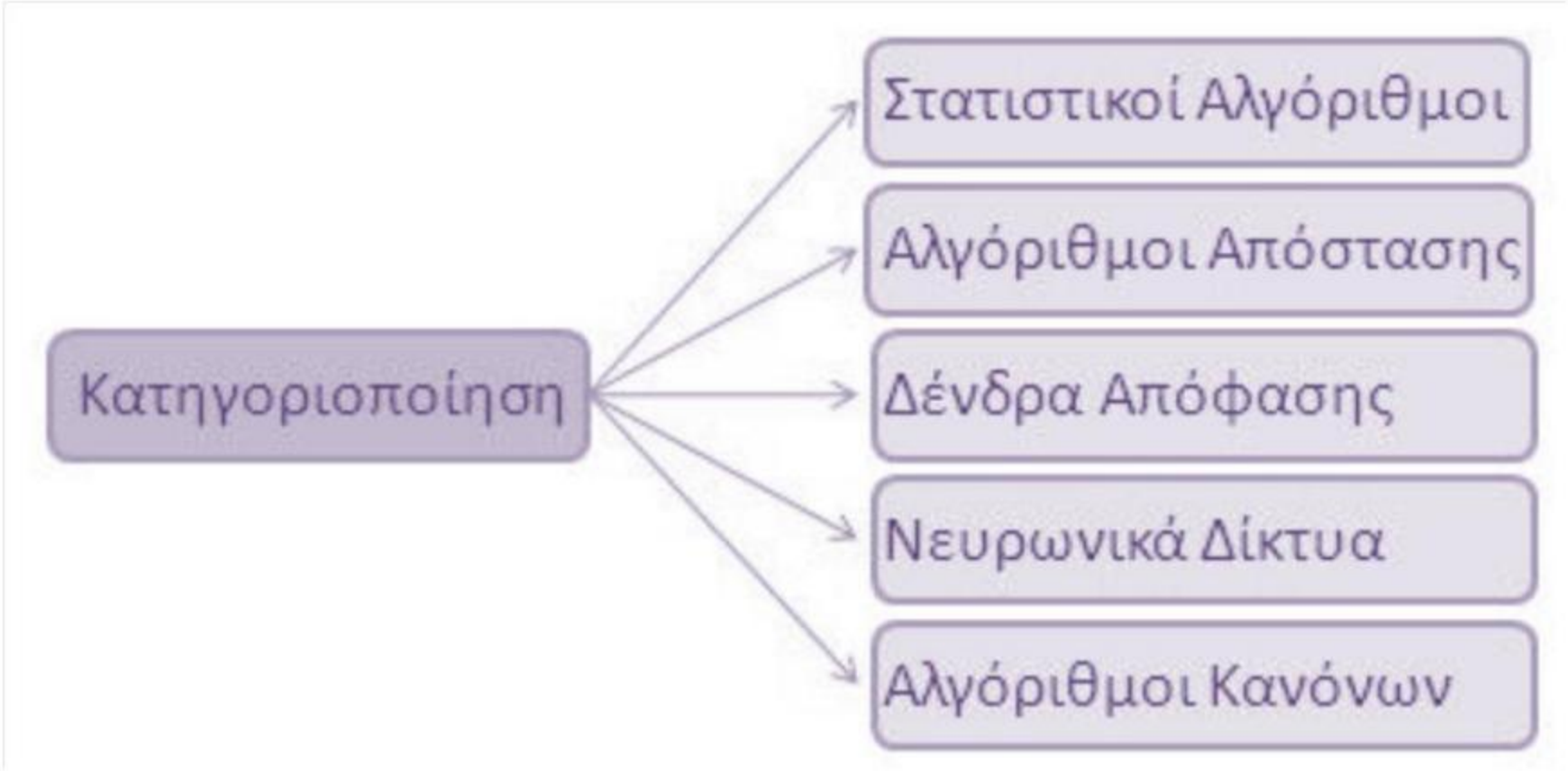
Στην παρακάτω εικόνα έχουμε έναν απλό διαχωρισμό των στοιχείων δανείου σε δύο περιοχές κατηγοριών. Η τράπεζα πιθανώς να θελήσει να χρησιμοποιήσει τις περιοχές ταξινόμησης για να αποφασίσει, εάν θα δοθεί δάνειο ή όχι στους μελλοντικούς υποψηφίους.



ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ κατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να οργανώσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί. Στις περισσότερες περιπτώσεις, υπάρχει ένα περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο βασικές κατηγορίες. Η πρώτη χρησιμοποιεί τα λεγόμενα δέντρα απόφασης (decision trees) ενώ η δεύτερη τα νευρωνικά δίκτυα (neural networks).

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ



ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Διαδικασία

Κατασκευή μοντέλου: περιγράφοντας ένα σύνολο προκαθορισμένων κατηγοριών

- Καθεμια από τις πλειάδες υποτίθεται ότι ανήκει σε μια προκαθορισμένη κατηγορία, όπως καθορίζεται από το γνώρισμα κατηγορίας(κλάσης) (supervised learning)
- Χρήση ενός training set για κατασκευήμοντέλου
- Αναπαράσταση μοντέλου σαν classification rules, decision trees, ή μαθηματικές εξισώσεις

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Χρήση μοντέλου: για κατηγοριοποίηση άγνωστων αντικειμένων

- Υπολογισμός ακρίβειας του μοντέλου κάνοντας χρήση ενός test set. Η γνωστή ετικέτα του test sample συγκίνεται με το αποτέλεσμα κατηγοριοποίησης του μοντέλου
- Η τιμή της ακρίβειας είναι το ποσοστό των test set samples που κατηγοριοποιήθηκαν σωστά απο το μοντέλο
- το test set είναι ανεξάρτητο από το training set, αλλιώς μπορεί να συμβεί overfitting

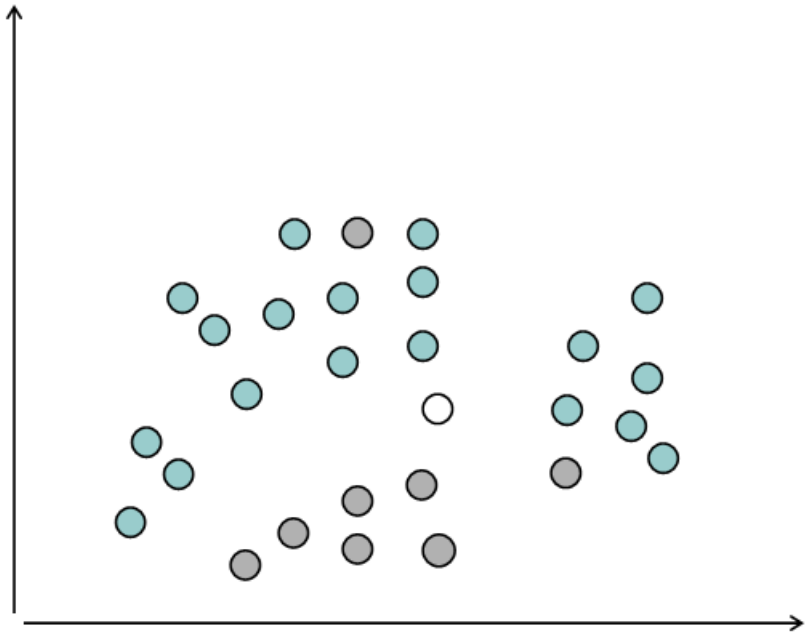
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- **Πρόβλημα κατηγοριοποίησης**:

δοσμένου ενός συνόλου δεδομένων εκπαίδευσης(Ν εγγραφές, με Μ γνωρίσματα, και ένα γνώρισμα ετικέτας) βρες κανόνες για την πρόβλεψη της ετικέτας(κλάσης) μιας νέας εγγραφής

Εκμάθηση μιας τεχνικής να προβλέπει την κλάση ενός στοιχείου επιλέγοντας από προκαθορισμένες τιμές

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ



Προσεγγίσεις:

- στατιστικές μέθοδοι
- δένδρα αποφάσεων
- νευρωνικά δίκτυα

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Η εκτίμηση της απόδοσης ενός μοντέλου κατηγοριοποίησης, βασίζεται στο πλήθος των εγγραφών ελέγχου που έχουν προβλεφθεί σωστά και λανθασμένα από το μοντέλο. Αυτές που μετρήσεις τοποθετούνται σε έναν πίνακα γνωστό ως μήτρα σύγχυσης (confusion matrix) .
- Ο παρακάτω πίνακας παριστάνει τη μήτρα σύγχυσης για ένα πρόβλημα δυαδική κατηγοριοποίησης .

