Ανάκτηση Πληροφοριών και Εξόρυξη Δεδομένων 2023-2024

2η Εργασία

(Ατομική εργασία)

Data Analysis and Machine Learning

Έχετε στη διάθεσή σας ένα σύνολο δεδομένων (dataset) με 4584 παρατηρήσεις (rows) και 53 μεταβλητές (features or columns). Το dataset αφορά τους αγώνες καλαθοσφαίρισης που πραγματοποιήθηκαν στην Euroleague στο χρονικό διάστημα 2000-2020 και οι παρατηρήσεις είναι τα στατιστικά που προκύπτουν στο τέλος κάθε αγώνα για την γηπεδούχο και την φιλοξενούμενη ομάδα. Στόχος είναι η εκπαίδευση ενός classification μοντέλου σε ένα (train set) για την πρόβλεψη της νίκης της γηπεδούχου ομάδας (τιμές από 0 ή 1, binary classification) σε άγνωστα αγώνες (test set).
Η εργασία έχει και τη μορφή διαγωνισμού όπου θα αναμετρηθείτε μεταξύ σας. Η σελίδα του διαγωνισμού είναι: https://www.kaggle.com/competitions/duth-ir-2023-2024 . Οδηγίες για τη χρήση της ιστοσελίδας θα σας δοθούν στο εργαστήριο. Για να συμμετέχετε στο διαγωνισμό πατήστε εδώ.

https://www.kaggle.com/t/e6708b5d23ca419ba9ce8e1f0367cfca

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 υποδεικνύει πως θα πρέπει να είναι μία υποβολή των προβλέψεών σας στο σύστημα.

Επιπλέον, σας δίνεται ένα script σε Jupiter notebook που περιέχει ένα ολοκληρωμένο απλό παράδειγμα, όπου φορτώνει τα δεδομένα, μετατρέπει τις κατηγορικές τιμές σε αριθμητικές, εκπαιδεύει ένα μοντέλο, το αξιολογεί 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 F1 score.

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 μιας μεταβλητής με το target
- Plots με πολλαπλές μεταβλητές.

Παρακάτω παρουσιάζονται οι κύριες μεταβλητές (features) του dataset (οι μεταβλητές παρουσιάζονται για την γηπεδούχο ομάδα, ισχύουν οι ίδιες ονομασίες για την φιλοξενούμενη αλλάζοντας το H με A στην κάθε μεταβλητή).

Feature	**Αντιστοιχία**
HTEAM	Κωδικοποιημένο όνομα ομάδας, από 2 έως 4 γράμματα
HOME TEAM	Ολόκληρο όνομα ομάδας
HOME WIN	0 για ήττα και 1 για νίκη της ομάδας
H MADE 2 POINTS	Συνολικά επιτυχημένα σουτ 2 πόντων
H TOTAL 2 POINTS	Συνολικά εκτελεσμένα σουτ 2 πόντων
H MADE 3 POINTS	Συνολικά επιτυχημένα σουτ 3 πόντων
H TOTAL 3 POINTS	Συνολικά εκτελεσμένα σουτ 3 πόντων
H MADE 1 POINTS	Συνολικές επιτυχημένες βολές (σουτ 1 πόντου)
H TOTAL 1 POINTS	Συνολικές εκτελεσμένες βολές (σουτ 1 πόντου)
H OFFENSIVE REBOUNDS	Αριθμός αμυντικών ριμπάουντ
H DEFENSIVE REBOUNDS	Αριθμός επιθετικών ριμπάουντ
H TOTAL REBOUNDS	Συνολικός αριθμός ριμπάουντ (επιθετικών και αμυντικών)
H ASSISTS	Αριθμός τελικών πασών
H STEALS	Αριθμός κλεψιμάτων
H TURNOVERS	Αριθμός λαθών
H BLOCKS FOR	Αριθμός μπλοκ που πραγματοποίησε η ομάδα
H BLOCKS AGAINST	Αριθμός μπλοκ που δέχθηκε η ομάδα
H FOULS COMMITTED	Αριθμός φάουλ που διέπραξε η ομάδα
H FOULS RECEIVED	Αριθμός φάουλ που δέχθηκε η ομάδα
H PIR	Συνολικός αριθμός αξιολόγησης της ομάδας
H Q1	Σύνολο πόντων που επιτεύχθηκαν στην $1^η$ περίοδο
H Q2	Σύνολο πόντων που επιτεύχθηκαν στην $2^η$ περίοδο
H Q3	Σύνολο πόντων που επιτεύχθηκαν στην $3^η$ περίοδο
H Q4	Σύνολο πόντων που επιτεύχθηκαν στην $4^η$ περίοδο
H O1	Σύνολο πόντων που επιτεύχθηκαν στην $1^η$ παράταση
H O2	Σύνολο πόντων που επιτεύχθηκαν στη $2^η$ παράταση

