

ΕΙΣΑΓΩΓΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ - ΒΑΣΙΚΕΣ ΕΦΑΡΜΟΓΕΣ

Χρήστος Ανδριανός

Σπύρος Βυθούλκας

Ζένια Φραγκάκη

Τμήμα Μηχανικών Βιοϊατρικής,
Πανεπιστήμιο Δυτικής Αττικής

Εισαγωγή – Μηχανική Μάθηση

Ανάπτυξη αλγορίθμων που επιτρέπουν στα συστήματα να μαθαίνουν από τα δεδομένα και να βελτιώνουν την απόδοσή τους

Στόχοι Μηχανικής Μάθησης:

- ✓ **Ταξινόμηση (Classification):**

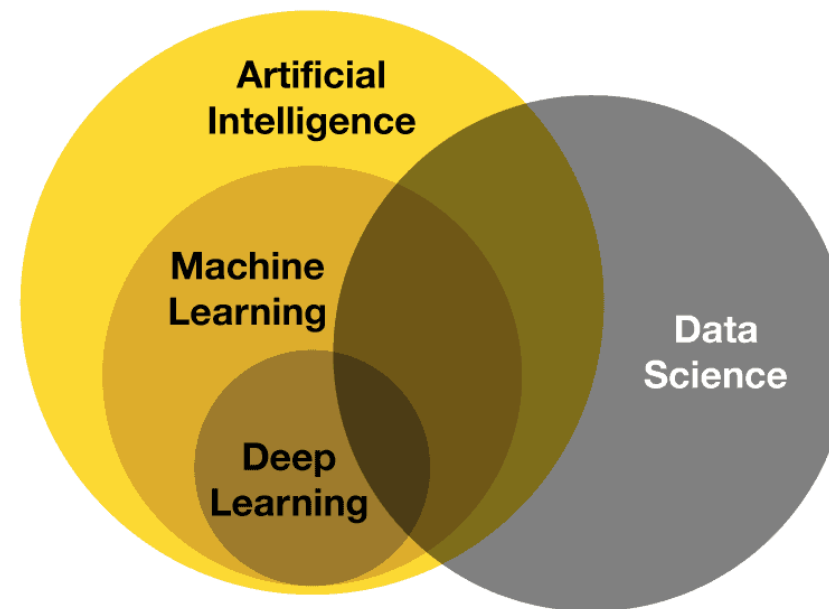
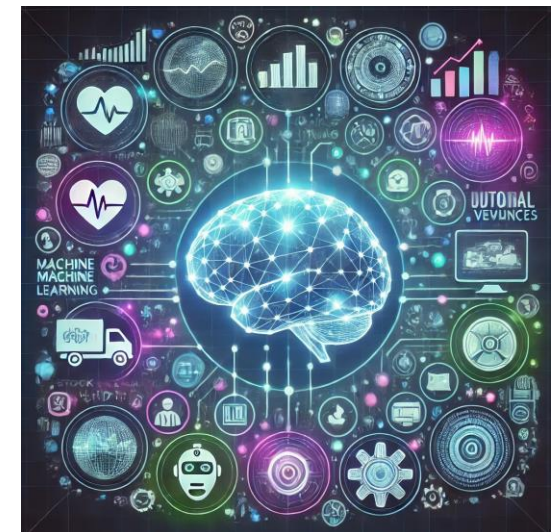
Κατηγοριοποίηση άγνωστων δεδομένων σε κλάσεις

- ✓ **Ομαδοποίηση (Clustering):**

Ομαδοποίηση με βάση κοινές ιδιότητες

- ✓ **Πρόβλεψη (Prediction)**

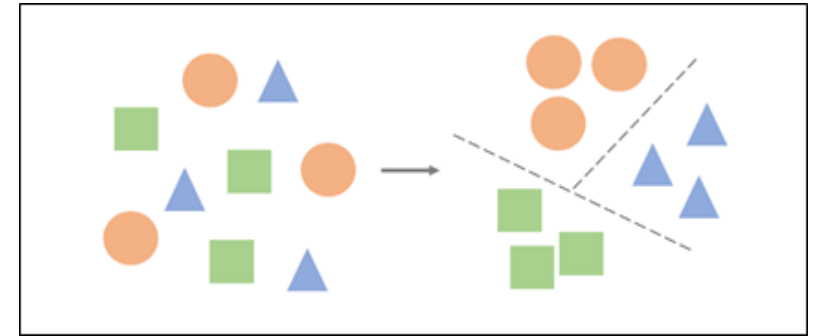
Εκτίμηση μελλοντικών τιμών αγνώστων μεταβλητών



