ΠΜΣ ΣΤΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΤΕΧΝΟΛΟΓΙΑ
Τμήμα Λογιστικής και Χρηματοοικονομικής
Πανεπιστήμιο Μακεδονίας

ΕΝΤΥΠΟ ΥΠΟΒΟΛΗΣ ΠΡΟΤΑΣΗΣ ΘΕΜΑΤΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

(για ένταξη στην Τράπεζα Θεμάτων Διπλωματικών Εργασιών του Π.Μ.Σ.)

1. ΣΤΟΙΧΕΙΑ ΠΡΟΤΑΣΗΣ

Πεδίο	Στοιχεία
Κωδικός Θέματος *(συμπληρώνεται από τη Γραμματεία μετά την έγκριση του θέματος από τη Συντονιστική Επιτροπή)*	
Ημερομηνία Υποβολής	28/06/2026
Προτείνων	Καθηγητής Νικόλαος Σαμαράς
Φορέας Προέλευσης Θέματος *(FAC, FTSAI, RES, IND, STU, EXT)*[1]	FAC
Κύρια Θεματική Περιοχή *(FINTECH, , RISK, AI-DATA,, DLT, REG, GOV, PROG, IND)*	AI-DATA
Δευτερεύουσα Θεματική Περιοχή *(προαιρετικά)*	FINTECH
Τριτεύουσα Θεματική Περιοχή *(προαιρετικά)*	--

[1] **FAC:** Μέλος Δ.Ε.Π. ή Διδάσκων του Π.Μ.Σ., **FTSAI:** Financial Technology and Strategic Artificial Intelligence Laboratory, **RES:** Άλλη ερευνητική δομή ή ερευνητικό έργο, **IND:** Επιχείρηση ή οργανισμός, **STU:** Πρόταση φοιτητή, **EXT:** Εξωτερικός συνεργάτης ή φορέας.

2. ΤΙΤΛΟΣ ΘΕΜΑΤΟΣ

Τίτλος στα Ελληνικά

Βιβλιοθήκη Machine (Supervised) Learning, για αξιοποίηση/βελτιστοποίηση μικρών πινακοποιημένων συνόλων δεδομένων

Title in English

A Python library for (supervised) machine learning, optimized for small tabular datasets

Acronym: **PETAL** — **P**ython **E**stimators for **TA**bular **L**earning

3. ΚΑΤΗΓΟΡΙΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

- ☒ Ερευνητική Διπλωματική Εργασία
- ☐ Εφαρμοσμένη Διπλωματική Εργασία
- ☒ Τεχνολογική Διπλωματική Εργασία
- ☐ Διπλωματική σε Συνεργασία με Οργανισμό ή Επιχείρηση
- ☐ Διπλωματική Ενταγμένη σε Ερευνητική Δράση

Εφόσον επιλεγεί η τελευταία κατηγορία — Όνομα Ερευνητικής Δράσης

4. ΜΕΘΟΔΟΛΟΓΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

- ☐ Βιβλιογραφική ή Θεωρητική Μελέτη
- ☐ Συστηματική Βιβλιογραφική Ανασκόπηση
- ☐ Εμπειρική ή Ποσοτική Ανάλυση
- ☐ Μελέτη Περίπτωσης
- ☒ Συγκριτική Ανάλυση
- ☒ Ανάπτυξη ή Αξιολόγηση Τεχνολογικού Συστήματος
- ☒ Σχεδιασμός Πλαισίου, Μεθοδολογίας ή Μοντέλου
- ☐ Μικτή Προσέγγιση

5. ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΘΕΜΑΤΟΣ

(ενδεικτική έκταση: 100 έως 250 λέξεις)

Στην παρούσα διπλωματική εργασία θα αναπτυχθεί μια βιβλιοθήκη σε Python, με σκοπό τη διαχείριση μικρού όγκου δεδομένων, καθιστώντας την Επιστήμη των Δεδομένων πιο προσιτή και αξιοποιήσιμη από μεγαλύτερο σύνολο της οικονομίας. Από Στατιστικής πλευράς , γνωρίζουμε ότι τα μεγάλα σύνολα δεδομένων επιλύουν μόνα τους ένα σημαντικό μέρος του προβλήματος της εκπαίδευσης, καθώς το μέγεθος του δείγματος n με τα μέτρα απόδοσης (discrimination metrics) ενός μοντέλου έχουν πολύ υψηλή συσχέτιση. Ωστόσο τα Big Data, πέραν του ότι είναι κοστοβόρα, είναι και δυσεύρετα από τις μικρομεσαίες επιχειρήσεις / οργανισμούς, καθώς αυτοί σπάνια έχουν την ευκαιρία να συλλέξουν και να συντηρήσουν ή να εξαγοράσουν βάσεις μεγάλου όγκου δεδομένων. Π.χ. είναι ένας manager και έχει ένα data set με 20 εγγραφές (raw data) και θέλει να πάρει μια απόφαση έστω και με 60-70% accuracy. (μια προσέγγιση που σε βάθος χρόνου υπερτερεί σημαντικά της τυχαίας / εμπειρικής λήψης αποφάσεων). Έτσι, η Επιστήμη των Δεδομένων παραμένει απλησίαστη για μεγάλο κομμάτι των δραστηριοτήτων σε ένα οικονομικό σύνολο.

