Web App for Data Analysis, 2024, pp. 1–3

doi:
Publication Date: Day Month Year 1/9/2024

Web App για Αξιολόγηση και Οπτικοποίηση Αλγορίθμων Μηχανικής Μάθησης

Author: Μπάρλας Ιωάννης Π2019009

Abstract

Αυτή η εργασία παρουσιάζει μια web εφαρμογή που αναπτύχθηκε για την αξιολόγηση και οπτικοποίηση αλγορίθμων machine learning (ML). Ξεκινά με την εισαγωγή και τον καθαρισμό των δεδομένων, ακολουθούμενη από τη μετατροπή και κανονικοποίηση τους.Στη συνέχεια, επιλέγονται τα χαρακτηριστικά που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου, και τα δεδομένα διαχωρίζονται σε εκπαιδευτικό και δοκιμαστικό σετ. Ο χρήστης εκπαιδεύει το μοντέλο με διάφορους αλγόριθμους ταξινόμησης, αξιολογεί την απόδοση του μέσω μετρικών, και βελτιστοποιεί τις παραμέτρους τους για καλύτερα αποτελέσματα.Το μοντέλο χρησιμοποιείται για να προβλέψει κατηγορίες για νέα δεδομένα, προσφέροντας πολύτιμα στοιχεία για λήψη αποφάσεων.

Εισαγωγή

Η μηχανική μάθηση αποτελεί κεντρικό πυλώνα της τεχνητής νοημοσύνης, με εφαρμογές που εκτείνονται από την αναγνώριση προτύπων και την επεξεργασία φυσικής γλώσσας, έως τη ρομποτική και την αυτόνομη οδήγηση. Η αυξανόμενη πολυπλοκότητα και ποικιλία των αλγορίθμων ΜΛ καθιστούν την επιλογή του βέλτιστου μοντέλου για ένα συγκεκριμένο πρόβλημα μια σημαντική πρόκληση. Η αξιολόγηση της απόδοσης των μοντέλων και η ερμηνεία των αποτελεσμάτων τους είναι κρίσιμα βήματα για την ανάπτυξη αποτελεσματικών συστημάτων ΜΛ. Παρά την αφθονία των διαθέσιμων εργαλείων ΜΛ, η πλειονότητα επικεντρώνεται στην εκπαίδευση και ανάπτυξη μοντέλων, αφήνοντας ένα κενό στην ολοκληρωμένη αξιολόγηση και οπτικοποίηση της απόδοσης τους. Επιπλέον, πολλά από αυτά τα εργαλεία απαιτούν εξειδικευμένες γνώσεις προγραμματισμού, περιορίζοντας την προσβασιμότητά τους. Για να αντιμετωπίσουμε αυτές τις προκλήσεις, αναπτύξαμε μια διαδικτυακή εφαρμογή που παρέχει ένα φιλικό προς το χρήστη περιβάλλον για την αξιολόγηση και την οπτικοποίηση αλγορίθμων ΜΛ. Η εφαρμογή επιτρέπει στους χρήστες να συγκρίνουν την απόδοση διαφορετικών αλγορίθμων σε ποικίλα σύνολα δεδομένων, αξιοποιώντας τεχνικές όπως η Principal Component Analysis (PCA) η UMAP (Uniform Manifold Approximation and Projection) και η (EDA) Exploratory Data Analysis για την οπτικοποίηση των αποτελεσμάτων ενώ ακόμα μπορούν να συγκρίνουν τους αλγορίθμους ταξινόμησης KNN,Random Forest.

Περιγραφή της Εφαρμογής

Η εφαρμογή έχει σχεδιαστεί με γνώμονα την ευχρηστία και την προσβασιμότητα. Αναπτύχθηκε χρησιμοποιώντας ένα συνδυασμό τεχνολογιών ωεβ, όπως ΗΤΜΛ, ΎΣΣ και ΘααΣςριπτ, για την παρουσίαση του περιβάλλοντος εργασίας και την διαχείριση της αλληλεπίδρασης με τον χρήστη. Για την υλοποίηση των αλγορίθμων μηχανικής μάθησης και των τεχνικών οπτικοποίησης, αξιοποιήσαμε τη βιβλιοθήκη Πψτηον σςικιτ-λεαρν, η οποία παρέχει ένα ευρύ φάσμα εργαλείων για ανάλυση δεδομένων και μοντελοποίηση.