Τύποι Μηχανικής Μάθησης

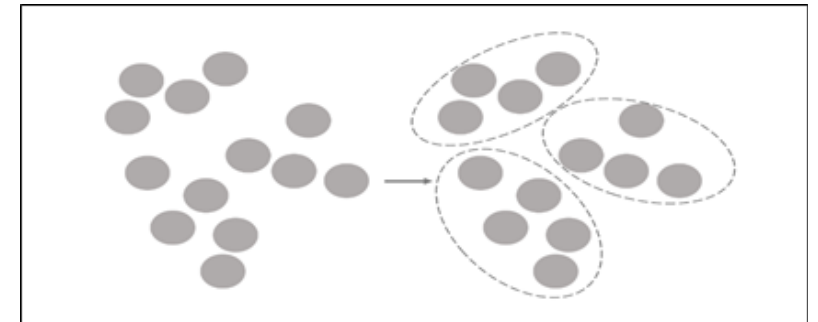
Εποπτευόμενη Μηχανική Μάθηση (Supervised ML)

Επισημασμένα Δεδομένα (Labeled)



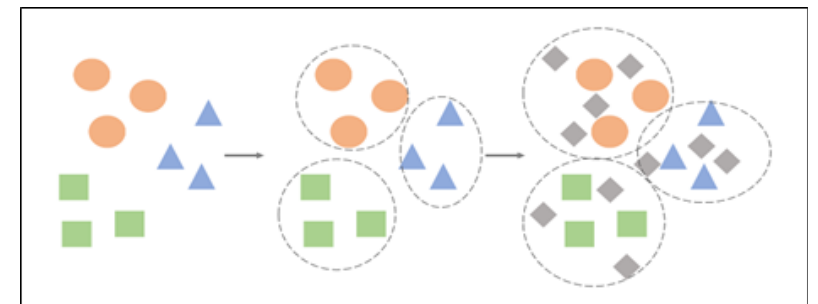
Μη Εποπτευόμενη Μηχανική Μάθηση (Unsupervised ML)

Μη επισημασμένα (Unlabeled)



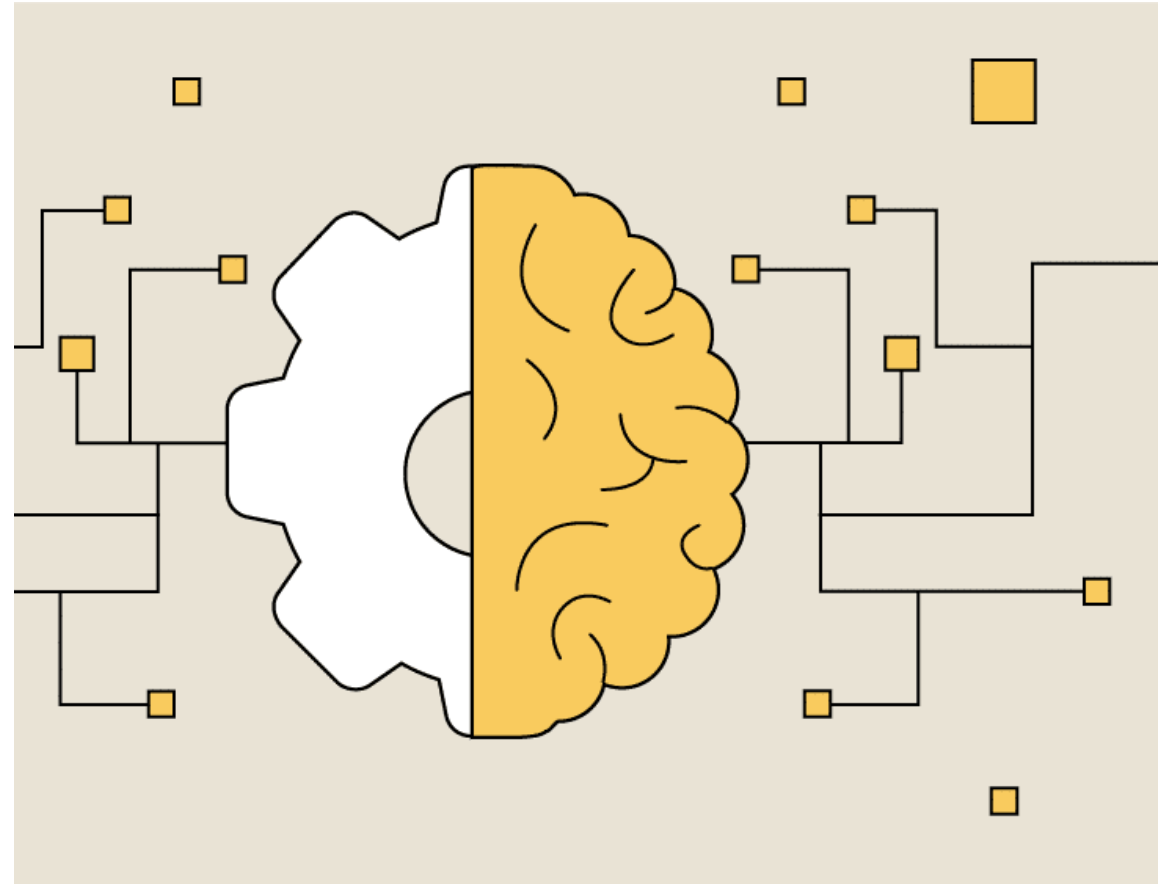
Ήμι-Εποπτευόμενη Μηχανική Μάθηση (Semi Supervised ML)