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Κάθε καταχώρηση f_{ij}, του πίνακα δηλώνει το πλήθος των εγγραφών της κατηγορίας i, που προβλέφθηκε ότι ανήκει στην κατηγορία j

μέτρα απόδοσης

ακρίβεια

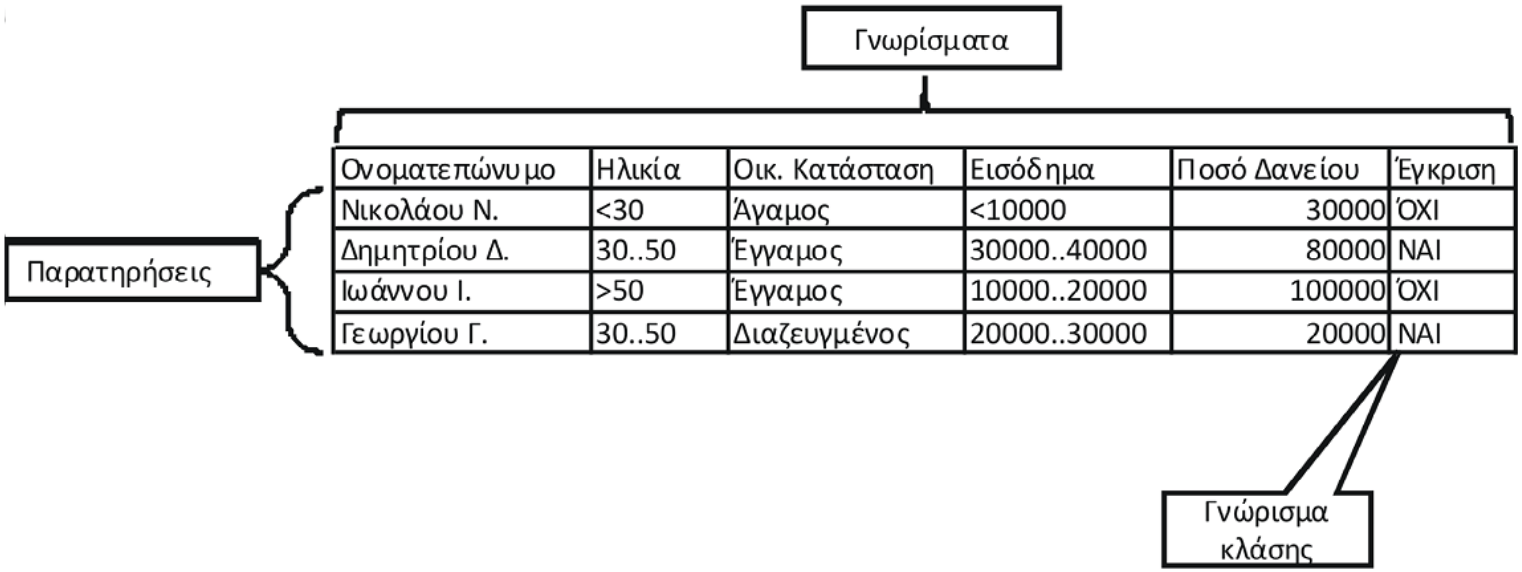
		Predicted Class	
		$Class = 1$	$Class = 0$
Actual Class	$Class = 1$	f_{11}	f_{10}
	$Class = 0$	f_{01}	f_{00}

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

ρυθμό σφάλματος

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}.$$

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ



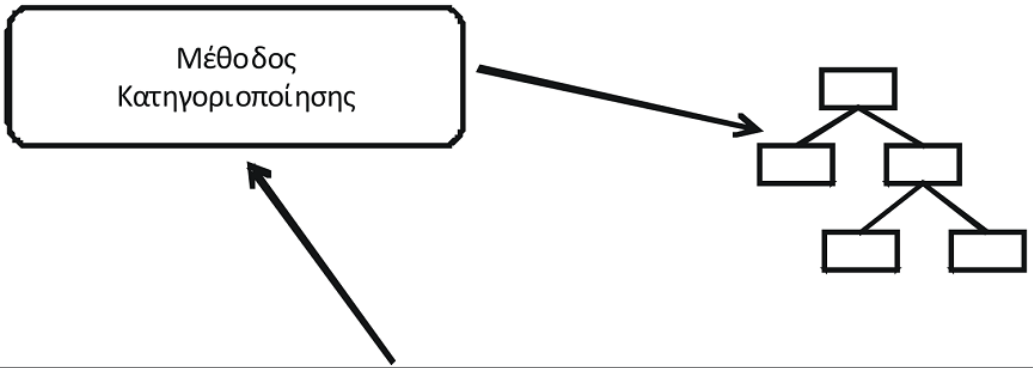
Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Νικολάου Ν.	<30	Άγαμος	<10000	30000	ΟΧΙ
Δημητρίου Δ.	30..50	Έγγαμος	30000..40000	80000	ΝΑΙ
Ιωάννου Ι.	>50	Έγγαμος	10000..20000	100000	ΟΧΙ
Γεωργίου Γ.	30..50	Διαζευγμένος	20000..30000	20000	ΝΑΙ

