

Άσκηση 4 (Ταξινομητής Naïve Bayes)

Η απόφαση ενός δρομέα αποστάσεων να προπονηθεί ή όχι εξαρτάται από τις επικρατούσες καιρικές συνθήκες. Για τις τελευταίες 14 μέρες, έχουμε καταγράψει τα καιρικά δεδομένα καθώς και την απόφασή του, όπως φαίνεται στον παρακάτω πίνακα

Ημέρα	Καιρός	Θερμοκρασία	Υγρασία	Αέρας	Προπόνηση
1	Βροχερός	Υψηλή	Υψηλή	Χαμηλός	Όχι
2	Βροχερός	Υψηλή	Υψηλή	Υψηλός	Όχι
3	Συννεφιασμένος	Υψηλή	Υψηλή	Χαμηλός	Ναι
4	Ηλιοφάνεια	Μέτρια	Υψηλή	Χαμηλός	Ναι
5	Ηλιοφάνεια	Χαμηλή	Κανονική	Χαμηλός	Ναι
6	Ηλιοφάνεια	Χαμηλή	Κανονική	Υψηλός	Όχι
7	Συννεφιασμένος	Χαμηλή	Κανονική	Υψηλός	Ναι
8	Βροχερός	Μέτρια	Υψηλή	Χαμηλός	Όχι
9	Βροχερός	Χαμηλή	Κανονική	Χαμηλός	Ναι
10	Ηλιοφάνεια	Μέτρια	Κανονική	Χαμηλός	Ναι
11	Βροχερός	Μέτρια	Κανονική	Υψηλός	Ναι
12	Συννεφιασμένος	Μέτρια	Υψηλή	Υψηλός	Ναι
13	Συννεφιασμένος	Υψηλή	Κανονική	Χαμηλός	Ναι
14	Ηλιοφάνεια	Μέτρια	Υψηλή	Υψηλός	Όχι

Εκπαιδευόντας με τα παραπάνω δεδομένα έναν ταξινομητή Naïve Bayes και αν υποθέσουμε ότι σήμερα ο καιρός έχει *Ηλιοφάνεια*, η θερμοκρασία είναι *Υψηλή*, η υγρασία *Κανονική* και ο αέρας *Χαμηλός*, τι είναι πιθανότερο να κάνει ο αθλητής; Να αθληθεί ή να μην αθληθεί;

Ενδεικτική Απάντηση

Παρατηρούμε ότι το συγκεκριμένο πρόβλημα είναι ένα πρόβλημα δυαδικής ταξινόμησης (Προπονείται/Δεν προπονείται ο δρομέας), ενώ τα δεδομένα μας περιγράφονται από 4 χαρακτηριστικά (Καιρός, Θερμοκρασία, Υγρασία και Αέρας).

Στο πρώτο βήμα υπολογίζουμε τη σχετική συχνότητα της κάθε κλάσης (πόσες μέρες προπονήθηκε/δεν προπονήθηκε ο δρομέας δηλαδή).

$$P(\text{Προπονήθηκε}) = \frac{9}{14} \text{ και } P(\text{Δεν Προπονήθηκε}) = \frac{5}{14}$$

Στη συνέχεια υπολογίζουμε τη σχετική συχνότητα $p(x^{(k)}|i)$ του χαρακτηριστικού $x^{(k)}$ στην κλάση i .

Καιρός

	Προπονήθηκε	Δεν προπονήθηκε	$P(\text{Προπονήθηκε})$	$P(\text{Δεν Προπονήθηκε})$
Ηλιοφάνεια	3	2	$\frac{3}{9}$	$\frac{2}{5}$
Συννεφιασμένος	4	0	$\frac{4}{9}$	0
Βροχερός	2	3	$\frac{2}{9}$	$\frac{3}{5}$
Σύνολο	9	5	1	1

Θερμοκρασία

	Προπονήθηκε	Δεν προπονήθηκε	$P(\text{Προπονήθηκε})$	$P(\text{Δεν Προπονήθηκε})$
Υψηλή	2	2	$\frac{2}{9}$	$\frac{2}{5}$
Μέτρια	4	2	$\frac{4}{9}$	$\frac{2}{5}$
Χαμηλή	3	1	$\frac{3}{9}$	$\frac{1}{5}$
Σύνολο	9	5	1	1

Υγρασία

	Προπονήθηκε	Δεν προπονήθηκε	$P(\text{Προπονήθηκε})$	$P(\text{Δεν Προπονήθηκε})$
Υψηλή	3	4	$\frac{3}{9}$	$\frac{4}{5}$
Κανονική	6	1	$\frac{6}{9}$	$\frac{1}{5}$
Σύνολο	9	5	1	1

Αέρας

	Προπονήθηκε	Δεν προπονήθηκε	$P(\text{Προπονήθηκε})$	$P(\text{Δεν Προπονήθηκε})$
Υψηλός	3	3	$\frac{3}{9}$	$\frac{3}{5}$
Χαμηλός	6	2	$\frac{6}{9}$	$\frac{2}{5}$
Σύνολο	9	5	1	1

Αν για σήμερα ισχύει $x = (\text{Ηλιοφάνεια}, \text{Υψηλή}, \text{Κανονική}, \text{Χαμηλός})$ τότε η πιθανότητα να προπονηθεί ο δρομέας είναι ίση με

$$P(\text{Προπονήθηκε}|x)$$

$$= \frac{P(\text{Προπονήθηκε}) \prod_{k=1}^4 p(x^{(k)}|\text{Προπονήθηκε})}{P(\text{Προπονήθηκε}) \prod_{k=1}^4 p(x^{(k)}|\text{Προπονήθηκε}) + P(\text{Δεν Προπονήθηκε}) \prod_{k=1}^4 p(x^{(k)}|\text{Δεν Προπονήθηκε})}$$

$$P(\text{Προπονήθηκε}|x) = \frac{\frac{9}{14} \cdot \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9}}{\frac{9}{14} \cdot \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} + \frac{5}{14} \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5}} \approx 0,82$$

Αντίστοιχα, η πιθανότητα να μην προπονηθεί ο δρομέας είναι ίση με

$$P(\text{Δεν Προπονήθηκε}|x)$$

$$= \frac{P(\text{Δεν Προπονήθηκε}) \prod_{k=1}^4 p(x^{(k)}|\text{Δεν Προπονήθηκε})}{P(\text{Προπονήθηκε}) \prod_{k=1}^4 p(x^{(k)}|\text{Προπονήθηκε}) + P(\text{Δεν Προπονήθηκε}) \prod_{k=1}^4 p(x^{(k)}|\text{Δεν Προπονήθηκε})}$$

$$P(\text{Δεν Προπονήθηκε}|x) = \frac{\frac{5}{14} \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5}}{\frac{9}{14} \cdot \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} + \frac{5}{14} \cdot \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5}} \approx 0,18$$

Συνεπώς, επειδή $P(\text{Προπονήθηκε}|x) > P(\text{Δεν Προπονήθηκε}|x)$, ο ταξινομητής Naive Bayes προβλέπει ότι με βάση τις σημερινές καιρικές συνθήκες, ο δρομέας θα προπονηθεί.