6. ΣΤΟΧΟΙ ΚΑΙ ΕΡΕΥΝΗΤΙΚΑ ΕΡΩΤΗΜΑΤΑ

Α. Στόχοι

- Ανάπτυξη βιβλιοθήκης στην Python
- Να παραχθεί τεκμηρίωση, παραδείγματα χρήσης και πακετάρισμα (π.χ. διάθεση μέσω PyPI), ώστε το αποτέλεσμα να είναι πραγματικά αξιοποιήσιμο
- Να γίνει συγκριτική αποτίμηση (benchmarking) έναντι υπαρχόντων εργαλείων σε καθιερωμένα σύνολα αναφοράς (OpenML, UCI)

Β. Ερευνητικά Ερωτήματα

- Συγκριτική απόδοση και επιλογή μοντέλου για μικρό όγκο δεδομένων (Εδώ το ερώτημα δεν είναι «ποιο μοντέλο είναι καλύτερο» γενικά, αλλά υπό ποιες *συνθήκες δεδομένων)*
- Ποιες οικογένειες μοντέλων (γραμμικά με ομαλοποίηση, δεντρικά/ensembles, gradient boosting, k-NN, πιθανοτικά) υπερτερούν συστηματικά σε μικρά πινακοποιημένα σύνολα, και πώς εξαρτάται αυτό από τα χαρακτηριστικά του συνόλου (λόγος δειγμάτων/χαρακτηριστικών n/p, ανισορροπία κλάσεων, τύπος μεταβλητών, θόρυβος);

- Πόσο αναξιόπιστες είναι οι διάφορες διαδικασίες εκτίμησης απόδοσης (απλή διαίρεση, k-fold, επαναλαμβανόμενη και εμφωλευμένη διασταυρωμένη επικύρωση) όσο μειώνεται το n, και ποιο είναι το «κόστος διασποράς» κάθε προσέγγισης;

7. ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΛΟΓΙΑ

System Architecture: Η Βιβλιοθήκη θα δέχεται το DataFrame, θα μπορεί να κάνει training το μοντέλο σε Γραμμική / Λογιστική Παλινδρόμηση και σε Δέντρα Αποφάσεων και να επιστρέφει τα σχετικά αποτελέσματα πρόβλεψης.

1. Backend:

- Στο backend , θα κυριαρχεί η NumPy στη μετατροπή για δύο λόγους. 1. Είναι γραμμένη σε C και θα τρέχει πολύ πιο γρήγορα από τα in / for loops της Python , και 2. Θα κάνει vectorisation των δεδομένων μας.
- Χρήση JIT(Just in Time) Compilerer της Python στο Training , επίσης για λόγους ταχύτητας.

2. Computational Logic:

Το Dataset θα μεταμορφώνεται σε μεγαλύτερο με στοχαστικές προσομοιώσεις χρησιμοποιώντας τον αλγόριθμο Metropolis Hastings (MCMC). Το μοντέλο θα ορίζει τα Priors βάση του αρχικού μας dataset, και ο αλγόριθμος θα εκτελεί τη στοχαστική διαδικασία δημιουργώντας κατανομές (weighted) για κάθε τιμή. Το μοντέλο θα αναλύει τα δεδομένα καταλήγοντας σε ένα ιστορικό κατανομών, το οποίο θα χαρτογραφεί την αβεβαιότητα (variance) του αρχικού dataset.
Μετά, με τη χρήση του Likelihood των κατανομών που δημιουργήθηκαν (Τις επικρατούσες τιμές, means) , η βιβλιοθήκη θα παράγει νέα δεδομένα μεγαλύτερου όγκου (posterior) που θα διατηρούν τα αρχικά descriptive statistics του small dataset (variance, correlations), με τον απαραίτητο θόρυβο της τυχαιότητας ώστε να μην έχουμε το κλασσικό overfitting των Small data.

3. User Interface - Outputs:

Χρήση συναρτήσεων του Scikit-Learn όπου ο χρήστης θα καλεί για να δομήσει / εκτελέσει το μοντέλο.