Περιγραφή Αρχικής Οθόνης

Η αρχική οθόνη της εφαρμογής μας αποτελεί την πύλη εισόδου για τους χρήστες, παρέχοντας μια φιλική και ελκυστική εμπειρία που διευκολύνει την αξιολόγηση και οπτικοποίηση αλγορίθμων μηχανικής μάθησης. Η οθόνη είναι δομημένη σε μια κύρια κεντρική περιοχή.
Στην κορυφή της οθόνης, υπάρχει μια γραμμή τίτλου που εμφανίζει το όνομα της εφαρμογής, ¨Εφαρμογή Εξόρυξης και Ανάλυσης Δεδομένων¨. Ακριβώς κάτω από αυτόν τον τίτλο, υπάρχει μια γραμμή πλοήγησης με πέντε ταβς: "Data Upload", "Visualization", "Feature Selection", "Classification", και "Info Tab".

"Data Upload": Εδώ, ο χρήστης μπορεί να ανεβάσει τα δεδομένα του μέσω ενός uploader που υποστηρίζει αρχεία CSV, Excel, TSV. Στην περίπτωση που ανέβει ένα αρχείο, εμφανίζεται ο πίνακας δεδομένων, και δίνεται η δυνατότητα επιλογής της στήλης στόχου. Το αναδιοργανωμένο dataset εμφανίζεται και είναι έτοιμο για περαιτέρω ανάλυση.

"Visualization": Σε αυτό το tab, παρουσιάζονται διάφορες οπτικοποιήσεις των δεδομένων. Περιλαμβάνονται 2D και 3D διαγράμματα PCA και UMAP, τα οποία βοηθούν στην απεικόνιση και κατανόηση των δεδομένων σε μειωμένες διαστάσεις. Επίσης, παρέχονται εργαλεία για Exploratory Data Analysis (EDA) όπως heatmaps.

"Feature Selection": Εδώ, ο χρήστης μπορεί να επιλέξει τον αριθμό των χαρακτηριστικών που επιθυμεί να κρατήσει για ανάλυση. Με την επιλογή του αριθμού, η εφαρμογή εφαρμόζει την τεχνική SelectKBest για να επιλέξει τα καλύτερα χαρακτηριστικά και να παρουσιάσει τα αποτελέσματα.

"Classification": Σε αυτό το tab, ο χρήστης μπορεί να επιλέξει παραμέτρους για δύο αλγόριθμους ταξινόμησης: K-Nearest Neighbors (KNN) και Random Forest. Οι αλγόριθμοι εκτελούνται πριν και μετά την επιλογή χαρακτηριστικών, και η εφαρμογή παρουσιάζει συγκριτικά αποτελέσματα, όπως accuracy, F1-score , ROC-AUC.

"Info": Αυτό το tab παρέχει πληροφορίες για την εφαρμογή, περιγράφοντας τη λειτουργία της και την ομάδα ανάπτυξης.

Εισαγωγή Δεδομένων

Η εφαρμογή δίνει τη δυνατότητα στον χρήστη να ανεβάσει αρχεία δεδομένων σε διάφορες μορφές, όπως:
CSV (Comma Separated Values)Excel (.xlsx) TSV (Tab Separated Values). Ο κώδικας για την εισαγωγή δεδομένων διαβάζει τα αρχεία χρησιμοποιώντας τη βιβλιοθήκη Πανδας, η οποία είναι ευρέως χρησιμοποιούμενη για την ανάλυση δεδομένων.Ο τρόπος ανάγνωσης του αρχείου γίνεται μέσω του pd.readcsv.

Διαδικασία: Ο χρήστης ανεβάζει το αρχείο μέσω της διεπαφής του Streamlit. Το πρόγραμμα αναγνωρίζει αυτόματα τον τύπο του αρχείου από την επέκταση και το διαβάζει ανάλογα. Μόλις τα δεδομένα φορτωθούν, εμφανίζονται στον χρήστη με την εντολή st.dataframe, δίνοντας τη δυνατότητα να δει το dataset.
Στήλη Στόχος (Target Column): Μετά την εισαγωγή των δεδομένων, ο χρήστης πρέπει να επιλέξει τη στήλη στόχο, δηλαδή τη μεταβλητή που θέλει να προβλέψει. Αυτή η στήλη μπορεί να είναι μια κατηγορική μεταβλητή (π.χ. τύπος πελάτη,) .

