Ανάκτηση Πληροφοριών και Εξόρυξη Δεδομένων 2020-2021

2η Εργασία

(Ατομική εργασία)

Data Analysis and Machine Learning

Έχετε στη διάθεσή σας ένα σύνολο δεδομένων (dataset) με 437 παρατηρήσεις (rows) και 6831 μεταβλητές (features or columns). Το dataset αφορά το είδος Engraulis encrasicolus (το ψάρι γνωστό στα ελληνικά ως γαύρος) και οι παρατηρήσεις είναι οι εμφανίσεις του στη Μεσόγειο θάλασσα. Στόχος είναι η εκπαίδευση ενός regression μοντέλου σε ένα (train set) για την πρόβλεψη της πιθανότητας εμφάνισής του (τιμές από 0 εως 1) σε άγνωστα σημεία της θάλασσας (test set).

Η εργασία έχει και τη μορφή διαγωνισμού όπου θα αναμετρηθείτε μεταξύ σας. Η σελίδα του διαγωνισμού είναι: https://www.kaggle.com/c/engraulis-encrasicolus-prediction/. Οδηγίες για τη χρήση της ιστοσελίδας θα σας δωθούν στο εργαστήριο. Για να συμμετέχετε στο διαγωνισμό πατήστε εδώ.

Τα απαραίτητα αρχεία σας δίνονται σε μορφή csv. Τα αρχεία αυτά είναι:

- **train.csv**: Αποτελεί το σύνολο εκπαίδευσης που θα χρησιμοποιηθεί για την εκπαίδευση και βελτιστοποίηση του μοντέλου μηχανικής μάθησης. Επίσης είναι το αρχείο που θα χρησιμοποιήσετε για οποιαδήποτε ανάλυση και διάγραμμα κάνετε.
- **test.csv:** Αποτελεί το σύνολο στο οποίο θα γίνει η πρόβλεψη. Για κάθε παρατήρηση την πιθανότητα εμφάνισης του συγκεκριμένου είδους.
- **sample_submission.csv:** Αρχείο που υποδυκνύει πως θα πρέπει να είναι μία υποβολή των προβλέψεών σας στο σύστημα.

Επιπλέον, σας δίνεται ένα script σε jupyter notebook που περιέχει ένα ολοκληρωμένο απλό παράδειγμα, όπου φορτώνει τα δεδομένα, μετατρέπει τις κατηγορικές τιμές σε αριθμητικές, εκπαιδεύει ένα μοντέλο, το αξιολογεί με 5-fold cross validation και δημιουργεί ένα αρχείο με τις προβλέψεις. Το αρχείο αυτό είναι στη μορφή που απαιτείται για να υποβληθεί στην ιστοσελίδα του διαγωνισμού.

Τα αποτελέσματα εκτέλεσής του αποτελούν το baseline, μία τιμή που έχετε ως στόχο να ξεπεράσετε. Η μετρική που θα χρησιμοποιηθεί είναι η root mean square error.

Α ΜΕΡΟΣ – Data Analysis

Το πρώτη βήμα στην αντιμετώπιση οποιουδήποτε προβλήματος machine learning είναι η ανάλυση των δεδομένων, ώστε να κατανοηθεί το πρόβλημα, να βρεθούν τυχόν ιδιαιτερότητες/λάθη στα δεδομένα και να διορθωθούν ώστε να αυξηθεί η απόδοση του μοντέλου. Να γίνει **ανάλυση**, **περιγραφή** και **οπτικοποίηση** των δεδομένων του dataset.

Συγκεκριμένα, να δημιουργήσετε 5 τουλάχιστον plots. Πιθανά plots είναι το distribution plot μίας μεταβλητής, plot που δείχνει τη συσχέτιση μίας μεταβλητής με το target, και plots με πολλαπλές μεταβλητές.

Έτοιμους κώδικες για plots θα βρείτε εδώ.

Παρακάτω παρουσιάζονται οι κύριες μεταβλητές (features) του dataset.

Όνομα	Αντιστοιχία
Center Lat	Γεωγραφικό πλάτος.
Center Long	Γεωγραφικό μήκος.
Temperature….	Μεταβλητές θερμοκρασίας του νερού.
Salinity….	Αλατότητα.
Secchi….	Καθαρότητα του νερού σε μέτρα.
DissolvedOxygen….	Διαλυμένο οξυγόνο.
Chlorophyll….	Χλωροφύλλη.
EuphoticDepth….	Το στρώμμα της θάλασσας όπου το φως είναι αρκετό για την παραγωγή της χλωροφύλλης.
WaveHeight…	Ύψος κύμματος.
MeridionalCurrent….	Θαλάσσια ρεύματα κατά το γεωγραφικό μήκος.
ZonalCurrent….	Θαλάσσια ρεύματα κατά το γεωγραφικό πλάτος.
Bathymetry	Βάθος θάλασσας.
MajorRiver	Αποσταση από αρχή ή τέλος ποταμού. Επίσης, μέγεθος ποταμού.
Substrate	Διάφορες μεταβλητές για το υπέδαφος.

Β ΜΕΡΟΣ – Machine Learning

Το επόμενο βήμα είναι η εκπαίδευση ενός μοντέλου μηχανικής μάθησης για την πρόβλεψη της πιθανότητας εμφάνισης του είδους σε άγνωστα σημεία. Σας έχει δωθεί ένα jupyter notebook που εκτελεί όλη τη διαδικασία της μηχανικής μάθησης. Σκοπός είναι να το βελτιώσετε με αλλαγές που θα κάνετε στα δεδομένα και στον αλγόριθμο. Αναφέρεται μάλιστα ότι το 70% του χρόνου που αφιερώνει κάποιος για τη βελτίωση του μοντέλου του είναι στην προεπεξεργασία των δεδομένων.