Επισημασμένα & Μη επισημασμένα (Labeled & Unlabeled)

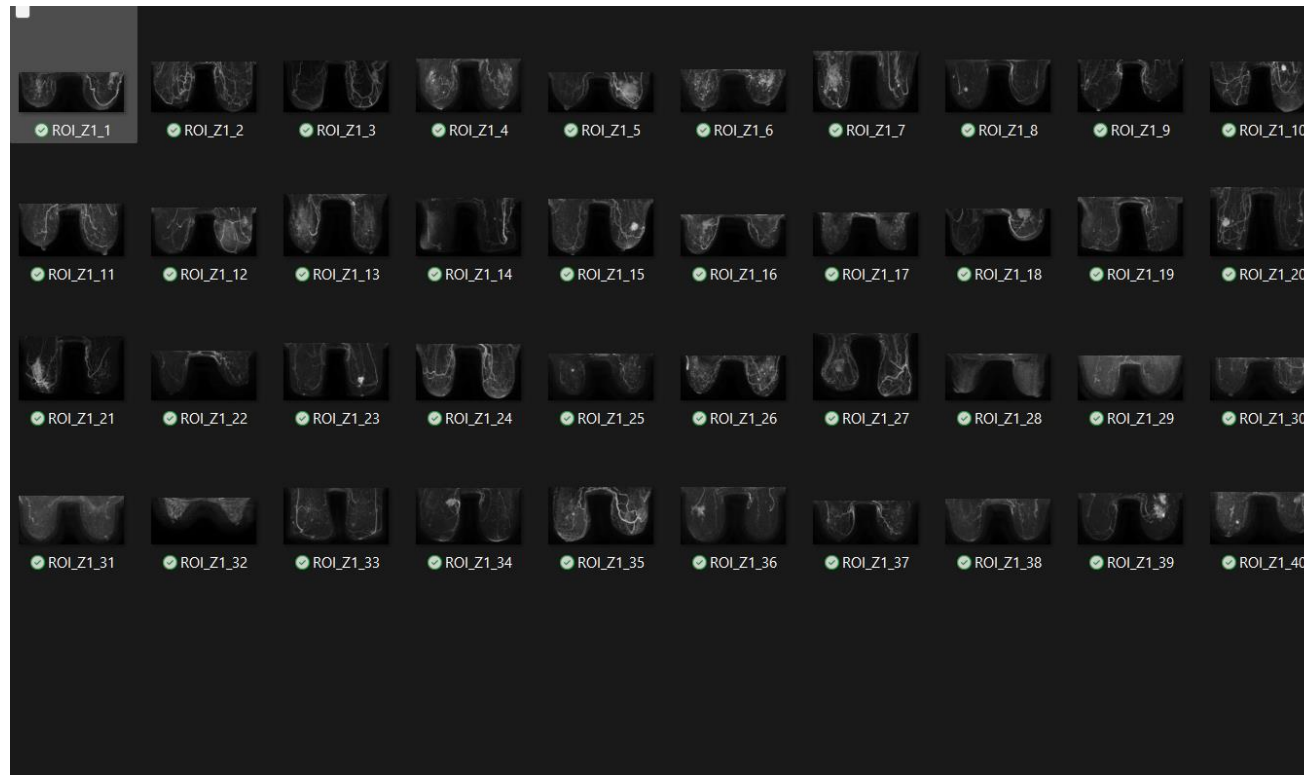


Βασική Μεθοδολογία Ανάπτυξης Μοντέλων ML

- I. Συλλογή Δεδομένων
- II. Προ-επεξεργασία Δεδομένων
- III. Εξαγωγή Χαρακτηριστικών
- IV. Επιλογή Αλγορίθμου Ταξινόμησης
- V. Εκπαίδευση - Αξιολόγηση Μοντέλου
- VI. Επικύρωση Αποτελεσμάτων Σε Άγνωστα Δεδομένα



I. Συλλογή Δεδομένων



Τύποι Δεδομένων

Σήματα (1D)

Εικόνες (2D/3D, Gray Scale/RGB)