Διαδικασία: Ο χρήστης επιλέγει από ένα dropdown menu τη στήλη που θέλει να ορίσει ως στήλη στόχο. Το dataset αναδιαμορφώνεται έτσι ώστε η στήλη στόχος να μετακινείται στο τέλος. Αυτό διευκολύνει την επεξεργασία στη συνέχεια, καθώς πολλές βιβλιοθήκες κατηγοριοποίησης υποθέτουν ότι η στήλη στόχος βρίσκεται στο τέλος.

Οπτικοποιήσεις **(Visualizations)**

Το πρόγραμμα προσφέρει ποικιλία από εργαλεία για την οπτικοποίηση των δεδομένων, που βοηθούν τον χρήστη να κατανοήσει καλύτερα τη δομή τους.

PCA (Principal Component Analysis) PCA είναι ένας αλγόριθμος μείωσης διάστασης που χρησιμοποιείται για να προβάλει τα δεδομένα σε λιγότερες διαστάσεις.Αυτό διευκολύνει την απεικόνιση πολύπλοκων δεδομένων και δίνει μια γενική αίσθηση της διάταξης των δεδομένων.

UMAP (Uniform Manifold Approximation and Projection) UMAP είναι μια άλλη τεχνική μείωσης διάστασης, που τείνει να διατηρεί καλύτερα την τοπική δομή των δεδομένων από το PCA. Συχνά χρησιμοποιείται σε μεγάλα και περίπλοκα δατασετς, καθώς είναι πιο ικανό να διατηρεί τις μη γραμμικές σχέσεις.

Boxplot Boxplot είναι ένα εργαλείο οπτικοποίησης που βοηθά στην απεικόνιση της κατανομής των δεδομένων και την ανίχνευση ανωμαλιών ή outliers.

Heatmap Heatmap είναι ένας χάρτης θερμότητας που χρησιμοποιείται για να απεικονίσει τη συσχέτιση μεταξύ των χαρακτηριστικών, βοηθώντας τον χρήστη να κατανοήσει ποιες μεταβλητές μπορεί να σχετίζονται περισσότερο με τη στήλη στόχο ή μεταξύ τους.

Επιλογή Χαρακτηριστικών **(Feature Selection)**

Η επιλογή χαρακτηριστικών είναι μια κρίσιμη διαδικασία στην ανάλυση δεδομένων, όπου ο στόχος είναι να διατηρηθούν μόνο τα πιο σημαντικά χαρακτηριστικά για την πρόβλεψη της στήλης στόχου. Εδώ, χρησιμοποιείται ο αλγόριθμος SelectKBest, που εφαρμόζει το ANOVA F-test.

ANOVA F-test: Αυτός ο στατιστικός έλεγχος υπολογίζει τη διαφορά μεταξύ των διακυμάνσεων μέσα σε ομάδες και μεταξύ ομάδων για να καθορίσει τη σημαντικότητα κάθε χαρακτηριστικού. Τα χαρακτηριστικά που έχουν τη μεγαλύτερη διαφορά μεταξύ των κατηγοριών στη στήλη στόχο βαθμολογούνται υψηλότερα.

Διαδικασία: Ο χρήστης μπορεί να επιλέξει πόσα χαρακτηριστικά θέλει να διατηρήσει (π.χ. τα 5 καλύτερα χαρακτηριστικά). Το πρόγραμμα αναλύει τη σημασία κάθε χαρακτηριστικού χρησιμοποιώντας το SelectKBest και εμφανίζει το νέο δατασετ με τα επιλεγμένα χαρακτηριστικά.

Αλγόριθμοι Κατηγοριοποίησης

Η εφαρμογή προσφέρει δύο δημοφιλείς αλγόριθμους κατηγοριοποίησης:

K-Nearest Neighbors (KNN) KNN είναι ένας απλός και κατανοητός αλγόριθμος κατηγοριοποίησης που βασίζεται στην απόσταση μεταξύ των δεδομένων. Κατά την πρόβλεψη μιας νέας τιμής, εξετάζει τους ¨k' κοντινότερους γείτονες και καθορίζει την κατηγορία ανάλογα με την πλειοψηφία των γειτόνων.

