Random mushroom or Random forest ?
Σύγκριση αλγορίθμων μηχανικής μάθησης για πρόβλεψη εδωδιμότητας άγριων μανιταριών

Εργασία εξαμήνου για το μάθημα : Μέθοδοι ανάλυσης στην Βιοπληροφορική

12/11/2025
Θοδωρής Γιωννάς-Μασσέλος
ΠΜΣ “Βιοπληροφορική”
Ιατρική Σχολή
Πανεπιστήμιο Κρήτης

Abstract

Within mushroom-foraging communities and citizen scientists recording species occurrences correct identification of mushroom species remains a challenge. Species identification and therefore mushroom edibility are still crucial for high quality citizen science data acquisition and foraging safety respectively. Applications of ML algorithms for identification have suggested the use of pictures for the correct identification of species, but identification is not always possible with high certainty. Here we'll use a public dataset made of 22 qualitative characteristics from 23 gilled mushroom species (*fam. Agaricus* and *Lepiota)* to build a classification scheme for mushroom profiles in regard to their edibility. We used kNN and Random Forest algorithms for classification and compared their efficacy. Our goal is to create a scheme that successfully identifies mushroom species of our dataset and explore the capabilities of our pipeline for cases where data might be missing due to lack of user input or knowledge in a real-filed-use scenario .

Επιλογή Δεδομένων και Μεθοδολογία

Όλα τα δεδομένα προήλθαν από το αποθετήριο μηχανικής μάθησης του Πανεπιστημίου της Καλιφόρνια στο Ιρβάιν. Το σετ δεδομένων “mushrooms” είχε 22 κατηγορικές μεταβλητές και 8124 δείγματα στο σύνολο. Τα δείγματα αποτελούν προσομοιώσεις βάσει των περιγραφών για 23 είδη μανιταριών των οικογενειών *Lepiota* και *Agaricus.* Η μεταβλητή-στόχος της ταξινόμησης, που είναι η εδωδιμότητα χωρίζεται σε δύο μεταβλητές, φαγώσιμο (=edible) και δηλητηριώδες (=poisonous) καθιστώντας το πρόβλημα ταξινόμησης δυαδικό. Από τα είδη που χρησιμοποιούνται όσα είναι άγνωστης εδωδιμότητας και δεν προτείνονται συμπεριλήφθηκαν στα δηλητηριώδη. Το σύνολο δεδομένων περιέχει απούσες τιμές και οι μεταβλητές είναι είτε δυαδικές είτε με πολλές κλάσεις αλλά μη διαβαθμισμένες. Οι μεταβλητές και οι τιμές τους φαίνονται στον *Πίνακα 1*.

Οι αναλύσεις πραγματοποιήθηκαν με χρήση λειτουργιών της βιβλιοθήκης Scikit-learn της Python. Δημιουργήθηκαν δύο αντίγραφα των δεδομένων, ένα για κάθε ανάλυση : KNN και Random Forest. Έγινε κωδικοποίηση των κατηγοριών των μεταβλητών και αφαιρέθηκαν οι απούσες τιμές από το αντίγραφο που προοριζόταν για την εφαρμογή KNN. Τα δεδομένα χωρίστηκαν στα δύο για την εκπαίδευση και την αξιολόγηση των αλγορίθμων αντίστοιχα.

Πίνακας 2

name	type	description
poisonous	Categorical	null
cap-shape	Categorical	bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
cap-surface	Categorical	fibrous=f,grooves=g,scaly=y,smooth=s
cap-color	Binary	brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
bruises	Categorical	bruises=t,no=f
odor	Categorical	almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
gill-attachment	Categorical	attached=a,descending=d,free=f,notched=n
gill-spacing	Categorical	close=c,crowded=w,distant=d
gill-size	Categorical	broad=b,narrow=n
gill-color	Categorical	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
stalk-shape	Categorical	enlarging=e,tapering=t
stalk-root	Categorical	bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
stalk-surface-above-ring	Categorical	fibrous=f,scaly=y,silky=k,smooth=s
stalk-surface-below-ring	Categorical	fibrous=f,scaly=y,silky=k,smooth=s
stalk-color-above-ring	Categorical	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
stalk-color-below-ring	Categorical	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
veil-type	Binary	partial=p,universal=u
veil-color	Categorical	brown=n,orange=o,white=w,yellow=y
ring-number	Categorical	none=n,one=o,two=t
ring-type	Categorical	cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
spore-print-color	Categorical	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
population	Categorical	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
habitat	Categorical	grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Random Forest

