



Πανεπιστήμιο Πειραιώς
ΕΚΕΦΕ “ΔΗΜΟΚΡΙΤΟΣ”
Δ.Π.Μ.Σ. στην Τεχνητή Νοημοσύνη
Τμήμα Ψηφιακών Συστημάτων Πανεπιστημίου Πειραιώς
Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών **ΕΚΕΦΕ**
«Δημόκριτος»

Ταξινόμηση δειγμάτων νερού επί τη βάσει της ποσιμότητας

Εργασία στο μάθημα “Μηχανική Μάθηση”

Νικόλαος.Π.Μακρής

Διδάσκων: Γιαννακόπουλος Θεόδωρος, Μεταδιδακτορικός Ερευνητής ΕΚΕΦΕ Δημόκριτος

Αθήνα, 2023

Εισαγωγικά Στοιχεία

- ✓ Πρόβλημα δυαδικής ταξινόμησης (binary classification) (0-μη πόσιμο, 1-πόσιμο)
- ✓ Γίνεται χρήση 9 features:
- ✓ ~3300 δείγματα

#	Feature	Μετάφραση	Περιγραφή
0	pH	pH	Μέτρο οξύτητας H_2O
1	Hardness	Σκληρότητα	Περιεκτικότητα H_2O σε Ca & Mg [mg/L]
2	Solids	Στερεά	Ολικά διαλυμένα στερεά (TDS) [ppm]
3	Chloramines	Χλωραμίνες	Ολικά διαλυμένη NH_2Cl [ppm]
4	Sulfate	Θειικά άλατα	SO_4^{2-} [mg/L]
5	Conductivity	Αγωγμότητα	Ηλεκτρική αγωγμότητα H_2O [$\mu S/cm$]
6	Organic_Carbon	Οργανικός άνθρακας	Συγκέντρωση οργανικού άνθρακα [ppm]
7	Trihalomethanes	Τριαλομεθάνια	Ομάδα χημικών ουσιών [$\mu g/L$]
8	Turbidity	Θολότητα	Μέτρο φωτεινότητας νερού [NTU]

- ✓ Περίπου 61% των δειγμάτων ανήκουν στην κλάση μη-πόσιμων δειγμάτων

2

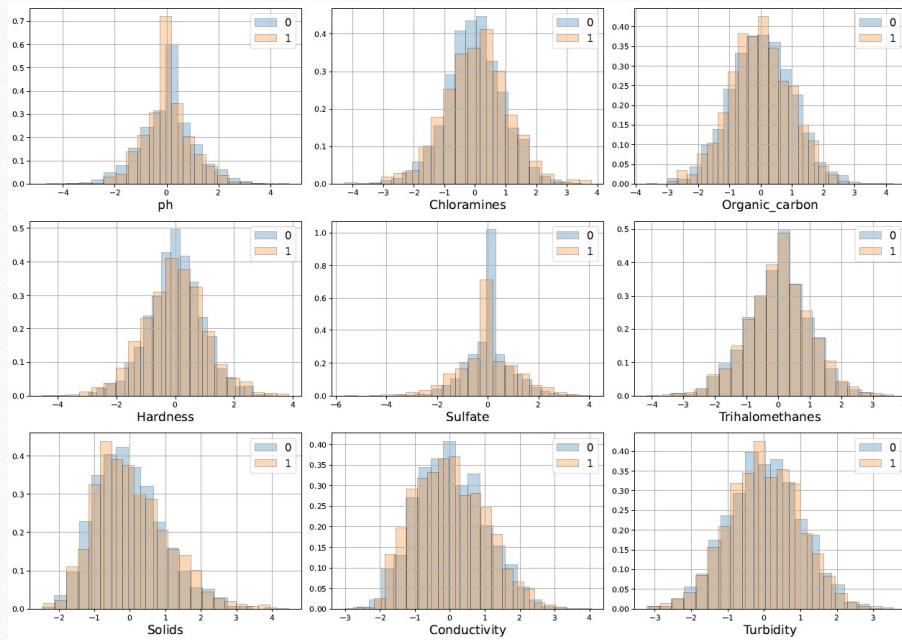
Ταξινόμηση δειγμάτων νερού επί τη βάσει της ποσιμότητας