Πλεονεκτήματα: Είναι εύκολος στην κατανόηση και δεν απαιτεί εκπαίδευση πριν την κατηγοριοποίηση.
Μειονεκτήματα: Δεν λειτουργεί καλά με μεγάλα datasets και είναι ευαίσθητος στα outliers.
Random Forest Random Forest είναι ένας πιο σύνθετος αλγόριθμος, που αποτελείται από πολλαπλά δέντρα απόφασης. Κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο των δεδομένων, και οι προβλέψεις συνδυάζονται για να δώσουν την τελική πρόβλεψη.

Πλεονεκτήματα: Είναι πολύ αποτελεσματικός σε πολλά προβλήματα, αποφεύγει το πρόβλημα της υπερεκπαίδευσης και λειτουργεί καλά με μεγάλα datasets.
Μειονεκτήματα: Μπορεί να είναι πιο αργός κατά την εκπαίδευση και την πρόβλεψη σε σύγκριση με απλούς αλγόριθμους όπως ο KNN.

Αποτελέσματα

Μετρικές Απόδοσης

Μετά την εκπαίδευση των αλγορίθμων, η εφαρμογή παρουσιάζει τα αποτελέσματα μέσω διαφόρων μετρικών απόδοσης:

Accuracy: Το ποσοστό των σωστών προβλέψεων επί του συνόλου των προβλέψεων.

F1-Score: Μια μετρική που συνδυάζει την ακρίβεια και την ανάκληση, ιδανική για προβλήματα με ανισομερείς κατηγορίες.

ROC-AUC: Η καμπύλη ROC (Receiver Operating Characteristic) δείχνει την ικανότητα του μοντέλου να διαχωρίζει τις κατηγορίες, και το AUC είναι η συνολική περιοχή κάτω από την καμπύλη, με τιμές που κυμαίνονται από 0.5 (τυχαιά επιλογή) έως 1 (τέλεια απόδοση).

Συγκρίσεις Μετά την Επιλογή Χαρακτηριστικών

Ένα από τα κύρια πλεονεκτήματα αυτής της εφαρμογής είναι ότι ο χρήστης μπορεί να συγκρίνει την απόδοση των αλγορίθμων πριν και μετά την επιλογή χαρακτηριστικών. Αυτό επιτρέπει την αξιολόγηση του κατά πόσο η μείωση των χαρακτηριστικών βελτιώνει ή μειώνει την ακρίβεια του μοντέλου.

Παράδειγμα Πίνακα Σύγκρισης

Αλγόριθμος	**Ακρίβεια**	**Ανάκληση**	**Φ1-Σcορε**
Δέντρα Αποφάσεων	0.85	0.80	0.82
Κ-Κοντινότεροι Γείτονες	0.82	0.83	0.82

Ταβλε 1. Παράδειγμα Συγκριτικής Απόδοσης Αλγορίθμων

User Interaction

Η εφαρμογή επιτρέπει στον χρήστη να επέμβει σε διάφορα σημεία:
Να ανεβάσει το dataset .
Να επιλέξει τη στήλη στόχο και τον αριθμό χαρακτηριστικών για επιλογή.
Να διαλέξει ποιον αλγόριθμο κατηγοριοποίησης θα χρησιμοποιήσει.
Να δει τα αποτελέσματα μέσω των οπτικοποιήσεων και των μετρικών απόδοσης.

Βελτιστοποίηση Μοντέλου

Εφόσον το μοντέλο έχει αξιολογηθεί, μπορεί να γίνει βελτιστοποίηση για καλύτερα αποτελέσματα:
Ρύθμιση Υπερπαραμέτρων: Ρύθμιση παραμέτρων του αλγορίθμου (π.χ. αριθμός γειτόνων στο KNN ή αριθμός δέντρων στο Random Forest).
Cross-validation: Εφαρμογή μεθόδων όπως το K-fold cross-validation για την καλύτερη εκτίμηση της γενικής ικανότητας του μοντέλου.

