ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Στατιστική Μηχανική Μάθηση

Τμήμα Στατιστικής

2η Εργασία

Ημερομηνία Παράδοσης: 9 Ιανουαρίου 2026

Περιγραφή Εργασίας

Σκοπός της παρούσας εργασίας είναι η δοκιμή και ανάλυση της συμπεριφοράς διαφορετικών μοντέλων μηχανικής μάθησης καθώς και η τελική επιλογή ενός μοντέλου για την πρόβλεψη καρδιακής νόσου. Συνεπώς, εστιάζουμε στην μάθηση με επίβλεψη (supervised learning) και καλείστε να συγκρίνετε διάφορους αλγορίθμους ταξινόμησης, τόσο σε επίπεδο απόδοσης ταξινόμησης όσο και ερμηνείας της συμπεριφοράς τους.
Το σύνολο δεδομένων αποτελείται από 918 εγγραφές και περιέχει συνολικά 11 μεταβλητές εισόδου και 1 μεταβλητή στόχου για πρόβλεψη. Συγκεκριμένα, οι μεταβλητές παρουσιάζονται στον Πίνακα 1:

Feature	Description	Values / Units
Age	Age of the patient	Years
Sex	Sex of the patient	M: Male, F: Female
ChestPainType	Chest pain type	TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic
RestingBP	Resting blood pressure	mm Hg
Cholesterol	Serum cholesterol	mg/dl
FastingBS	Fasting blood sugar	1: FastingBS > 120 mg/dl, 0: otherwise
RestingECG	Resting electrocardiogram results	Normal: Normal, ST: ST-T wave abnormality (T wave inversions and/or ST elevation/depression > 0.05 mV), LVH: Left ventricular hypertrophy by Estes' criteria
MaxHR	Maximum heart rate achieved	Numeric (60–202)
ExerciseAngina	Exercise-induced angina	Y: Yes, N: No
Oldpeak	ST depression induced by exercise	Numeric value (mm)
ST_Slope	Slope of the peak exercise ST segment	Up: upsloping, Flat: flat, Down: downsloping
HeartDisease	Target variable	1: Heart disease, 0: Normal

Πίνακας 1: Περιγραφή χαρακτηριστικών του συνόλου δεδομένων Καρδιακής Νόσου με μονάδες και τιμές

1 Διερευνητική Ανάλυση Δεδομένων (EDA) και Feature Engineering

Στο πρώτο στάδιο της εργασίας σας θα πραγματοποιήσετε διερευνητική ανάλυση δεδομένων (Exploratory Data Analysis – EDA) και βασικό feature engineering στο παρεχόμενο σύνολο δεδομένων, το οποίο αφορά δημογραφικά και κλινικά χαρακτηριστικά ασθενών σε σχέση με την εμφάνιση καρδιακής νόσου. Αρχικά, μελετήστε τη δομή του συνόλου δεδομένων και ελέγξτε για πιθανές ελλιπές τιμές και ανωμαλίες. Ιδιαίτερη προσοχή θα πρέπει να δοθεί και στη μεταβλητή-στόχο HeartDisease, εξετάζοντας την κατανομή των τιμών της και σχολιάζοντας αν υπάρχει ανισορροπία μεταξύ των δύο κατηγοριών.

Παράλληλα, διερευνήστε την παρουσία ακραίων τιμών, ιδιαίτερα σε συνεχείς μεταβλητές και συζητήστε αν αυτές οι τιμές πρέπει να διατηρηθούν, να μετασχηματιστούν ή να αποκλειστούν από την ανάλυση. Υπολογίστε βασικά περιγραφικά στατιστικά μέτρα για τις ποσοτικές μεταβλητές, όπως ο μέσος όρος, η διάμεσος, η τυπική απόκλιση και τα ακραία σημεία, τόσο συνολικά όσο και ξεχωριστά για ασθενείς με και χωρίς καρδιακή νόσο και οπτικοποιείστε τα αποτελέσματα (ιστογράμματα, boxplots ή ραβδογράμματα), με στόχο την ανάδειξη μοτίβων, διαφορών μεταξύ ομάδων και πιθανών ενδείξεων συσχέτισης με τη μεταβλητή-στόχο.

Στο πλαίσιο της διερεύνησης των σχέσεων μεταξύ των μεταβλητών, υπολογίστε μέτρα συσχέτισης για τις ποσοτικές μεταβλητές και παρουσιάστε τα αποτελέσματα με κατάλληλα γραφήματα, όπως πίνακες ή heatmaps συσχετίσεων. Εξετάστε ποιες μεταβλητές εμφανίζουν ισχυρή συσχέτιση μεταξύ τους ή με την ύπαρξη καρδιακής νόσου και σχολιάστε τυχόν ενδείξεις πολυσυγγραμμικότητας. Μπορείτε επίσης να εξετάσετε τη δημιουργία νέων μεταβλητών, αναφέροντας πάντα τους λόγους για τους οποίους θεωρείτε ότι θα βοηθήσουν την ανάλυση.

Στο τέλος αναφέρετε τα βασικά ευρήματα και ποιες μεταβλητές θα χρησιμοποιήσετε για την μοντελοποίηση.

