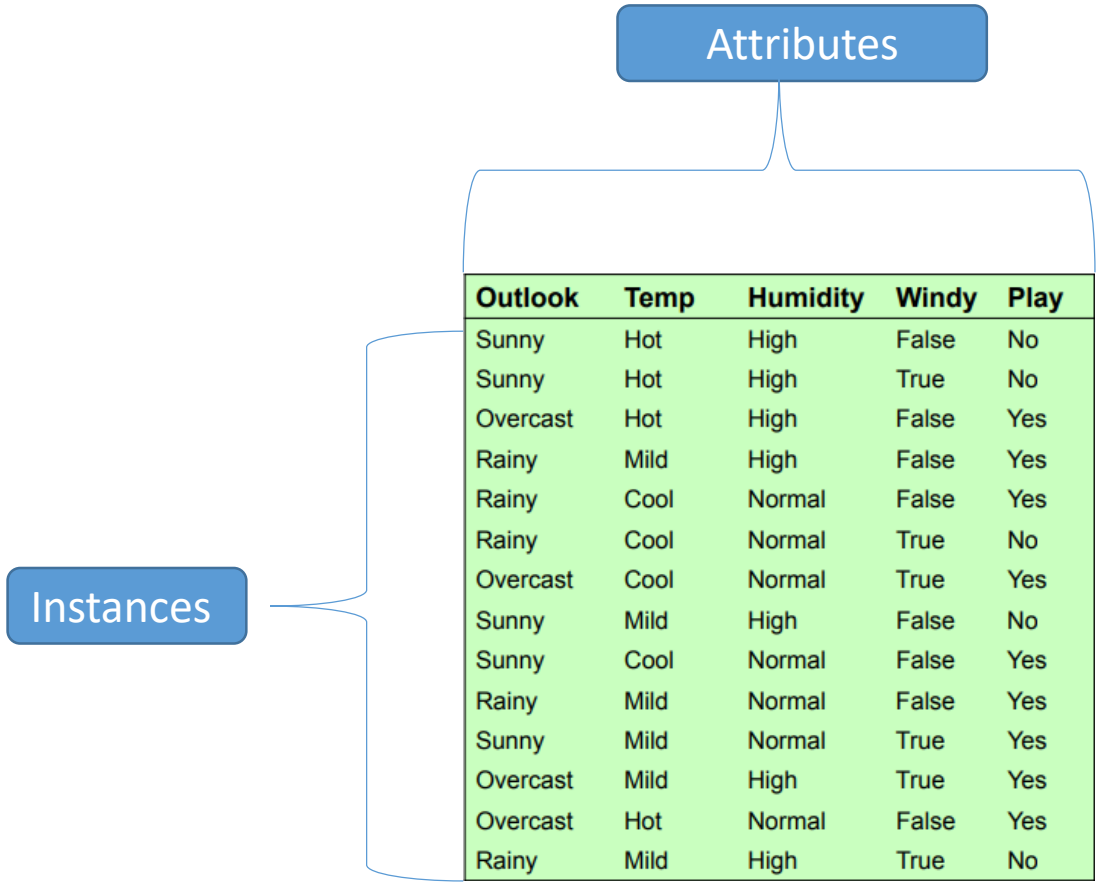
Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Επεξεργασία και διαχείριση δεδομένων και μεταβλητών

Περιγραφή των δεδομένων



Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Attributes
1. Outlook
2. Temp
3. Humidity
4. Windy
5. Play

Περιγραφή των δεδομένων

- Τα δεδομένα περιγράφονται υπό τη μορφή πινάκων με n γραμμές και d στήλες.
- Οι γραμμές ονομάζονται εγγραφές, **instances,…**
- Οι στήλες ονομάζονται ιδιότητες, χαρακτηριστικά,**attributes**, properties, features, variables, fields…

$$\mathbf{D} = \begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}$$

Τύποι ιδιοτήτων(attribute)

- Αριθμητικές (**Numeric**)

