Τεχνολόγία Λογισμικού

Σούρλας Ιωάννης

February 7, 2025

Abstract

Στην εργασία αυτή παρουσιάζεται μια web bassed εφαρμογή η οποία αναπτύχθηκε για την οπτικοποίηση και αξιολόηση αλγορίθμων machine learning (ML). Στην αρχική σελίδα γίνεται φόρτωση και αναγνώριση των δεδομένων. Στις επόμενες σελίδες (Visualisation in 2D & 3D , Machine Learning , Classification, How it works) μπορούν να επιλεχθούν χαρακτηριστικά βάση των οποίων εκπαιδεύεται το μοντέλο και διαχωρίζονται σε training και validation set. Ο χρήστης μπορεί να αξιολογήσει την απόδοση των μοντέλων μέσω των μετρικών που εμφανίζονται και να διαλέξει τις καταλληλότερες παραμέτρους για βέλτιστα αποτελέσματα

1 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

2 Περιγραφή της εφαρμογής.

Η εφαρμογή έχει φτιαχτεί έχοντας στο μυαλό την ευχρηστία και την εύκολη πρόσβαση. Αναπτύχθηκε χρησιμοποιώντας ένα συνδυασμό τεχνολογιών web, όπως html, css και javascript, 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 Python , Sklearn η οποία παρέχει ένα ευρύ φάσμα εργαλείων για ανάλυση δεδομένων και μοντελοποίης

3 Πλοήγηση

Στην αρχική οθόνη (Home Page) παρουσιάζονται κάποιες πληροφορίες σχετικά με την εργασία και έπειτα πιο κάτω υπάρχει το πεδίο για το ανέβασμα του αρχείου CSV ή Excel που θα χρησιμοποιήσουμε

3.1 Visualisation in 2D & 3D

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.

3.2 Machine Learning

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

3.3 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.

3.4 How it works

Μια σύντομη επεξήγηση για τη λειτουργία της εφαρμογής

4 Αναλυτικότερα

4.1 Data upload

Ο χρήστης έχει την δυνατότητα να φορτώσει στην εφαρμογή δεδομένα τύπου CSV , Excel ή TSV. Για την ανάγνωση των δεδομένων χρησιμοποιείται η βιβλιοθήκη pandas.

4.2 Visualisation

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.

4.3 Machine learning

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. 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 ¨κ' 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).

5 Γλώσσα προγραμματισμού και libraries που χρησιμοποιήθηκαν

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.

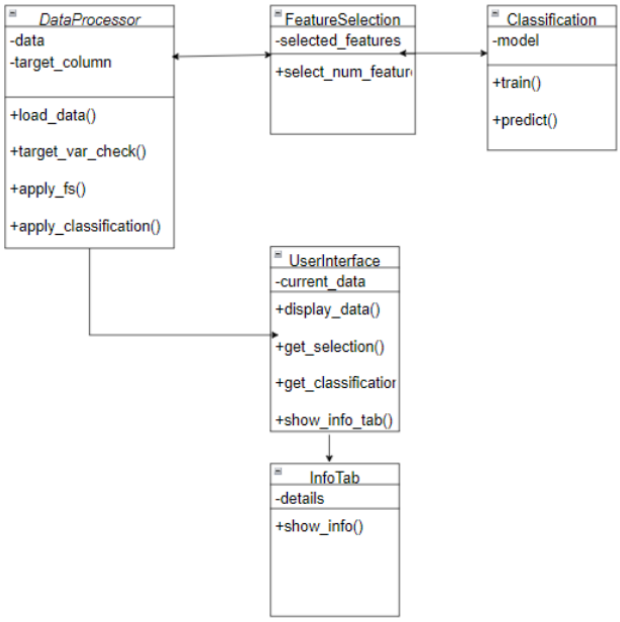
Docker: Χρησιμοποιήθηκε για την ανάπτυξη και διανομή της εφαρμογής, επιτρέποντας την εύκολη δημιουργία κοντέινερ που περιλαμβάνουν όλα τα απαραίτητα στοιχεία για την εκτέλεση της εφαρμογής σε διαφορετικά περιβάλλοντα.

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.

6 Υλοποίηση της εφαρμογής

Η εφαρμογή αυτή δημιουργήθηκε απο τον Σούρλα Ιωάννη Π2016102.

7 UML Diagrams



8 Ευχαριστώ

Ελπίζω η εφαμοργή να σας φάνηκε χρήσιμη και ενδιαφέρουσα Αν έχετε παρατηρήσεις ή προτάσεις βελτίωσης μπορείτε να με βρείτε στο **p16sour@ionio.gr**


ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ 1984