Κλινικά Δεδομένα

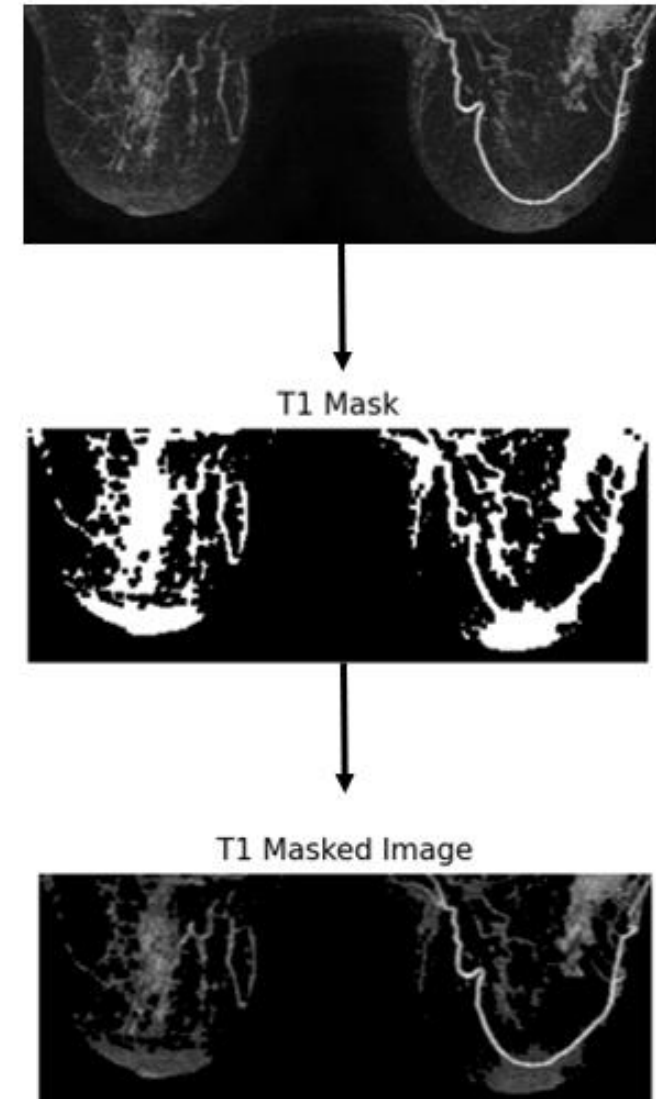
Γενετικά Δεδομένα (DNA/RNA)

Προέλευση βάσης δεδομένων:
MEDISP Laboratory, University of West Attica

II. Προ-Επεξεργασία Δεδομένων

- I. Δημιουργία συνθετικών δεδομένων από τα υπάρχοντα (data augmentation) μέσω περιστροφής, μεγέθυνσης και αλλαγής συντεταγμένων
- II. Τμηματοποίηση (segmentation) για τον διαχωρισμό επιμέρους περιοχών ενδιαφέροντος πάνω στην εικόνα
- III. Dataset Cleaning, αφαιρώντας ανακριβή/ελλιπή δεδομένα και λανθασμένες καταγραφές

Στο παράδειγμα της διπλανής εικόνας η περιοχή ενδιαφέροντος διαχωρίζεται από την υπόλοιπη εικόνα μέσω τεχνικής Otsu, που αποτελεί αυτόματη τεχνική κατωφλίωσης.



III. Εξαγωγή χαρακτηριστικών


I. Βασικά χαρακτηριστικά Υφής 1^{ης} Τάξης:

- Μέση Τιμή (Mean)
- Τυπική Απόκλιση (Standard deviation)
- Λοξότητα (Skewness)
- Κυρτότητα (Kurtosis)

II. Κάποια από τα χαρακτηριστικά Υφής 2^{ης} τάξης:

- Αντίθεση (Contrast)
- Ανομοιογένεια (Dissimilarity)
- Ενέργεια (Energy)
- Ομοιογένεια (Homogeneity)

```
164 # Εξαγωγή χαρακτηριστικών από το train και test set
165 X_train_features = extract_features(X_train)
166 X_test_features = extract_features(X_test)
167
168 # Standardization των χαρακτηριστικών
169 scaler = StandardScaler()
170 X_train_features = scaler.fit_transform(X_train_features)
171 X_test_features = scaler.transform(X_test_features)
```