Discrimination Metrics: Χρήση μετρικών Bayesian Performance Metrics [Posterior Probabilities, Likelihood Ratios, Misclassification Error Rate, Posterior Predictive Power] και Discrimination Measures [Bayes Risk, Bayes Factor]

Model Validation (Benchmarking): Χρήση στατιστικών για το έλεγχο της εγκυρότητας των αποτελεσμάτων.

Ideas for Deployment:

- **Live Web Dashboards:** Άμεσο upload δεδομένων και output χρήσιμων μετρικών στη λήψη αποφάσεων.
- **REST APIs:** Δυνατότητα κλήσης του μοντέλου από εξωτερικές εφαρμογές ή ενσωμάτωση σε λογισμικό
- **SmartPhone py-app:** Έξυπνες και άμεσες αποφάσεις εισάγοντας λίγα νούμερα στο Κινητό μας μέσα σε λίγα δευτερόλεπτα / λεπτά.

8. ΔΕΔΟΜΕΝΑ ΚΑΙ ΠΗΓΕΣ ΔΕΔΟΜΕΝΩΝ

https://www.kaggle.com/

https://archive.ics.uci.edu/

https://www.kaggle.com/datasets?tags=11108-Finance

https://finance.yahoo.com/

9. ΠΡΟΑΠΑΙΤΟΥΜΕΝΕΣ ΓΝΩΣΕΙΣ Ή ΔΕΞΙΟΤΗΤΕΣ

A. Hard skills

- Βασικές Γνώσεις Στατιστικής
- Βασικές Γνώσεις Γραμμικής Άλγεβρας
- Γνώσεις αλγορίθμων και μεθοδολογιών Μηχανικής Μάθησης
- Θεωρητικό υπόβαθρο μεθόδων παλινδρόμησης
- Γνώσεις Προγραμματισμού σε Python
- Συγγραφή κειμένου σε LaTeX

B. Soft skills

- Σύνταξη κατάλληλων prompts σε AI tools
- Ορθή διαχείριση χρόνου
- Ανάληψη πρωτοβουλιών
- Κριτική σκέψη

10. ΑΝΑΜΕΝΟΜΕΝΑ ΠΑΡΑΔΟΤΕΑ

Μια ολοκληρωμένη βιβλιοθήκη σε Python με φιλικό προς τον χρήστη interface (Graphical User Interface-GUI).
Manual χρήστη
Κείμενο διπλωματικής εργασίας
Paper με καταγραφή των αποτελεσμάτων σε επιστημονικό Conference/Journal

11. ΕΚΤΙΜΗΣΗ ΔΥΣΚΟΛΙΑΣ

Κλίμακα 1 (χαμηλή) έως 5 (υψηλή)

Κατηγορία	1	2	3	4	5
Θεωρητική Δυσκολία	☐	☐	☐	☐	☒
Προγραμματιστική Δυσκολία	☐	☐	☐	☐	☒
Δυσκολία Συλλογής Δεδομένων	☐	☐	☐	☒	☐

12. ΕΚΤΙΜΗΣΗ ΔΥΣΚΟΛΙΑΣ

☐ Χαμηλή

☐ Μέτρια

☒ Υψηλή

Σύντομη αιτιολόγηση

Ο χώρος είναι ώριμος και ανταγωνιστικός. Ο μεταπτυχιακός φοιτητής οφείλει να τοποθετηθεί σε σχέση με ισχυρά υπάρχοντα εργαλεία/βιβλιοθήκες, όπως: τις οικογένειες gradient boosting (XGBoost, LightGBM, CatBoost) που κυριαρχούν στα πινακοποιημένα δεδομένα, τα πλαίσια AutoML (auto-sklearn, TPOT, FLAML), καθώς και προσεγγίσεις σχεδιασμένες ειδικά για μικρά πινακοποιημένα σύνολα όπως το TabPFN. Υπάρχει επίσης σχετική βιβλιογραφία για το γιατί τα δεντρικά μοντέλα συχνά υπερτερούν του deep learning στα πινακοποιημένα δεδομένα.

Συντελεστές που *μειώνουν* τη δυσκολία: δόμηση πάνω στο scikit-learn αντί για υλοποίηση από το μηδέν.

Συντελεστές που *αυξάνουν* τη δυσκολία: στόχευση σε νέες μεθόδους με θεωρητική θεμελίωση, ευρύ εύρος αλγορίθμων, αυστηρή εμπειρική επικύρωση με στατιστικούς ελέγχους και φιλοδοξία η βιβλιοθήκη να είναι όντως ανταγωνιστική.