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Στάδια κατηγοριοποίησης

Επιβλεπόμενη μάθηση. Στο στάδιο αυτό, μια μέθοδος κατηγοριοποίησης αναλύει ένα σύνολο δεδομένων. Η μέθοδος θα ανακαλύψει σχέσεις μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Το αποτέλεσμα αυτής της επεξεργασίας είναι η κατασκευή ενός μοντέλου.

Μια μέθοδος κατηγοριοποίησης επεξεργάζεται ένα σύνολο εκπαίδευσης, το οποίο περιέχει στοιχεία δανείων και κατασκευάζεται ένα μοντέλο

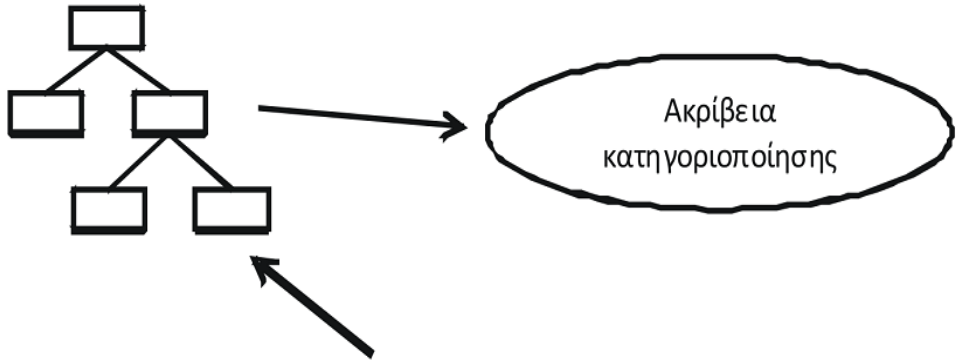


Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Νικολάου Ν.	<30	Άγαμος	<10000	30000	ΟΧΙ
Δημητρίου Δ.	30..50	Έγγαμος	30000..40000	80000	ΝΑΙ
Ιωάννου Ι.	>50	Έγγαμος	10000..20000	100000	ΟΧΙ
Γεωργίου Γ.	30..50	Διαζευγμένος	20000..30000	20000	ΝΑΙ

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Επικύρωση μοντέλου Στο στάδιο αυτό δοκιμάζεται η ακρίβεια του μοντέλου, η ικανότητα του δηλαδή να προβλέπει σωστά την κλάση των παρατηρήσεων. Το μοντέλο τροφοδοτείται με παρατηρήσεις, των οποίων η κλάση είναι γνωστή. Αναλύοντας τα στοιχεία των ανεξάρτητων μεταβλητών κάθε παρατήρησης, το μοντέλο προβλέπει την κλάση της παρατήρησης και στη συνέχεια συγκρίνεται η πρόβλεψη του μοντέλου με την πραγματική τιμή της κλάσης.

Το μοντέλο τροφοδοτείται με περιπτώσεις δανείων διαφορετικές από αυτές που χρησιμοποιήθηκαν για την εκπαίδευση. Για κάθε δάνειο, το μοντέλο πραγματοποιεί μια πρόβλεψη και η πρόβλεψη αυτή συγκρίνεται με την πραγματική απόφαση έγκρισης ή απόρριψης του δανείου.

