

# Εργασία 1 – Επιβλεπόμενη Μάθηση

## Μέρος 1 – Λογιστική Παλινδρόμηση

Ιδιωτική κλινική θέλει να κατασκευάσει σύστημα πρόβλεψης διαβήτη, ώστε να αυτοματοποιήσει την ανίχνευση διαβητικών ασθενών. Επιπλέον, διαθέτει στη διάθεση του ένα σύνολο εξετάσεων 768 ασθενών, που περιλαμβάνουν:

1. **Pregnancies:** Εγκυμοσύνες.
2. **Glucose:** Συγκέντρωση γλυκόζης.
3. **Blood Pressure:** Αρτηριακή πίεση.
4. **Skin Thickness:** Πλάχος δέρματος στους τρικέφαλους.
5. **Insulin:** Συγκέντρωση ινσουλίνης.
6. **BMI:** Δείκτης μάζας-σώματος.
7. **Diabetes Pedigree Function:** Σκορ που εκφράζει πιθανότητα εμφάνισης διαβήτη με βάση το οικογενειακό ιστορικό.
8. **Age:** Ηλικία
9. **Outcome:** Αποτέλεσμα εξετάσεων (0: Αρνητικός, 1: Θετικός).

## Ερωτήματα

1. Φορτώστε το σύνολο δεδομένων [diabetes.csv](#) σε ένα DataFrame μέσω της βιβλιοθήκης pandas. Στη συνέχεια, περιγράψτε το κάθε χαρακτηριστικό με μέση τιμή, τυπική απόκλιση, ελάχιστη και μέγιστη τιμή (df.describe()). Τέλος, δημιουργήστε το ιστόγραμμα (Histogram) για κάθε χαρακτηριστικό. Για το Outcome, να δημιουργηθεί ραβδόγραμμα (Bar Plot), εφόσον έχει μόνο 2 τιμές.
2. Θεωρείτε πως η ποιότητα των δεδομένων είναι καλή ή κακή? Αιτιολογείστε, αξιοποιώντας τις πληροφορίες από ερώτημα (1). Δώστε τουλάχιστον 2 επιχειρήματα.
3. Τί κατανομή ακολουθεί η μεταβλητή Age? Είναι καλή η κατανομή αυτή για τη κατασκευή της συγκεκριμένης εφαρμογής? Αιτιολογείστε.
4. Σύμφωνα με κλινικές μελέτες, αν κάποιος ασθενής έχει υψηλά επίπεδα γλυκόζης, είναι πιολύ πιθανό να εμφανίσει διαβήτη. Να Δείξετε αν και πως διαπιστώνεται αυτό από τα δεδομένα.
5. Δημιουργήστε Numpy arrays με inputs (x) και targets (y), όπου στο x περιλαμβάνονται όλα τα χαρακτηριστικά εκτός του outcome και y το outcome. Στη συνέχεια, χωρίστε το σύνολο δεδομένων σε σύνολα εκπαίδευσης-επικύρωσης (train-validation) με ποσοστό 70-30%, χρησιμοποιώντας ως seed (random state) το 0. Εμφανίστε το πλήθος των παραδειγμάτων εκπαίδευσης και επικύρωσης. Θα χρειαστείτε να φορτώσετε τη συνάρτηση train\_test\_split της scikit-learn.
6. Εκπαιδεύστε ταξινομητή (Classifier) Λογιστικής Παλινδρόμησης (Logistic Regression) στο train set και μετρήστε την ακρίβεια (accuracy score) του στα σύνολα train και test. Χρησιμοποιήστε ως seed το 0. Μπορείτε να συμβουλευτείτε το documentation της scikit-learn [https://scikit-learn.org/1.5/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html). Τι παρατηρείτε για την ακρίβεια του test set σε σχέση με του train set?
7. Δουλεύει καλά ο ταξινομητής σας για όλες τις ηλικίες? Παρουσιάστε ραβδόγραμμα (bar plot) με την ακρίβεια ανά ομάδα ηλικιών 0 ως 25, 25 ως 50 και >50 του test set.