Μέθοδοι Μηχανικής Μάθησης

Αρχικά, εφαρμόστε έναν ταξινομητή Naive Bayes σχολιάζοντας τις βασικές υποθέσεις, ιδιαίτερα την υπόθεση ανεξαρτησίας των χαρακτηριστικών, και συζητήστε σε ποιο βαθμό αυτές είναι ρεαλιστικές για το συγκεκριμένο σύνολο δεδομένων. Παρουσιάστε τα αποτελέσματα της ταξινόμησης και αξιολογήστε την απόδοση του μοντέλου χρησιμοποιώντας κατάλληλα μέτρα, όπως accuracy, precision, recall, F1-score, confusion matrices και ROC καμπύλες.

Στη συνέχεια, εφαρμόστε μεθόδους ensemble βασισμένες στο bagging. Εκπαιδεύστε ένα μοντέλο bagging (π.χ. Bagging Classifier με δέντρα αποφάσεων ως βασικούς ταξινομητές ή random forests), σχολιάστε την επιλογή σας και εξετάστε την επίδραση παραμέτρων όπως ο αριθμός των βασικών μοντέλων και το βάθος των δέντρων στην απόδοση του ταξινομητή και συγκρίνετε τα αποτελέσματα με αυτά του Naive Bayes.

Τέλος, εφαρμόστε μία μέθοδο boosting, όπως AdaBoost, Gradient Boosting, XGBoost, κλπ. Σχολιάστε την επιλογή σας και μελετήστε την επίδραση βασικών υπερπαραμέτρων (π.χ. learning rate, αριθμός επαναλήψεων) και αξιολογήστε την απόδοση του τελικού μοντέλου τόσο αυτοτελώς όσο και σε σχέση με τις προηγούμενες μεθόδους.

Για όλες τις μεθόδους, χωρίστε τα δεδομένα σε σύνολα εκπαίδευσης και ελέγχου ή χρησιμοποιήστε cross-validation, εξηγώντας τη διαδικασία και τις επιλογές που ακολουθήσατε. Συγκρίνετε συστηματικά τις μεθόδους ως προς την προγνωστική τους ικανότητα, τη σταθερότητα των αποτελεσμάτων και την ερμηνευσιμότητα των μοντέλων. Ολοκληρώστε την ανάλυση σχολιάζοντας ποια μέθοδος αποδίδει καλύτερα στο συγκεκριμένο πρόβλημα και γιατί, λαμβάνοντας υπόψη τόσο τα στατιστικά αποτελέσματα όσο και τις υποκείμενες υποθέσεις κάθε μεθόδου.

Παραδοτέα

Ως παραδοτέο της εργασίας θα πρέπει να υποβληθεί 1)ο πλήρης κώδικας που χρησιμοποιήθηκε για την ανάλυση, γραμμένος στη γλώσσα R, καθώς και 2) μία αναφορά σε μορφή αρχείου PDF.

Ο κώδικας θα πρέπει να είναι πλήρως λειτουργικός, κατάλληλα σχολιασμένος και οργανωμένος, έτσι ώστε να είναι σαφές ποια βήματα αντιστοιχούν στη διερευνητική ανάλυση δεδομένων, στο feature engineering και στην εφαρμογή των μεθόδων μηχανικής μάθησης. Θα πρέπει να αποφεύγεται η απλή παράθεση εντολών χωρίς τεκμηρίωση και να διασφαλίζεται ότι όλα τα αποτελέσματα της αναφοράς μπορούν να αναπαραχθούν από τον παρεχόμενο κώδικα.

Η αναφορά θα πρέπει να είναι δομημένη και να καλύπτει όλα τα στάδια της ανάλυσης που ζητήθηκαν. Θα πρέπει να παρουσιάζονται και να σχολιάζονται τα αποτελέσματα της διερευνητικής ανάλυσης, όπως περιγραφικά στατιστικά, γραφήματα και συσχετίσεις, καθώς και να τεκμηριώνονται οι επιλογές που αφορούν τη διαχείριση ελλιπών τιμών, την αντιμετώπιση ανωμαλιών, τους μετασχηματισμούς των μεταβλητών και τη δημιουργία νέων χαρακτηριστικών. Στο μέρος της μηχανικής μάθησης, η αναφορά θα πρέπει να περιλαμβάνει περιγραφή των μεθόδων που εφαρμόστηκαν (Naive Bayes, bagging και boosting), της διαδικασίας εκπαίδευσης και αξιολόγησης των μοντέλων, καθώς και συγκριτική παρουσίαση των αποτελεσμάτων με κατάλληλες μετρικές και γραφήματα απόδοσης. Η αναφορά θα πρέπει να συνδυάζει κείμενο, πίνακες και γραφήματα με σαφή και συνεκτικό τρόπο, να περιλαμβάνει ερμηνεία των αποτελεσμάτων και να καταλήγει σε συμπεράσματα που συνοψίζουν τα βασικά ευρήματα και αιτιολογούν την επιλογή της καλύτερης μεθόδου.

Καλή επιτυχία!