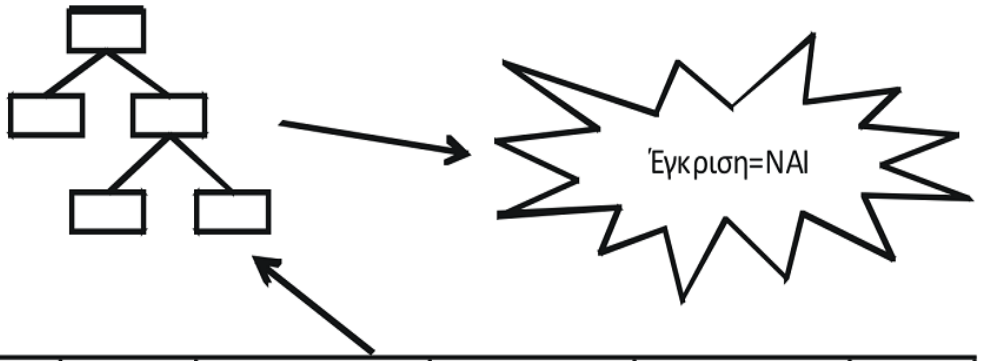


Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Κωνσταντίνου Κ.	<30	Άγαμος	<10000	40000	ΟΧΙ
Παναγιώτου Π.	30..50	Έγγαμος	30000..40000	100000	ΝΑΙ

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Χρήση του μοντέλου. Το μοντέλο, αφού εκπαιδευτεί και επικυρωθεί, χρησιμοποιείται για τη διατύπωση προβλέψεων. Μια νέα παρατήρηση, της οποίας η κλάση είναι άγνωστη, εισάγεται στο μοντέλο. Το μοντέλο χρησιμοποιώντας τις τιμές των ανεξάρτητων μεταβλητών υπολογίζει την τιμή της κλάσης.

το μοντέλο χρησιμοποιείται για την πρόβλεψη έγκρισης νέων δανείων



Ονοματεπώνυμο	Ηλικία	Οικ. Κατάσταση	Εισόδημα	Ποσό Δανείου	Έγκριση
Χρήστου Χ.	30..50	Έγγαμος	30000..40000	90000	

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Η επίδοση των αλγορίθμων εξετάζεται με την εκτίμηση της ακρίβειας (accuracy) της κατηγοριοποίησης, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Η εκτίμηση της ακρίβειας είναι ένα πολύ σημαντικό ζήτημα στο χώρο της κατηγοριοποίησης αφού κάτι τέτοιο μας δείχνει το πόσο καλά ανταποκρίνεται ο αλγόριθμος μας για δεδομένα με τα οποία δεν έχει εκπαιδευτεί. Η εκτίμηση της ακρίβειας είναι επίσης θεμιτή αφού μας επιτρέπει την σύγκριση των διαφόρων αλγορίθμων κατηγοριοποίησης.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Ταχύτητα (Speed): Κόστος υπολογισμός (συμπεριλαμβανομένου την παραγωγή και τη χρήση του μοντέλου)

Ανθεκτικότητα (Robustness): Σωστή πρόβλεψη με ελλιπή δεδομένα ή δεδομένα με θόρυβο

Επεκτασιμότητα (Scalability): Αποδοτική κατασκευή του μοντέλου δοθέντος μεγάλη ποσότητα δεδομένων (μπορεί να εκτιμηθεί μετρώντας τις λειτουργίες I/O που απαιτεί ο αλγόριθμος)

Ερμηνευσιμότητα (Interpretability): Επίπεδο κατανόησης και γνώση που παρέχεται από το μοντέλο. (Μπορεί να εκτιμηθεί μετρώντας το πόσο πολύπλοκο είναι το μοντέλο π.χ. αριθμός κόμβων στα δένδρα απόφασης, αριθμός επιπέδων στα νευρωνικά δίκτυα κ.ά.)

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Εκτίμηση απόδοσης κατηγοριοποίησης

Μπορούμε να χρησιμοποιήσουμε ένα σύνολο δεδομένων αρχικά για να εκπαιδεύσουμε τον αλγόριθμο μας και στην συνέχεια να χρησιμοποιήσουμε το ίδιο σύνολο δεδομένων για να εκτιμήσουμε την ακρίβεια του αλγορίθμου. Μια τέτοια επιλογή θα μας οδηγούσε σε μια πολύ αισιόδοξη εκτίμηση της ακρίβειας αφού ο αλγόριθμος εκπαιδεύεται αλλά και δοκιμάζεται με το ίδιο σύνολο δεδομένων.