8. Τι θεωρείτε πως είναι χειρότερο για τον ταξινομητή σας στη συγκεκριμένη εφαρμογή, να προβλέπει ότι κάποιος ασθενής έχει διαβήτη, χωρίς να έχει, ή ότι κάποιος ασθενής δεν έχει διαβήτη, ενώ έχει? Αιτιολογήστε.
9. Επαναλάβετε τις διαδικασίες 5-6 (με μία for loop), χρησιμοποιώντας seed από 0 ως 9. Υπολογίστε μέσο όρο και τυπική απόκλιση της ακρίβειας σας.
10. Εφαρμόστε κανονικοποίηση των δεδομένων Min-Max:  $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$  (εκτός της μεταβλητής Target). Στη συνέχεια, να αναφέρετε τους 3 πιο σημαντικούς παράγοντες (συντελεστές) που επηρεάζουν την πρόβλεψη του διαβήτη. Αιτιολογείστε.

## Μέρος 2 – Γραμμική Παλινδρόμηση

Η ίδια ιδιωτική κλινική θέλει να εντοπίζει πιθανούς διαβητικούς μέσω μιας έξυπνης εφαρμογής τηλεφώνου. Οι χρήστες θα μπορούν να εισάγουν δεδομένα που απαιτούν ελάχιστο ιατρικό εξοπλισμό, όπως **εγκυμοσύνες, αρτηριακή πίεση, BMI και ηλικία** και η εφαρμογή θα προβλέπει το επίπεδο γλυκόζης στο αίμα. Αν είναι πάνω από 170, τότε θα τους προτείνει να προσέλθουν για εξετάσεις.

### Ερωτήματα

1. Φορτώστε το σύνολο δεδομένων [diabetes.csv](#). Δημιουργήστε Numpy arrays με inputs (x) και targets (y), όπου x: (Pregnancies, Blood Pressure, BMI, Age) και y η μεταβλητή Glucose. Χωρίστε το σύνολο δεδομένων σε σύνολα εκπαίδευσης-επικύρωσης (train-validation) με ποσοστό 70-30% με 0 seed.
2. Χρησιμοποιείστε Γραμμική παλινδρόμηση της scikit-learn [https://scikit-learn.org/dev/modules/generated/sklearn.linear\\_model.LinearRegression.html](https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.LinearRegression.html), ώστε να προβλέψετε την ποσότητα γλυκόζης στο test set και μετρήστε την ακρίβεια με κατάλληλη μετρική. Ποια μετρική είναι καταλληλότερη: Mean Squared Error (MSE) ή Mean Absolute Error (MAE)? Αιτιολογείστε.
3. Επαναλάβετε το ερώτημα (2) χρησιμοποιώντας Lasso Regression [https://scikit-learn.org/dev/modules/generated/sklearn.linear\\_model.Lasso.html](https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.Lasso.html) (Linear + L1). Δοκιμάστε τιμές (0.2, 0.4, 0.6, 0.8, 1.0) για το βάρος λ (alpha στην scikit-learn) και κατασκευάστε πινακάκι με την ακρίβεια, χρησιμοποιώντας τη μετρική του ερωτήματος 2.
4. Επαναλάβετε το ερώτημα 2, αφαιρώντας από το X το χαρακτηριστικό με τη χαμηλότερη βαρύτητα (συντελεστή) και συγκρίνεται την επίδοση του μοντέλου σας με το (2). Για να βρείτε αυτό το χαρακτηριστικό, υλοποιήστε το ερώτημα 10, εφαρμόζοντας όμως γραμμική παλινδρόμηση.

### Οδηγίες

- Χρησιμοποιείστε την πλατφόρμα [Google Colab](#) για την υλοποίηση της άσκησης.
- Τα plots/πινακάκια να εμφανίζονται επάνω στο Colab. Επίσης, μπορείτε να εισάγετε κελιά για κείμενο και πινακάκια όπως φαίνεται στον παρακάτω σύνδεσμο:  
[https://colab.research.google.com/notebooks/markdown\\_guide.ipynb](https://colab.research.google.com/notebooks/markdown_guide.ipynb)

- Δώστε έμφαση στην παρουσίαση της εργασίας. Copy-Paste από το ChatGPT θα μηδενίζονται.
- Στο elearning θα υποβάλλετε το link της εργασίας σας στο Google Colab.