Για την επιλογή των βέλτιστων υπερπαραμέτρων για το μοντέλο έγινε δοκιμή όλων των συνδυασμών παραμέτρων από ένα σύνολο παραμέτρων εξάγωντας τον μέσο όρο επίδοσης από πέντε διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης (Grid Search Cross Validation). Τέλος υπολογίστηκαν ο δείκτες : Accuracy, Precision, Recall για το επιλεγμένο μοντέλο.

Στόχος της ανάλυσης μας είναι και η εύρεση των πιο σημαντικών χαρακτήρων των μανιταριών που φαίνεται να επηρεάζουν την ικανότητα του αλγορίθμου για πρόβλεψη. Για να εστιάσουμε στις πιο επιδραστικούς χαρακτήρες υπολογίσαμε την σημαντικότητα τους με δύο μεθόδους: Mean Decrease in Impurity (MDI) και Permutation Importance. Χρησιμοποιήθηκαν δύο μέθοδοι, καθώς αναφέρεται ότι η πρώτη μπορεί να αυξάνει αδικαιολόγητα την σημαντικότητα κάποιων μεταβλητών (1). Τελικά, οι πέντε πιο σημαντικές μεταβλητές, όπως προέκυψαν και από την κάθε μέθοδο χρησιμοποιήθηκαν για εκ νέου εκπαίδευση των μοντέλων και υπολογίστηκαν οι ίδιοι δείκτες αποτελεσματικότητας. Οι μεταβλητές που επιλέχθηκαν για την εκπαίδευση των μοντέλων μετά την εκτίμηση σημαντικότητας είναι οι ['cap-shape', 'cap-color', 'cap-surface', 'odor', 'bruises'] βάσει της μεθόδου MDI και οι ['odor', 'spore-print-color', 'gill-size', 'habitat', 'bruises'] βάσει της μεθόδου Permutation.

K-Nearest Neighbors

Χρησιμοποιήθηκε η λειτουργία KNeighborsClassifier() για την δημιουργία του μοντέλου. Εξετάστηκε η απόδοση του μοντέλου για διαφορετικά K, από K=2 έως K=10 και υπολογίστηκαν και πάλι οι ίδιοι δείκτες: Accuracy, Precision, Recall. Ο δείκτης Accuracy υπολογίστηκε εξάγωντας τον μέσο όρο από πέντε διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης (Cross Validation).

Πέραν την κλασσικής λειτουργία KNN, δοκιμάστηκε και ο συνδυασμός KNN με την μέθοδο Neighborhood Component Analysis. Αυτή η μέθοδος μπορεί να προσφέρει ένα πιο φιλοσοφημένο μέτρο καθορισμού των γειτόνων στην μέθοδο KNN σε σχέση με την “Ευκλείδεια απόσταση” που χρησιμοποιείται συνήθως (2). Στην περίπτωση μας η εφαρμογή της μεθόδου έγινε εξερευνητικά.

Αποτελέσματα

Οι δύο αλγόριθμοι πέτυχαν πολύ υψηλές τιμές στην ακρίβεια πρόβλεψης της εδωδιμότητας, αποδεικνύοντας ότι για το συγκεκριμένο σύνολο δεδομένων η κατηγοριοποίηση σε εδώδιμα ή μη μπορεί να γίνει με αξιοπιστία. Στον *Πίνακα 2* φαίνονται οι δείκτες αποτελεσματικότητας για το Random Forest ενώ στους *Πίνακες 3* & *4* για το kNN και το kNN με Neighbor Component Analysis αντίστοιχα. Στα *Γραφήματα 1* και *2* φαίνονται οι εκτιμήσεις για την σημαντικότητα των μεταβλητών.