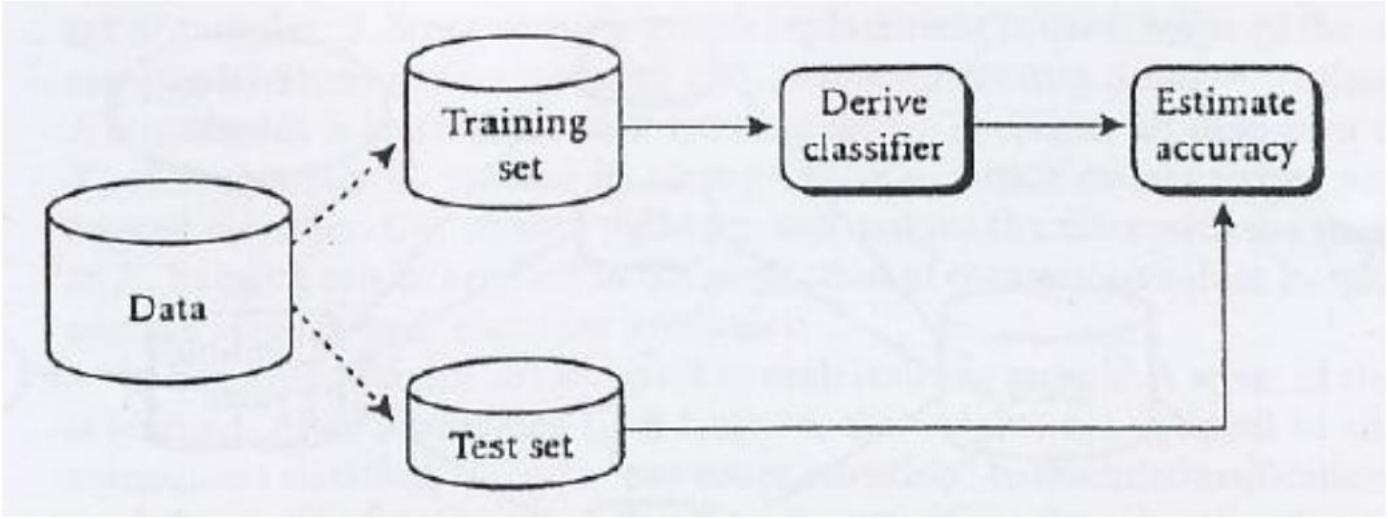
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Άλλος ένας τρόπος εκτίμησης της ακρίβειας ενός αλγορίθμου κατηγοριοποίησης είναι η **μέθοδος της κατακράτησης** (**holdout method**)

Χρησιμοποιώντας αυτή την μέθοδο, το σύνολο δεδομένων που έχουμε στην διάθεση μας χωρίζεται με τυχαίο τρόπο σε δυο ανεξάρτητα σύνολα δεδομένων. Το πρώτο ονομάζεται σύνολο δεδομένων εκπαίδευσης και χρησιμοποιείται για την εκπαίδευση του αλγορίθμου κατηγοριοποίησης και το δεύτερο ονομάζεται σύνολο δεδομένων δοκιμής που χρησιμοποιείται για την δοκιμή του αλγορίθμου και την εκτίμηση της ακρίβειας.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Η τελική εκτίμηση της ακρίβειας είναι μέσος όρος των εκτιμήσεων ακρίβειας των επαναλήψεων.



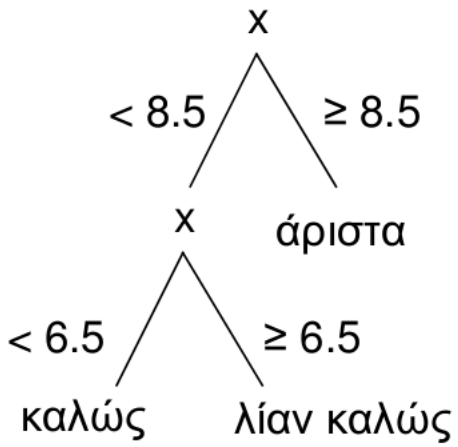
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

- Ένας λίγο πιο σύνθετος τρόπος εκτίμησης της απόδοσης είναι η k - διπλή διασταυρούμενη επικύρωση (k-fold cross-validation). Σύμφωνα με αυτήν, το αρχικό σύνολο δεδομένων αρχικά χωρίζεται σε k υποσύνολα, $S_1,S_2,\ldots, S_k$ κάθε ένα από τα οποία είναι ίδιου μεγέθους.. Η εκπαίδευση και η δοκιμή εκτελείται k φορές. Στην i επανάληψη, το υποσύνολο S_i παίζει τον ρόλο το συνόλου δοκιμής, ενώ τα υπόλοιπα k-1 υποσύνολα χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου. Αυτό σημαίνει ότι στην πρώτη επανάληψη, το S_1 λειτουργεί σαν σύνολο δοκιμής ενώ τα $\{S_2,S_3,\ldots,S_k\}$ σαν σύνολο εκπαίδευσης. Αντίστοιχα στην δεύτερη επανάληψη, το S_2 λειτουργεί σαν σύνολο δοκιμής, ενώ τα $\{S_1,S_3,\ldots, S_k\}$ σαν σύνολο εκπαίδευσης. Η ακρίβεια διαιρώντας το συνολικό αριθμό τον σωστών κατηγοριοποιήσεων με τον αριθμό των πλειάδων του αρχικού συνόλου δεδομένων. Όπως γίνεται εύκολα κατανοητό, αυτή η προσέγγιση απαιτεί k φορές περισσότερο χρόνο από την μέθοδο της κατακράτησης.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