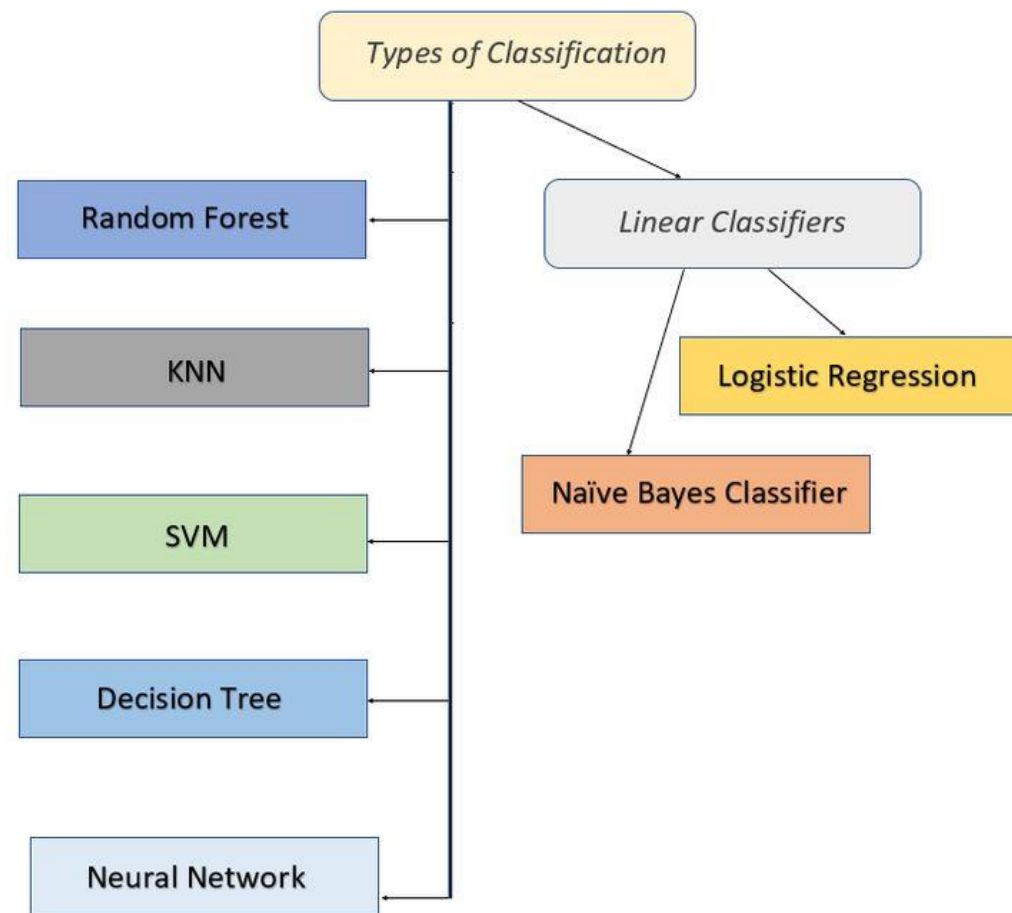


	mean	std	skewness	kurtosis	...	LBP_7	LBP_8	LBP_9	LBP_10
0	230.315552	75.402648	-2.727190	8.437568	...	985.0	484.0	376.0	868.0
1	177.522583	117.280990	-0.853064	1.727719	...	971.0	394.0	283.0	703.0
2	209.911194	97.289377	-1.694197	3.870305	...	845.0	422.0	287.0	661.0
3	224.728088	82.482508	-2.357616	6.558355	...	944.0	481.0	310.0	701.0
4	206.471558	100.101722	-1.577874	3.489687	...	1046.0	510.0	331.0	693.0
5	16.918030	63.467501	3.484786	13.143737	...	1091.0	597.0	386.0	801.0

IV. Επιλογή Αλγορίθμου Ταξινόμησης

Κάποιοι από τους πιο συνηθισμένους ταξινομητές σε μοντέλα μηχανικής μάθησης είναι:

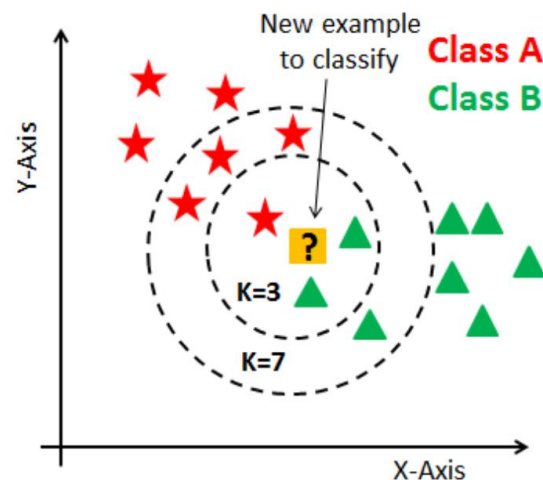
- K-Nearest Neighbors (KNN): Ταξινομεί ένα πρότυπο βάσει της **πλειοψηφίας** των **k-πλησιέστερων** γειτόνων του στον χώρο
- Logistic Regression: Προβλέπει την πιθανότητα να ανήκει το πρότυπο σε μία κατηγορία μέσω της **λογιστικής συνάρτησης (sigmoid)**
- Random Forest: Αποτελείται από **πολλαπλά δέντρα απόφασης** και η ταξινόμηση γίνεται με **ψηφοφορία**
- Support Vector Machine (SVM): Βρίσκει ένα **υπερ-επίπεδο (hyperplane)** που διαχωρίζει τις κατηγορίες με **μέγιστο περιθώριο**



Λίγα Μαθηματικά...