Πίνακας 2

	Random Forest	Random Forest with selected features (MDI based)	Random Forest with selected features (Permutation based)
Accuracy	100%	99.38%	99.92%
Precision	100%	100%	100%
Recall	100%	98.73%	99.83%

Πίνακας 3

kNN model with Cross-Validation

K	training	test	C-V_mean_accuracy	Precision	Recall
K = 2 :	100.00%	100.00%	100.00%	100.00%t	100.00%
K = 3 :	100.00%	100.00%	100.00%	100.00%t	100.00%
K = 4 :	100.00%	100.00%	100.00%	100.00%t	100.00%
K = 5 :	100.00%	100.00%	100.00%	100.00%t	100.00%
K = 6 :	100.00%	100.00%	100.00%	100.00%t	100.00%
K = 7 :	100.00%	100.00%	100.00%	100.00%t	100.00%
K = 8 :	99.98%	100.00%	100.00%	100.00%t	100.00%
K = 9 :	99.98%	100.00%	100.00%	100.00%t	100.00%
K = 10 :	99.98%	100.00%	99.99%	100.00%t	100.00%

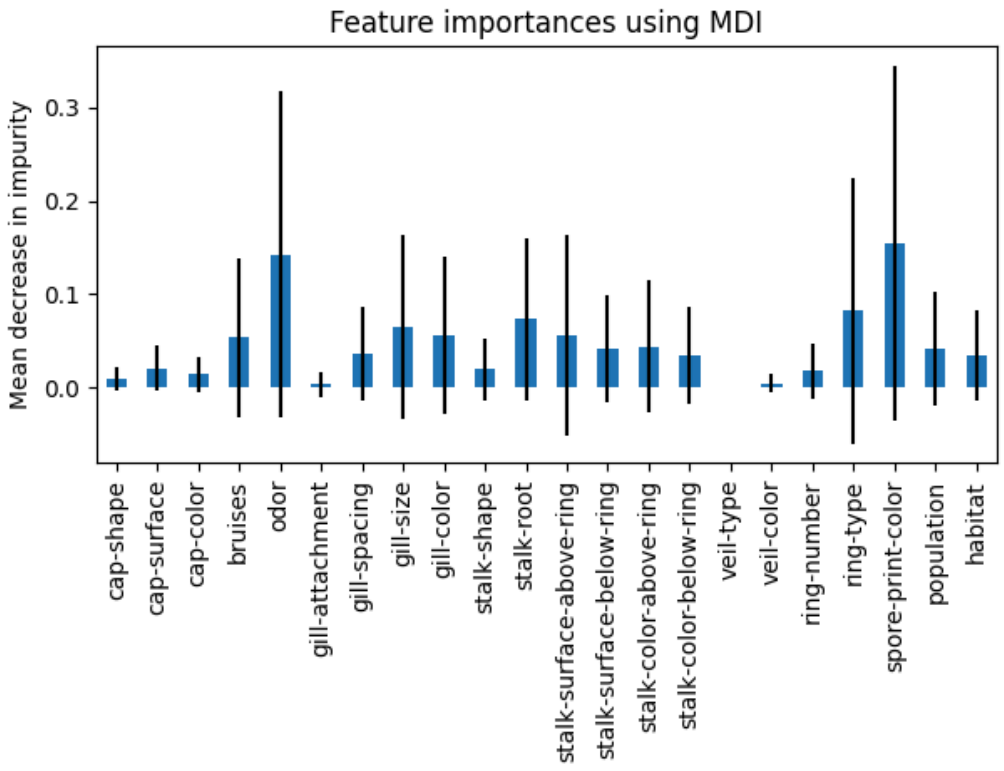
Πίνακας 4

---

kNN model NCA with Cross-Validation

	Mean CV accuracy (train)	Accuracy (test)	Precision	Recall
K = 2 :	100.000%	100.000%	100.000%	100.000%
K = 3 :	100.000%	100.000%	100.000%	100.000%
K = 4 :	99.965%	100.000%	100.000%	100.000%
K = 5 :	99.965%	100.000%	100.000%	100.000%
K = 6 :	99.930%	100.000%	100.000%	100.000%
K = 7 :	99.965%	100.000%	100.000%	100.000%
K = 8 :	99.930%	99.877%	100.000%	99.747%
K = 9 :	99.965%	99.877%	100.000%	99.747%
K = 10 :	99.965%	99.877%	100.000%	99.747%