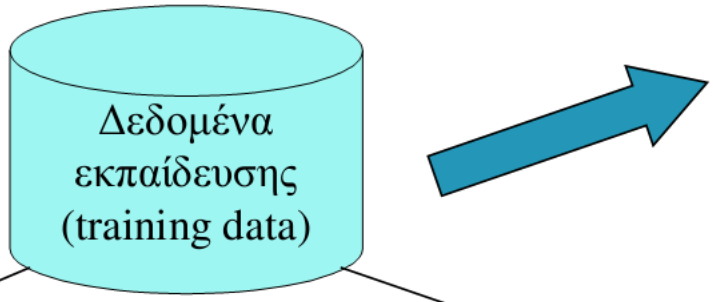
Βαθμολογία πτυχίου

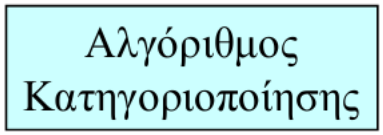
- If $x \geq 8.5$ then
  grade = «άριστα».
- If $6.5 \leq x < 8.5$ then
  grade = «λίαν καλώς».
- If $x < 6.5$ then
  grade = «καλώς».

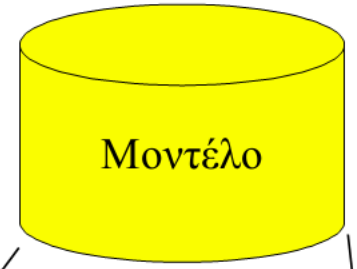


ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

1ο βήμα: Δημιουργία μοντέλου





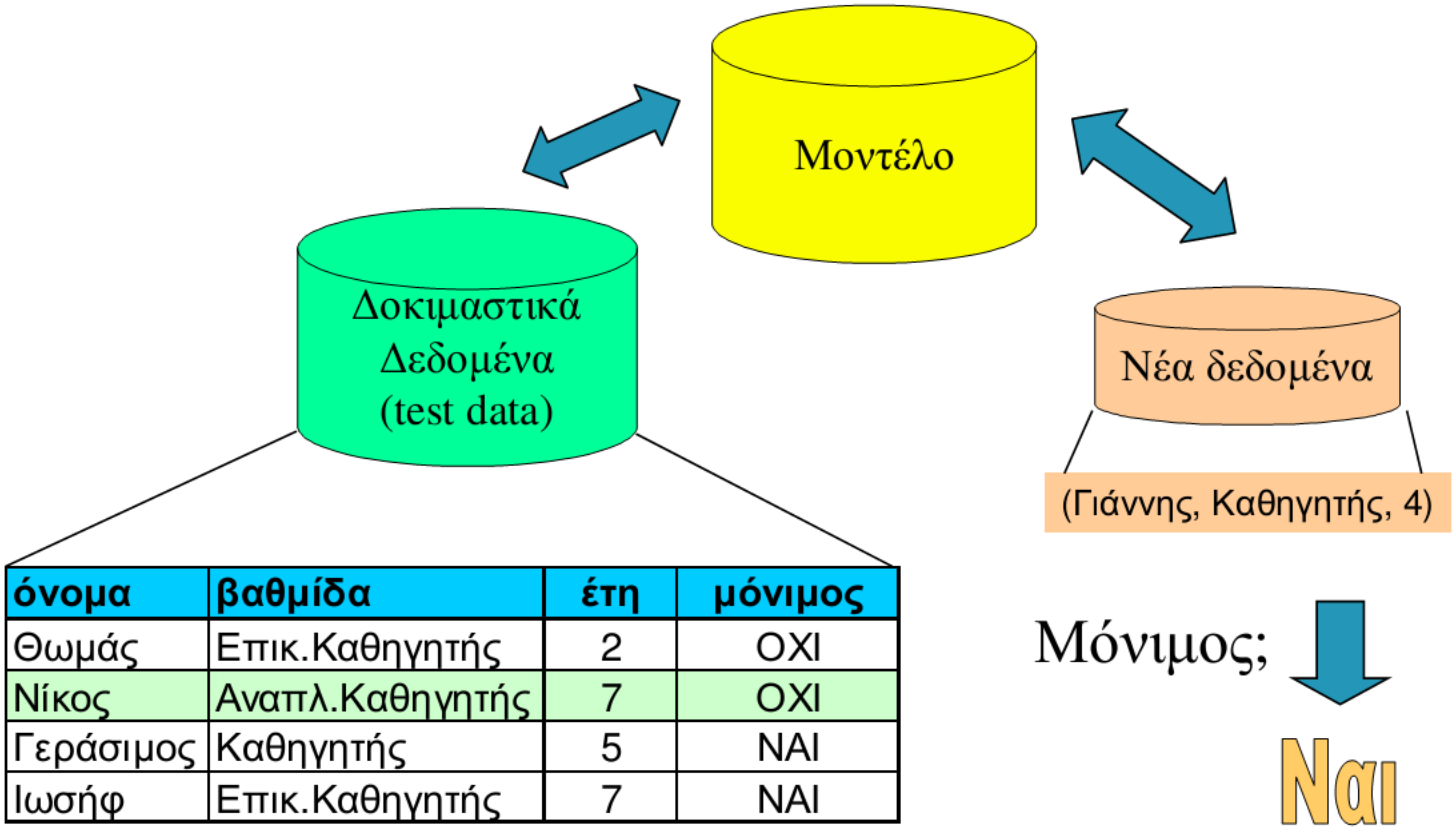


όνομα	βαθμίδα	έτη	μόνιμος
Μιχάλης	Επικ.Καθηγητής	3	ΟΧΙ
Νίκος	Επικ.Καθηγητής	7	ΝΑΙ
Βασίλης	Καθηγητής	2	ΝΑΙ