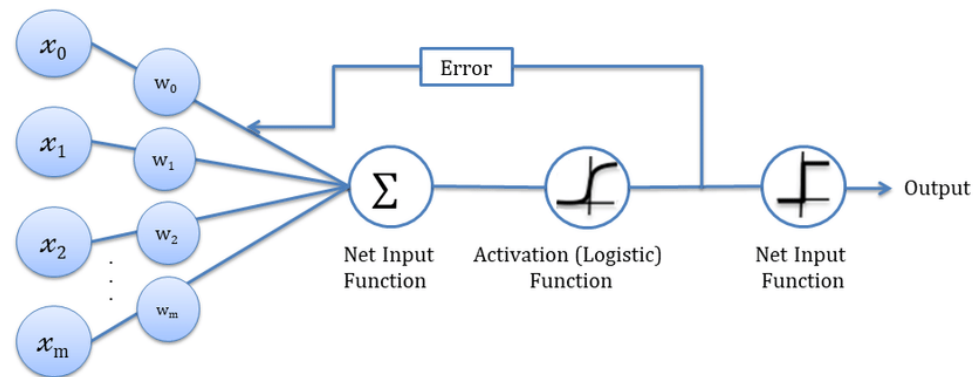
1. K-Nearest Neighbors (KNN) με Ευκλείδεια απόσταση:

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2}$$



2. Logistic Regression:

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$



Λίγα Μαθηματικά...

3. Radom Forrest:

$$Gini\ Impurity = 1 - \sum_{i=1}^K p_i^2$$

$$= 1 - Gini\ Index$$

where K is the number of class labels,

p_i is the proportion of i^{th} class label

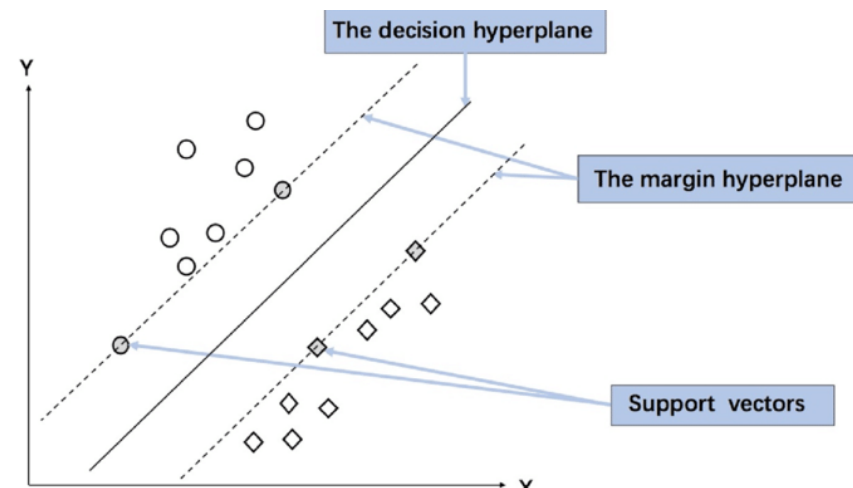
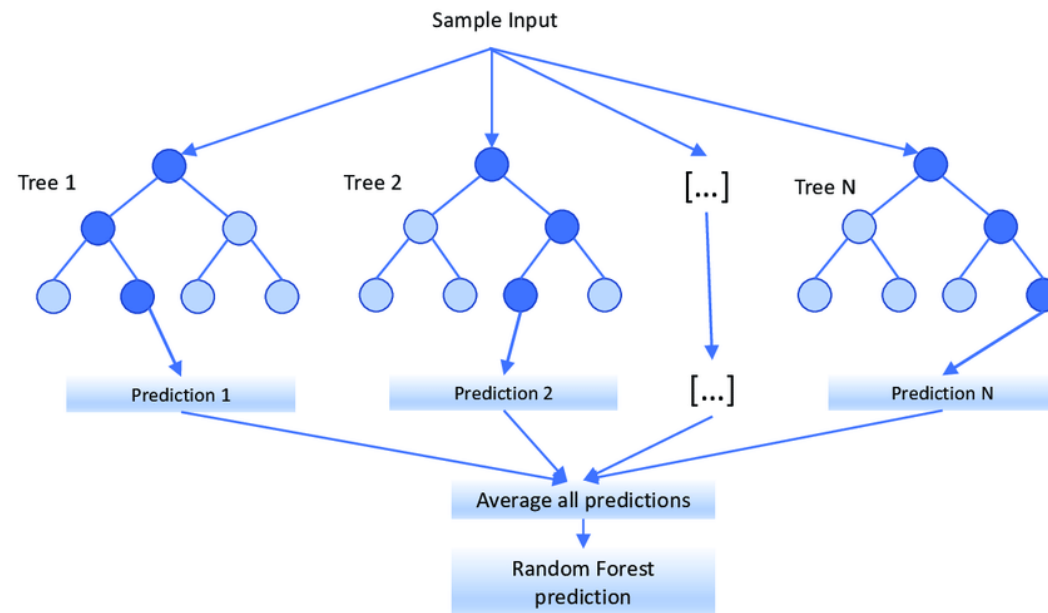
4. Support Vector Machine (SVM):