Προεπεξεργασία Δεδομένων

Συμπέρασμα από έλεγχο Ιστογράμματος Συχνοτήτων:

Οι κατανομές των μετρήσεων κάθε ιδιότητας προσεγγίζουν την κανονική κατανομή

Εμφανής η αλληλοεπικάλυψη των δεδομένων → μικρή συσχέτιση μεταβλητής με κλάση



3

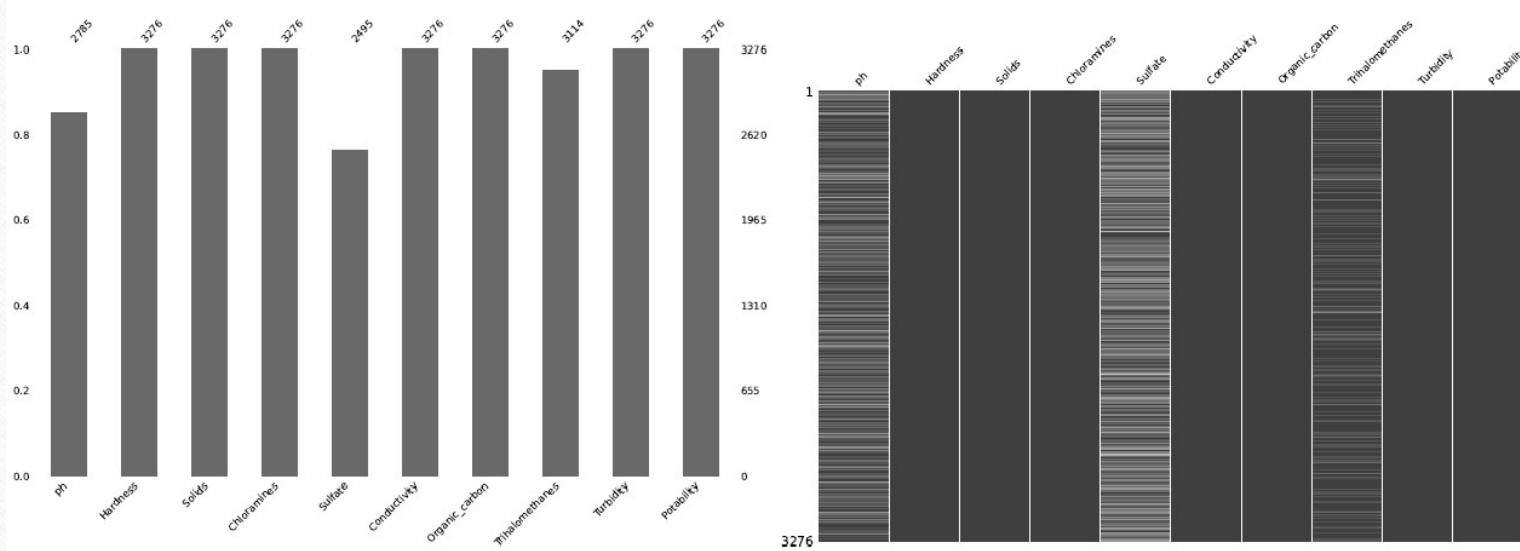
Ταξινόμηση δειγμάτων νερού επί τη βάσει της ποσιμότητας

Προεπεξεργασία Δεδομένων

Διαχείριση μη διαθέσιμων τιμών:

απόδοση τιμής της διαμέσου (median) της στήλης

ή παράλειψη της συγκεκριμένης γραμμής δεδομένων



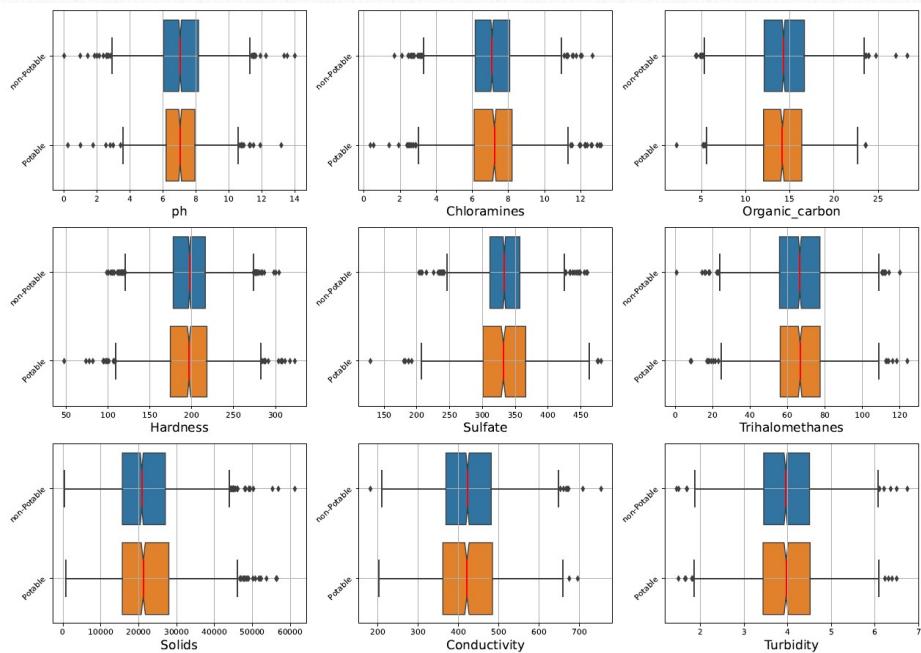
4

Ταξινόμηση δειγμάτων νερού επί τη βάσει της ποσιμότητας

Προεπεξεργασία Δεδομένων

Εντοπισμός ακραίων τιμών: Μέθοδος IQR

Διαχείριση ακραίων τιμών: Παράλειψη τους και αντικατάσταση από τη διάμεση τιμή



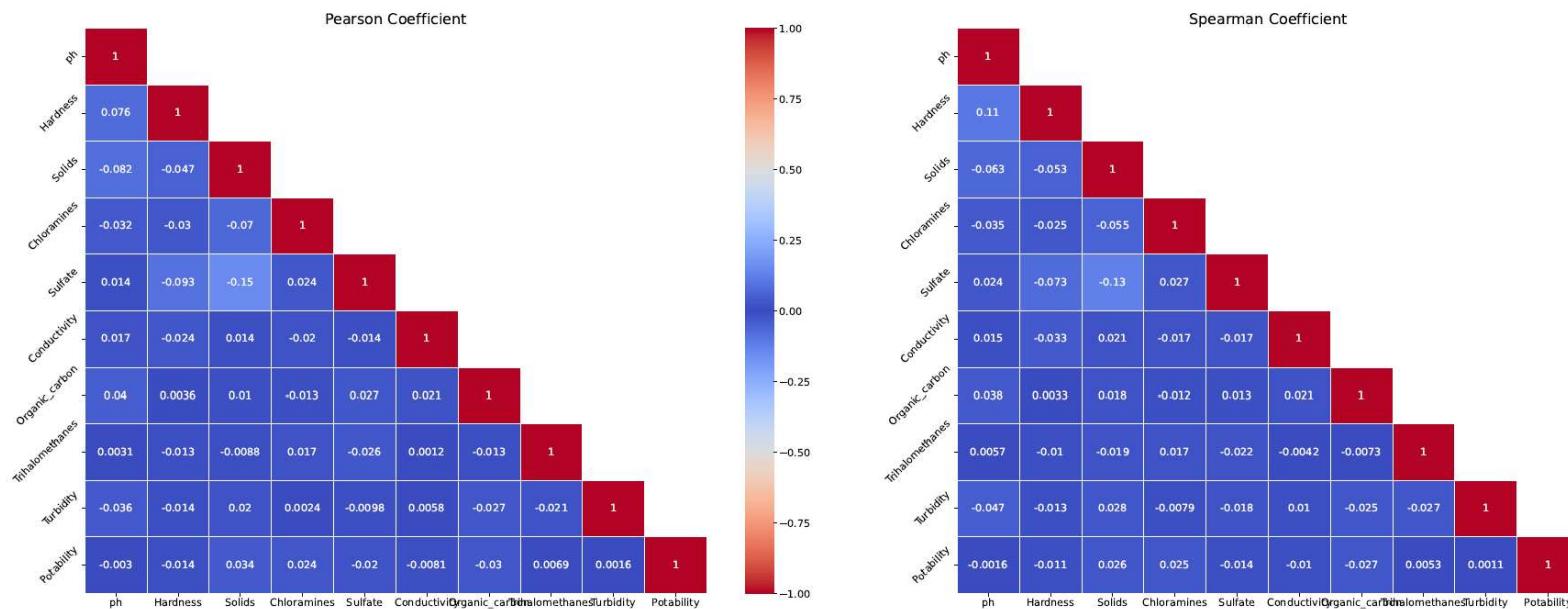
Ταξινόμηση δειγμάτων νερού επί τη βάσει της ποσιμότητας