Δημήτρης	Αναπλ.Καθηγητής	7	ΝΑΙ
Γιώργος	Επικ.Καθηγητής	6	ΟΧΙ
Κώστας	Αναπλ.Καθηγητής	3	ΟΧΙ

IF βαθμίδα = 'Καθηγητής'
OR έτη > 6
THEN μόνιμος = 'ΝΑΙ'

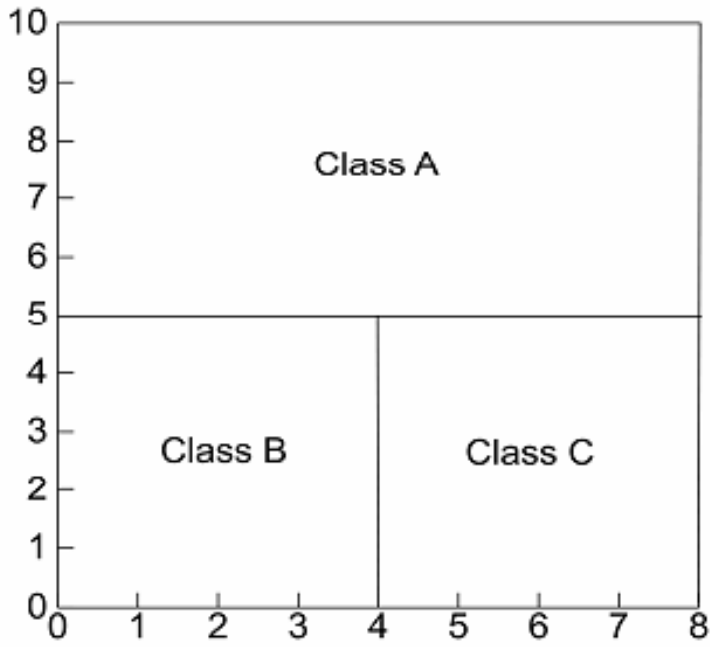
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

2ο βήμα: Εφαρμογή μοντέλου



όνομα	βαθμίδα	έτη	μόνιμος
Θωμάς	Επικ.Καθηγητής	2	ΟΧΙ
Νίκος	Αναπλ.Καθηγητής	7	ΟΧΙ
Γεράσιμος	Καθηγητής	5	ΝΑΙ
Ιωσήφ	Επικ.Καθηγητής	7	ΝΑΙ

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ



με βάση τη διαμέριση

με βάση την απόσταση