Distance from a Data Point to the Hyperplane:

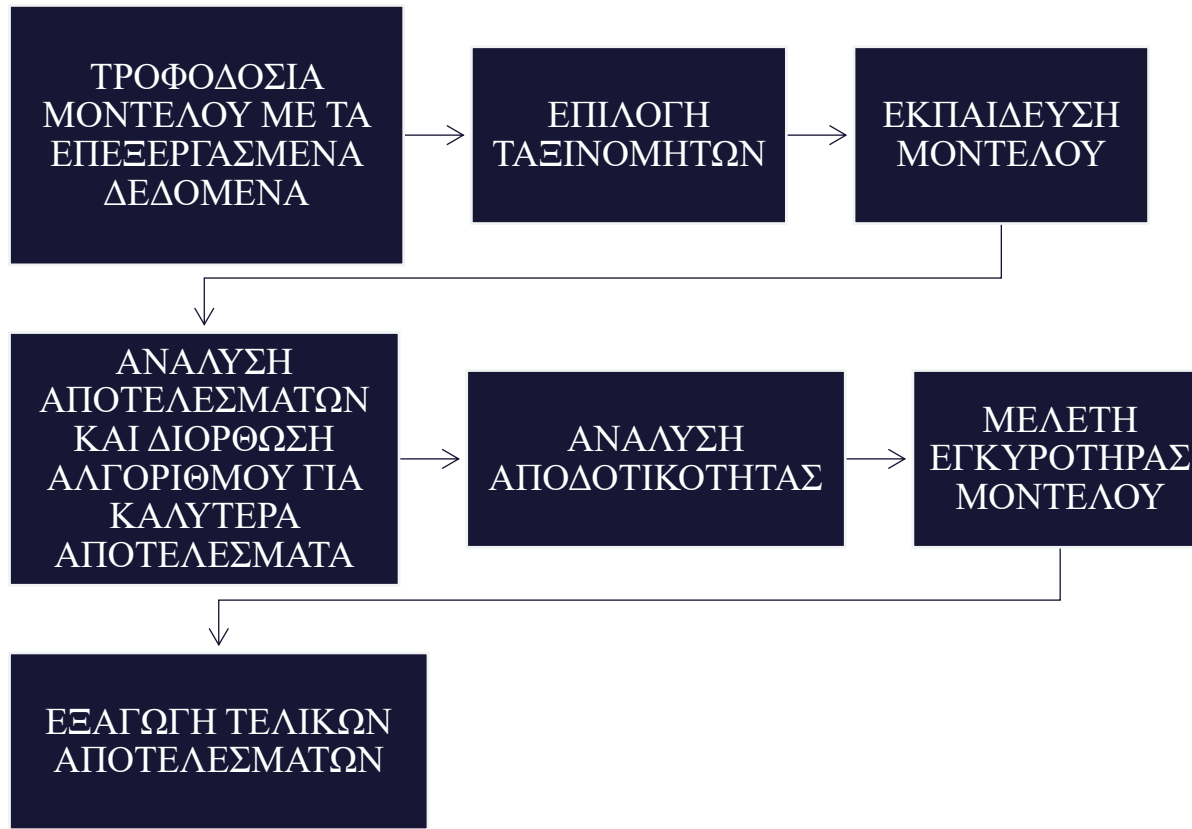
$$\hat{y} = \begin{cases} 1 & : w^T x + b \geq 0 \\ 0 & : w^T x + b < 0 \end{cases}$$

Where \hat{y} is the predicted label of a data point.