Τεχνολογίες Υλοποίησης

Για την υλοποίηση της web εφαρμογής, χρησιμοποιήθηκαν οι εξής τεχνολογίες:
Python: Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε για την ανάπτυξη του backend της εφαρμογής. Προσφέρει εργαλεία και βιβλιοθήκες για την επεξεργασία και ανάλυση δεδομένων, καθώς και για την εφαρμογή αλγορίθμων machine learning.
Streamlit: Το κύριο frameworkγια την ανάπτυξη του frontend. Streamlit είναι ένα εύκολο στη χρήση εργαλείο που επιτρέπει τη δημιουργία διαδραστικών web εφαρμογών απευθείας από Python scripts.Υποστηρίζει τη δυναμική εμφάνιση δεδομένων, γραφημάτων, και την αλληλεπίδραση με τον χρήστη μέσω αρχείων, πινάκων και επιλογών.
Pandas: Βιβλιοθήκη για την επεξεργασία και διαχείριση των δεδομένων. Χρησιμοποιήθηκε για την ανάγνωση και το χειρισμό των αρχείων CSV, Excel . TSV, καθώς και για την ανάλυση των δεδομένων σε μορφή DataFrame.
Scikit-learn: Βασική βιβλιοθήκη για την εφαρμογή αλγορίθμων μαςηινε λεαρνινγ. Χρησιμοποιήθηκε για την υλοποίηση της επιλογής χαρακτηριστικών (SelectKBest), καθώς και των αλγορίθμων ταξινόμησης όπως K-Nearest Neighbors (KNN) ,Random Forest.
. **Matplotlib , Seaborn:** Βιβλιοθήκες για την οπτικοποίηση δεδομένων. Χρησιμοποιήθηκαν για την δημιουργία γραφημάτων όπως boxplots, heatmaps scatter plots, παρέχοντας εργαλεία για τη διερεύνηση των δεδομένων.
Plotly: Βιβλιοθήκη που επιτρέπει τη δημιουργία διαδραστικών γραφημάτων. Χρησιμοποιήθηκε για την απεικόνιση των αποτελεσμάτων των αναλύσεων, συμπεριλαμβανομένων των διαγραμμάτων PCA UMAP.
NumPy: Χρησιμοποιήθηκε για την εκτέλεση αριθμητικών υπολογισμών σε πίνακες και πολυδιάστατα δεδομένα.
Docker: Χρησιμοποιήθηκε για την ανάπτυξη και διανομή της εφαρμογής, επιτρέποντας την εύκολη δημιουργία κοντέινερ που περιλαμβάνουν όλα τα απαραίτητα στοιχεία για την εκτέλεση της εφαρμογής σε διαφορετικά περιβάλλοντα.
GitHub: Η πλατφόρμα αυτή χρησιμοποιήθηκε για την αποθήκευση και διαχείριση του κώδικα της εφαρμογής.

Η επιλογή αυτών των τεχνολογιών εξασφάλισε μια ολοκληρωμένη και αποδοτική υλοποίηση, επιτρέποντας την ανάπτυξη, οπτικοποίηση και εύκολη διανομή της εφαρμογής machine learning σε ευρύτερο κοινό.

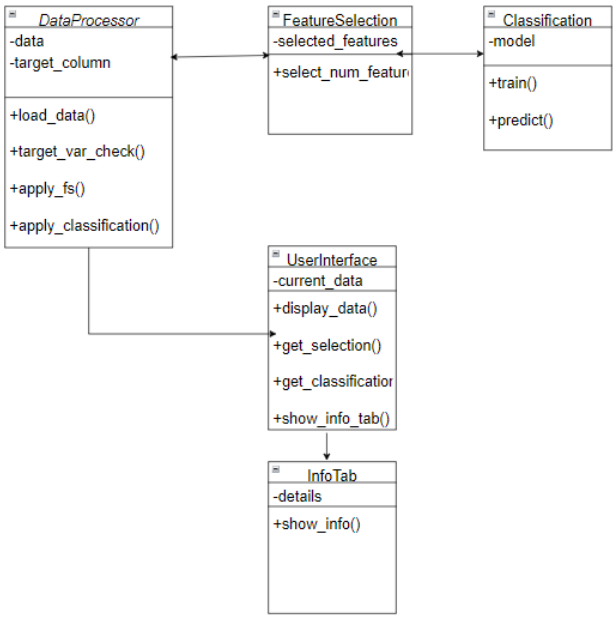
Ομάδα Υλοποίησης

Η εφαρμογή σχεδιάστηκε και υλοποιήθηκε από τον **Μπάρλα Ιωάννη (Π2019009).**Ο ίδιος ήταν υπεύθηνος τόσο για τον σχεδιασμό του front-end όσο και του back-end της εφαρμογής.