Προεπεξεργασία Δεδομένων

Συντελεστές συσχέτισης:

Εξαιρετικά μικρή συσχέτιση μεταξύ των διαφόρων μεταβλητών

Κανονικοποίηση (Standardization) των δεδομένων



Ταξινόμηση δειγμάτων νερού επί τη βάσει της ποσιμότητας

Εκπαίδευση Μοντέλων

- Dataset 80% train – 20% test
- Εκπαίδευση πολλών ταξινομητών → k-fold
- Εύρεση των καλύτερων υπερπαραμέτρων
- Κριτήριο η ακρίβεια (accuracy)

5 Σενάρια:

- I. Χρήση όλων των δεδομένων και μη επεξεργασία ακραίων τιμών
- II. Χρήση όλων των δεδομένων εκτός των δεδομένων Conductivity και μη επεξεργασία ακραίων τιμών
- III. Χρήση όλων των δεδομένων εκτός των δεδομένων Solids, Turbidity και μη επεξεργασία ακραίων τιμών
- IV. Χρήση όλων των δεδομένων εκτός των δεδομένων Hardness, Sulfate, Organic_Carbon, Turbidity και μη επεξεργασία ακραίων τιμών
- V. Χρήση όλων των δεδομένων και επεξεργασία ακραίων τιμών

#	Ταξινομητής	case 1	case 2	case 3	case 4	case 5
1	Ridge Regression	0.606	0.606	0.605	0.605	0.606
2	Logistic Regression	0.605	0.606	0.605	0.605	0.605
3	Linear Perceptron	0.505	0.519	0.513	0.518	0.505
4	LDA	0.605	0.606	0.605	0.605	0.605
5	QDA	0.667	0.674	0.664	0.626	0.667
6	Gaussian Process	0.671	0.680	0.662	0.615	0.671
7	Gaussian Naive Bayes	0.621	0.619	0.622	0.602	0.621
8	Linear SVM	0.605	0.605	0.605	0.605	0.605
9	Polynomial SVM	0.605	0.605	0.605	0.605	0.605
10	Gaussian RBF SVM	0.669	0.674	0.666	0.616	0.669
11	k-Neighbors	0.628	0.641	0.623	0.583	0.628
12	Random Forest	0.662	0.661	0.657	0.613	0.661
13	Gradient Boosting	0.662	0.661	0.657	0.613	0.661
14	Ada Boost	0.662	0.661	0.657	0.613	0.661

Δεν υπάρχει εμφανής βελτίωση με κάποια από τις προηγούμενες ενέργειες

Εκπαίδευση Μοντέλων

- Dataset 80% train – 20% test
- Εκπαίδευση πολλών ταξινομητών → k-fold
- Εύρεση των καλύτερων υπερπαραμέτρων
- Κριτήριο η ακρίβεια (accuracy)

5 Σενάρια:

- I. Χρήση όλων των δεδομένων και μη επεξεργασία ακραίων τιμών
- II. Χρήση όλων των δεδομένων εκτός των δεδομένων Conductivity και μη επεξεργασία ακραίων τιμών
- III. Χρήση όλων των δεδομένων εκτός των δεδομένων Solids, Turbidity και μη επεξεργασία ακραίων τιμών
- IV. Χρήση όλων των δεδομένων εκτός των δεδομένων Hardness, Sulfate, Organic_Carbon, Turbidity και μη επεξεργασία ακραίων τιμών
- V. Χρήση όλων των δεδομένων και επεξεργασία ακραίων τιμών