  Οι μεταβλητές είναι αριθμοί. Οι τιμές μπορεί να είναι

  - Συνεχόμενες (**Continuous**). Σε αυτή την κατηγορία ανήκουν οι μεταβλητές που μπορούν να δεχθούν οποιαδήποτε αριθμητική τιμή
  - Διακριτές (**Discrete**). Οι διακριτές τιμές έχουν προκύψει από συνεχόμενες τιμές τις οποίες έχουμε χωρίσει σε τμήματα. Για παράδειγμα αν έχουμε την μεταβλητή ύψος με τιμές από 1.60 μέχρι 1.93 τότε μπορούμε να δημιουργήσουμε 3 ομάδες. Οι οποίες θα είναι [1.60-1.70], [1.70-1.80], [1.80-1.93]. Οπότε αντί να έχουμε το ύψος κάθε ανθρώπου, θα έχουμε την ομάδα στην οποία ανήκει.
  - Το είδος του αλγορίθμου και οι περιορισμοί που έχει μας οδηγούν στην μετατροπή ή όχι μιας μεταβλητής από συνεχόμενη σε διακριτή.

Τύποι ιδιοτήτων (attribute)

Μεταβλητές από ένα συγκεκριμένο σύνολο τιμών (**Nominal**).

- Δέχονται σαν τιμές περιγραφές ή ονόματα.
- Για παράδειγμα η μεταβλητή Φύλο μπορεί να πάρει 2 τιμές {M, F} ή

η μεταβλητή σπουδές 3 τιμές { Πτυχίο, Μεταπτυχιακό, Διδακτορικό}

Ordinal μεταβλητές. Μοιάζουν με τις nominal μεταβλητές με την διαφορά ότι υπάρχει ιεραρχία μεταξύ των τιμών που μπορούν να αντιστοιχηθούν στις μεταβλητές.

- Ορίζεται διάταξη στις τιμές και όχι απόσταση μεταξύ τους
- Για παράδειγμα η μεταβλητή για την θερμοκρασία {ζεστό, μέτριο, κρύο}.

Σε αυτή την περίπτωση ισχύει ότι ζεστό >μέτριο>κρύο αλλά δεν μπορούμε να κάνουμε πρόσθεση και αφαίρεση.

Τύποι ιδιοτήτων (attribute)

Interval ιδιότητες. Σε αυτή την περίπτωση οι τιμές είναι διατεταγμένες αλλά μετριούνται με κάποια συγκεκριμένη κλίμακα (π.χ. η μεταβλητή θερμοκρασία, την οποία μετράμε σε βαθμούς κελσίου).

Αναλογικές (**ratio**) ιδιότητες. Οι μεταβλητές που έχουν αναλογικές ιδιότητες είναι αυτές για τις οποίες έχει οριστεί το σημείο μηδέν. Για παράδειγμα όταν μετράμε την απόσταση. Τότε αυτή η μεταβλητή μπορεί να είναι τύπου ratio.

Binary ιδιότητες. Σε αυτή την περίπτωση η μεταβλητή μπορεί να πάρει 2 τιμές. True ή False

Γιατί χρειαζόμαστε διαφορετικούς τύπους χαρακτηριστικών;

- Κάποιοι Αλγόριθμοί χρειάζονται συγκεκριμένους τύπους δεδομένων, οπότε φροντίζουμε να έχουμε τα δεδομένα στην κατάλληλη μορφή.
- Αναδεικνύουν καλύτερα πιθανές λύσεις σε προβλήματα. Για παράδειγμα Επιδόσεις _Αυτοκινήτου= Καλές (δεν βοηθάει στην ανάλυση..).
- Αλλά αν είχαμε ότι Ταχύτητα αυτοκινήτου=300 km/h και κατανάλωση καυσίμων 7 lt στα 100 Km τότε θα είχαμε πιο περιγραφικά δεδομένα για το τι σημαίνει Επιδόσεις Αυτοκινήτου= Καλές
- Βοηθούν στην εξαγωγή καλύτερων απαντήσεων

ΤΕΛΟΣ