Συγκεκριμένα, μπορείτε να ασχοληθείτε με τα παρακάτω (με+ παρουσιάζεται η δυσκολία) :

+ 1. Να χειριστείτε τα outliers. Παραδείγματα: αλλαγή των τιμών τους με τη μέση τιμή της στήλης, αλλαγή των τιμών τους με μία άλλη τιμή, αφαίρεση των γραμμών που περιέχουν outliers.
+ 2. Να γίνει μετασχηματισμός τιμών που θεωρήθηκαν λανθασμένες. Παραδείγματος χάρη για τη μεταβλητή bathymetry.
+++ 3. Να δημιουργηθούν νέες μεταβλητές από τις υπάρχουσες. Δηλαδή μεταβλητές που προκύπτουν από πράξεις ή ομαδοποιήσεις (groupby με transform) μεταξύ των ήδη υπάρχοντων. Ή ακόμα μπορεί να γίνει apply κάποια function σε τιμές μία μεταβλητής για να δημιουργήσει μία νέα. Ή ακόμα να δημιουργηθούν binary στήλες όπως isBathymetryZero. Η διαδικασία δημιουργίας νέων μεταβλητών ονομάζεται feature engineering και είναι αυτή που μπορεί να καθορίσει την ομάδα με το καλύτερο σκορ. Απαιτεί φαντασία. Παραθέτεται ένας πίνακας με διάφορα χαρακτηριστικά του είδους, τα οποία μπορούν να χρησιμοποιηθούν για την κατασκευή νέων

μεταβλητών:

Πεδίο	**Τιμή**
Environment	Marine, brackish, pelagic-neritic, oceanodromous
DepthRange	0-400
Distribution	Subtropical
DistributionCoordinates	62 North - 37 South, 18 West - 42 East
BiologyInfo	Mainly a coastal marine species, forming large schools. Tolerates salinities of 5-41 ppt and in some areas, enters lagoons, estuaries and lakes, especially during spawning. Tends to move further north and into surface waters in summer, retreating and descending in winter. Feeds on planktonic organisms. Spawns from April to November with peaks usually in the warmest months. Eggs are ellipsoidal to oval, floating in the upper 50 m and hatching in 24-65 hours. Marketed fresh, dried, smoked, canned and frozen; made into fish meal.
LifeAndMateInfo	Pelagic spawners. Gametogenesis is continuous, multiple spawning. Spawning peaks are usually in the warmer months which makes this species a spring-summer spawner. The limits of the spawning season is dependent on temperature and is therefore more restricted in northern areas. Sex ratio: 45% female.
preferredTemperatureMin	7.1
preferredTemperatureMean	10.8
preferredTemperatureMax	18

+++ 4. Με μεθόδους feature selection (π.χ. Recursive Feature Elimination) να βρεθούν οι μεταβλητές που δεν συνεισφέρουν στο μοντέλο και να αποκλειστούν από την εκπαίδευση. Έτσι, θα αυξηθεί περισσότερο η απόδοση.

+++ 5. Με μεθόδους dimensionality reduction να μειωθούν οι μεταβλητές ώστε να αυξηθεί η απόδοση.

\+ 6. Να δοκιμαστούν διάφοροι αλγόριθμοι της βιβλιοθήκης sklearn.

\+ 7. Να δοκιμαστούν διάφορες τιμές για τις παραμέτρους των αλγορίθμων.

++ 8. Να δοκιμαστεί ensembling αλγορίθμων και stacking.

\+ 9. Να δοκιμαστούν διάφορες τεχνικές cross validation.

++ 10. Να γίνει normalization and standardization των μεταβλητών (link).

Να δημιουργηθεί ένας πίνακας όμοιος με τον παρακάτω για την επισύναψη όλων των δοκιμών που κάνατε για τη βελτίωση του μοντέλου:

Public Score	**CV Score**	**Algorithm**	**Parameters**	**Features**	**Experiment**
0.17058	0.19538	Random Forest	min_weight_fraction_leaf=0.05, n_jobs=-2, random_state=0, max_depth=4, n_estimators=100	initial features	(π.χ. replace outliers with mean column value)

Από τα παραπάνω **να υλοποιήσετε τουλάχιστον 4**. Μπορείτε να υλοποιήσετε και περισσότερα αν θέλετε να είστε ανταγωνιστικοί στο leaderboard του διαγωνισμού και να ασχοληθείτε/μάθετε περισσότερα για το machine learning.

Παράδοση και Εξέταση:

Η παράδοση και εξέταση θα γίνει ατομικά. Το παραδοτέο θα είναι ένα zip αρχείο που θα περιλαμβάνει τα jupyter notebooks με τον κώδικα, σχήματα και κείμενο (σχολιασμός, παρατηρήσεις, συμπεράσματα), καθώς για τον πίνακα με τις δοκιμές που κάνατε σε ένα excel ή άλλο τύπο αρχείου. Θα πρέπει να το ανεβάσετε στο e-class μέχρι τις 23:55 της Δευτέρας 31/05/2021. Η εξέταση θα πραγματοποιηθεί την Τρίτη 01/06/2021 σύμφωνα με το πρόγραμμα που θα ανακοινωθεί στο e-class.

Βοηθητικοί σύνδεσμοι:

- https://www.tutorialspoint.com/python/index.htm
- https://machinelearningmastery.com/machine-learning-in-python-step-by-step/
- http://scikit-learn.org/stable/
- https://seaborn.pydata.org/
- https://plot.ly/
- https://deffro.github.io/data%20processing/pandas-tutorial/