Β ΜΕΡΟΣ – Machine Learning

Το επόμενο βήμα είναι η εκπαίδευση ενός μοντέλου μηχανικής μάθησης για την πρόβλεψη της νίκης της γηπεδούχου ομάδας σε άγνωστους αγώνες. Σας έχει δοθεί 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 τις ακραίες τιμές (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 χειριστείτε των κενών τιμών (missing values).

+ 3. Να γίνει μετασχηματισμός τιμών που θεωρήθηκαν λανθασμένες. Παραδείγματος χάρη για τη μεταβλητή bathymetry.

+++ 4. Να δημιουργηθούν νέες μεταβλητές από τις υπάρχουσες. Δηλαδή μεταβλητές που προκύπτουν από πράξεις ή ομαδοποιήσεις (groupby με transform) μεταξύ των ήδη υπαρχόντων. Ή ακόμα μπορεί να γίνει εφαρμογή κάποιας συνάρτησης σε τιμές μιας μεταβλητής για να δημιουργήσει μία νέα. Ή ακόμα να δημιουργηθούν binary στήλες όπως η HOME WIN / AWAY WIN. Η διαδικασία δημιουργίας νέων μεταβλητών ονομάζεται feature engineering και είναι αυτή που μπορεί να καθορίσει την ομάδα με το καλύτερο σκορ. Απαιτεί φαντασία. Παραθέτετε ένας πίνακας με διάφορα χαρακτηριστικά, τα οποία μπορούν να χρησιμοποιηθούν για την κατασκευή νέων μεταβλητών:

Μεταβλητές	**Τυπικές τιμές**
MADE 2 POINTS	0-40
TOTAL 2 POINTS	0-40
MADE 3 POINTS	0-30
TOTAL 3 POINTS	0-30
MADE 1 POINTS	0-50
TOTAL 1 POINTS	0-50
ASSISTS	0-40
TURNOVERS	0-30
Q1	0-70
Q2	0-40
Q3	0-40
Q4	>0

+++ 5. 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.

+++ 6. Με μεθόδους dimensionality reduction να μειωθούν οι μεταβλητές ώστε να αυξηθεί η απόδοση.

+7. Να δοκιμαστούν διάφοροι αλγόριθμοι της βιβλιοθήκης sklearn.

++ 8. Να προσπαθήσετε να βελτιστοποιήσετε τις μεταβλητές των αλγορίθμων. Η τεχνική αυτή ονομάζεται hyperparameter tuning και μπορεί να συνεισφέρει ουσιαστικά στην βελτίωση του μοντέλου. Υπάρχουν αυτοματοποιημένες τεχνικές για την επίτευξη αυτού του στόχου.

++ 9. Να δοκιμαστεί ensembling αλγορίθμων και stacking.

+ 10. Να γίνει normalization and standardization των μεταβλητών.

Να δημιουργηθεί ένας πίνακας όμοιος με τον παρακάτω πίνακα

Public score	**Algorithm**	**Parameters**	**Features**	**Experiment**
0.909	K Nearest Neighbors		Initial features	π.χ χρήση standardization τενχικής

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 17/06/2024. Η εξέταση θα πραγματοποιηθεί την Τρίτη 18/06/2024.

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/