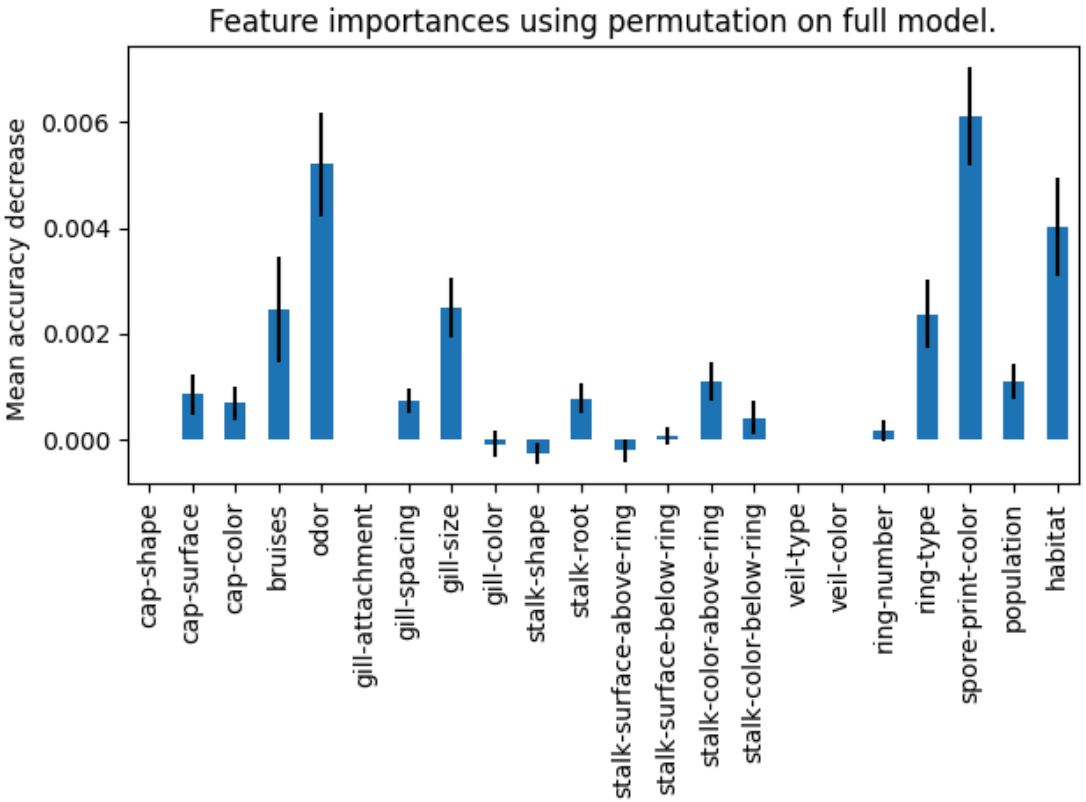
Γράφημα 1



Συζήτηση:

Τα δύο μοντέλα είχαν πολύ υψηλή ακρίβεια, κάτι που δείχνει ότι η εφαρμογή των συγκεκριμένων μεθόδων μάλλον επαρκεί για την ταξινόμηση ως προς την εδωδιμότητα, που ήταν και ο αρχικός στόχος αυτής της ανάλυσης. Οι δύο αλγόριθμοι χρησιμοποιήθηκαν χωρίς πρότερη γνώση του πεδίου εφαρμογής τους και χάρη σε αυτό βγαίνουν σημαντικά συμπεράσματα για το πρότζεκτ και τις δυνατότητες εφαρμογής των αλγορίθμων “σε πραγματικές συνθήκες”.