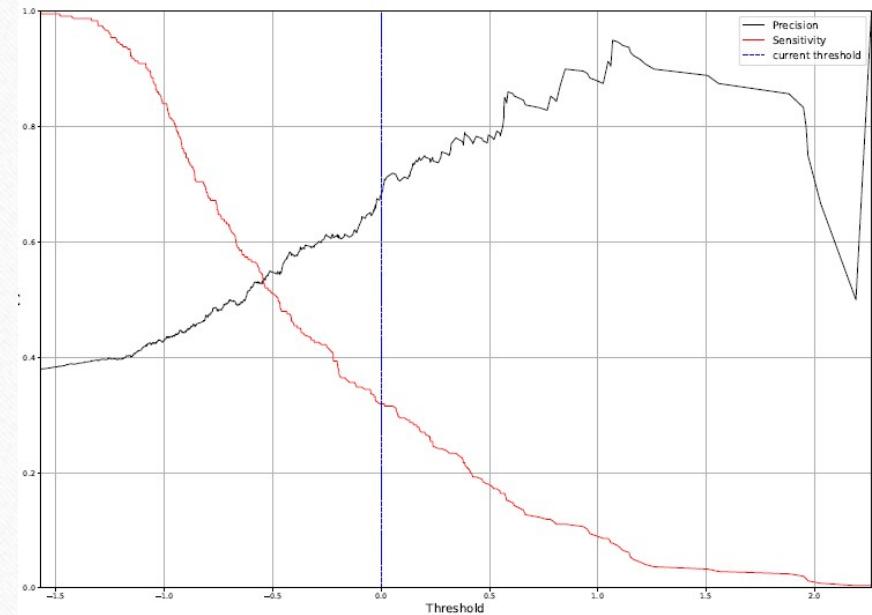
#	Ταξινομητής	case 1	case 2	case 3	case 4	case 5
1	Ridge Regression	0.606	0.606	0.605	0.605	0.606
2	Logistic Regression	0.605	0.606	0.605	0.605	0.605
3	Linear Perceptron	0.505	0.519	0.513	0.518	0.505
4	LDA	0.605	0.606	0.605	0.605	0.605
5	QDA	0.667	0.674	0.664	0.626	0.667
6	Gaussian Process	0.671	0.680	0.662	0.615	0.671
7	Gaussian Naive Bayes	0.621	0.619	0.622	0.602	0.621
8	Linear SVM	0.605	0.605	0.605	0.605	0.605
9	Polynomial SVM	0.605	0.605	0.605	0.605	0.605
10	Gaussian RBF SVM	0.669	0.674	0.666	0.616	0.669
11	k-Neighbors	0.628	0.641	0.623	0.583	0.628
12	Random Forest	0.662	0.661	0.657	0.613	0.661
13	Gradient Boosting	0.662	0.661	0.657	0.613	0.661
14	Ada Boost	0.662	0.661	0.657	0.613	0.661

Δεν υπάρχει εμφανής βελτίωση με κάποια <70% από τις προηγούμενες ενέργειες

Εκπαίδευση Μοντέλων

- I. Gaussian RBF SVM
 - II. Random Forest
 - III. Ada Boost
 - IV. Gradient Boosting
- fine-tuning
- 

Ταξινομητής	Ακρίβεια (Accuracy)	Λεπτομέρεια (Precision)	Ευαισθησία (Recall)	f_1
Gradient Boosting	0.677	0.654	0.279	0.391
Random Forest	0.686	0.698	0.275	0.394
RBF SVM	0.692	0.684	0.320	0.436
Voting Classifier	0.684	0.683	0.283	0.400



9

Ταξινόμηση δειγμάτων νερού επί τη βάσει της ποσιμότητας

Εκπαίδευση Μοντέλων Συμπεράσματα

- Χαμηλή ακρίβεια γραμμικών μοντέλων
- ~8min εκπαίδευσης για ~3000 δείγματα
- Καλύτεροι ταξινομητές: 1.Gaussian RBF SVM, 2.Random Forest, 3.Gradient Boosting
- Voting Classifier → ακρίβεια 68.4% (~baseline 61%)
- CV accuracy σχεδόν ταυτίζεται με test accuracy → ένδειξη καλής γενίκευσης μοντέλου