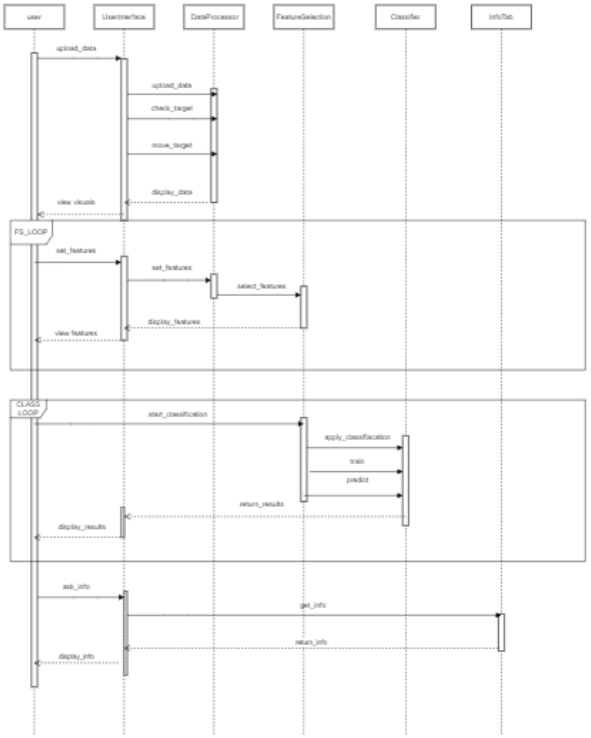
Github Link

Παρακάτω αναγράφεται ο σύνδεσμος github όπου βρίσκεται ο κώδικας της εφαρμογής και οδηγίες για την διάθεση λογισμικού μέσω docker image.
GITHUB LINK.

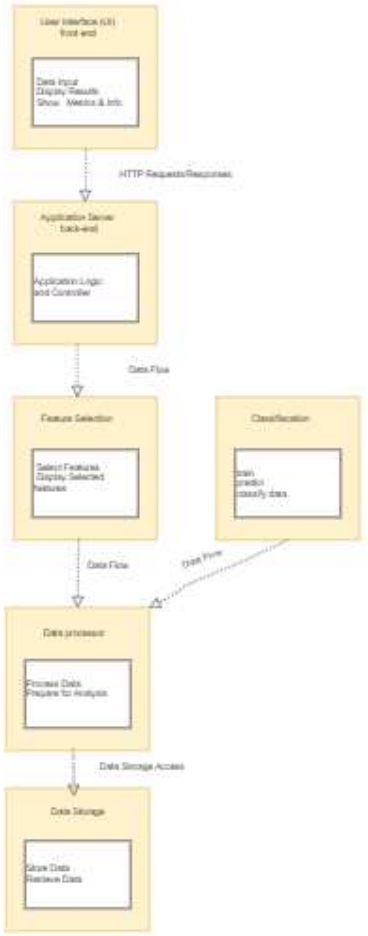
UML Diagrams



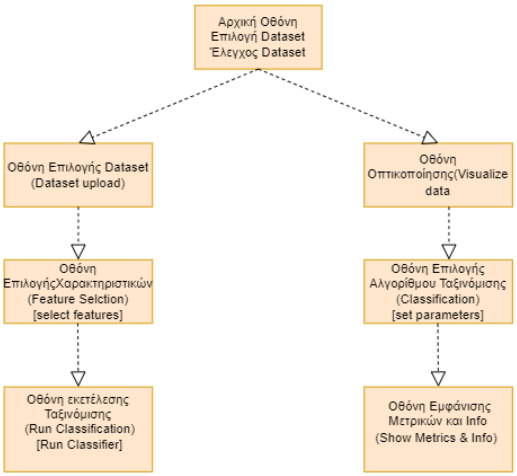
Σχήμα 1. Διάγραμμα κλάσης



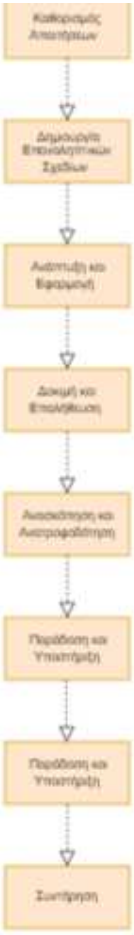
Σχήμα 2. Διάγραμμα ακολουθίας



Σχήμα 3. Διάγραμμα αρχιτεκτονικής της εφαρμογής



Σχήμα 4. Διάγραμμα διεπαφής χρήστη



Σχήμα 5. Διάγραμμα κύκλου λογισμικού