Γράφημα 2



Το γεγονός ότι έχουμε πολύ υψηλή απόδοση στην ταξινόμηση δείχνει ότι το δείγμα μας ήταν αρκετά προβλέψιμο και είχε συγκεκριμένα μοτίβα. Δημιουργείται λοιπόν ένα ερώτημα αν με ένα πιο ευρύ δείγμα που να αφορά σε περισσότερα είδη μανιταριών θα είχαμε εξίσου καλά αποτελέσματα. Αντίστοιχα και με ένα σετ δεδομένων με περισσότερες απούσες τιμές που θα μοιάζει περισσότερες με πραγματικές καταγραφές από ανθρώπους που δεν μπορούν πάντα να δώσουν τιμές για όλους τους χαρακτήρες.

Αξίζει να σημειωθεί ότι οι δύο μέθοδοι ανάλυσης της σημαντικότητας των χαρακτήρων (Feature Importance) έδειξαν διαφορετικούς χαρακτήρες ως πιο σημαντικούς κάτι που επιβεβαιώνει την ανάγκη για επίβλεψη του μοντέλου από ειδικούς που να μπορούν να επιβεβαιώσουν ότι οι μεταβλητές που επιλέγονται έχουν βιολογική, οικολογική και

γενικότερα εμπειρική βάση. Αυτό ακολουθεί την τάση σε πολλές εφαρμογές της μηχανικές μάθησης να υπάρχει “λογοδοσία” για το πως ένα μοντέλο βγάζει τα αποτελέσματά του. Να μην είναι δηλαδή το μοντέλο ένα “black box” που δεν καταλαβαίνουμε τι προκαλεί τα διαφορετικά αποτελέσματα. Σε αυτό το πλαίσιο ο αλγόριθμος KNN φαίνεται να είναι ακατάλληλος αφού δεν μπορεί να υποδείξει ποιές μεταβλητές είναι πιο σημαντικές για το διαχωρισμό των δεδομένων. Επίσης, ο KNN δεν επιτρέπει την εισαγωγή δεδομένων με απούσες τιμές, κάτι που θα εμπόδιζε την εφαρμογή του σε πραγματικές συνθήκες όπου ένας χρήστης ενδέχεται να μην μπορεί να καταχωρήσει κάθε φορά τιμές για όλους τους χαρακτήρες.

Η χρήση σύνθετων μεθόδων μηχανικής μάθησης όπως το Random Forest είναι υποσχόμενη για την ταξινόμηση ως προς την εδωδιμότητα. Γενικά η προσπάθεια για ταξινόμηση των μανιταριών μπορεί να προχωρήσει πολύ περισσότερο από την εξέταση της εδωδιμότητας και ιδανικά θα ήταν να μπορούσε να προβλεφθεί το είδος του μανιταριού. Ένα τέτοιο πόνημα περιλαμβάνει πολύ περισσότερες κλάσεις για την μεταβλητή στόχο αλλά αποτελεί πιο ολοκληρωμένη προσέγγιση αφού στην μυκητολογία η εδωδιμότητα μελετάται σε επίπεδο είδους. Επίσης μια τέτοια προσέγγιση μπορεί να συμβάλλει σημαντικά στην εξαγωγή μεγάλου όγκου δεδομένων από βάσεις δεδομένων της επιστήμης των πολιτών. Στο ίδιο πλαίσιο πρέπει να γίνεται αξιοποίηση φωτογραφιών σε συνδυασμό με τις περιγραφικές μεθόδους για τις ταυτοποιήσεις καθώς οι φωτογραφίες αποτελούν πιο συχνό μέσο καταγραφής δειγμάτων.

Βιβλιογραφία και πηγές:

Δεδομένα : https://archive.ics.uci.edu/dataset/73/mushroom

1) https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html
2) https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NeighborhoodComponentsAnalysis.html

Κώδικας:
https://github.com/bozydar-masselos/mushroom_edibility_classifier