$$\underset{w,b}{\text{minimize}} \frac{1}{2} \|w\|^2$$



ΣΧΕΔΙΑΓΡΑΜΜΑ ΚΑΙ ΒΑΣΗ ΚΩΔΙΚΑ

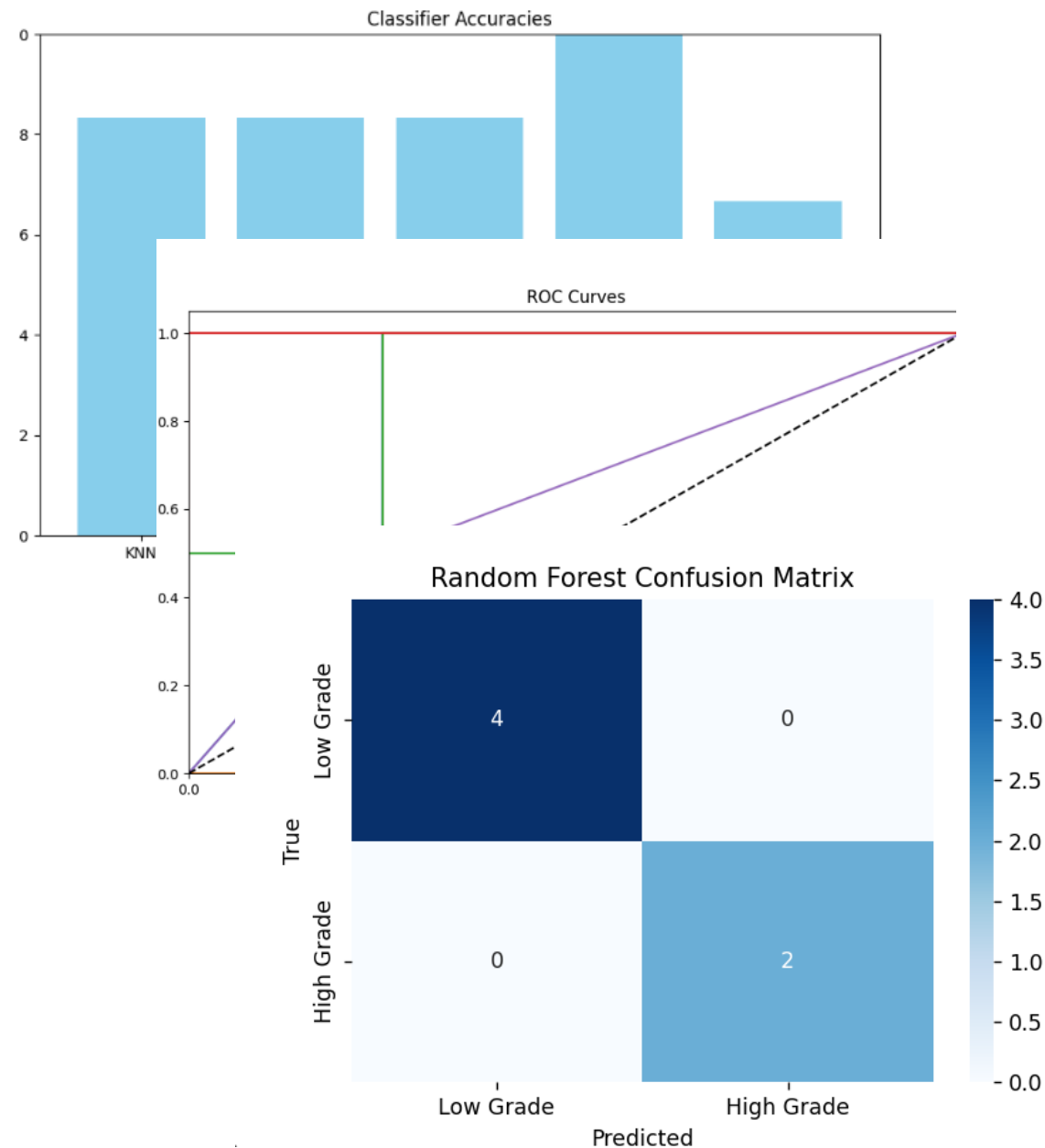


3.5.2 Code structure

```
1. # Loading Images and Split to train&test
2. X, y = load_images(DATA_PATH_LOW,
DATA_PATH_HIGH, DIMENSION)
5. X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.3,
random_state=SEED)
8. X_train_features=extract_features(X_train)
9. X_test_features = extract_features(X_test)
11. # Standardization of features
12. scaler = StandardScaler()
13. X_train_features =
scaler.fit_transform(X_train_features)
14. X_test_features =
scaler.transform(X_test_features)
16. # --- Train and Test --- #
17. classifiers = {
18.     "KNN": (KNeighborsClassifier(),
{'n_neighbors': [3, 5, 7]}),
19.     "SVM": (SVC(probability=True), {'C':
[0.1, 1, 10], 'kernel': ['linear', 'rbf']}),
20.     "Logistic Regression":
(LogisticRegression(), {'C': [0.1, 1, 10]}),
21.     "Random Forest":
(RandomForestClassifier(), {'n_estimators':
[50, 100, 200]}),
22.     "XGBoost": (XGBClassifier(),
{'n_estimators': [50, 100, 200],
'learning_rate': [0.01, 0.1, 0.2]})
23. }
25. best_accuracy = 0
26. best_classifier_name = None
27. best_classifier = None
31. cv_results = {}
```

ΜΕΤΡΗΣΗ ΑΠΟΔΟΤΙΚΟΤΗΤΑΣ

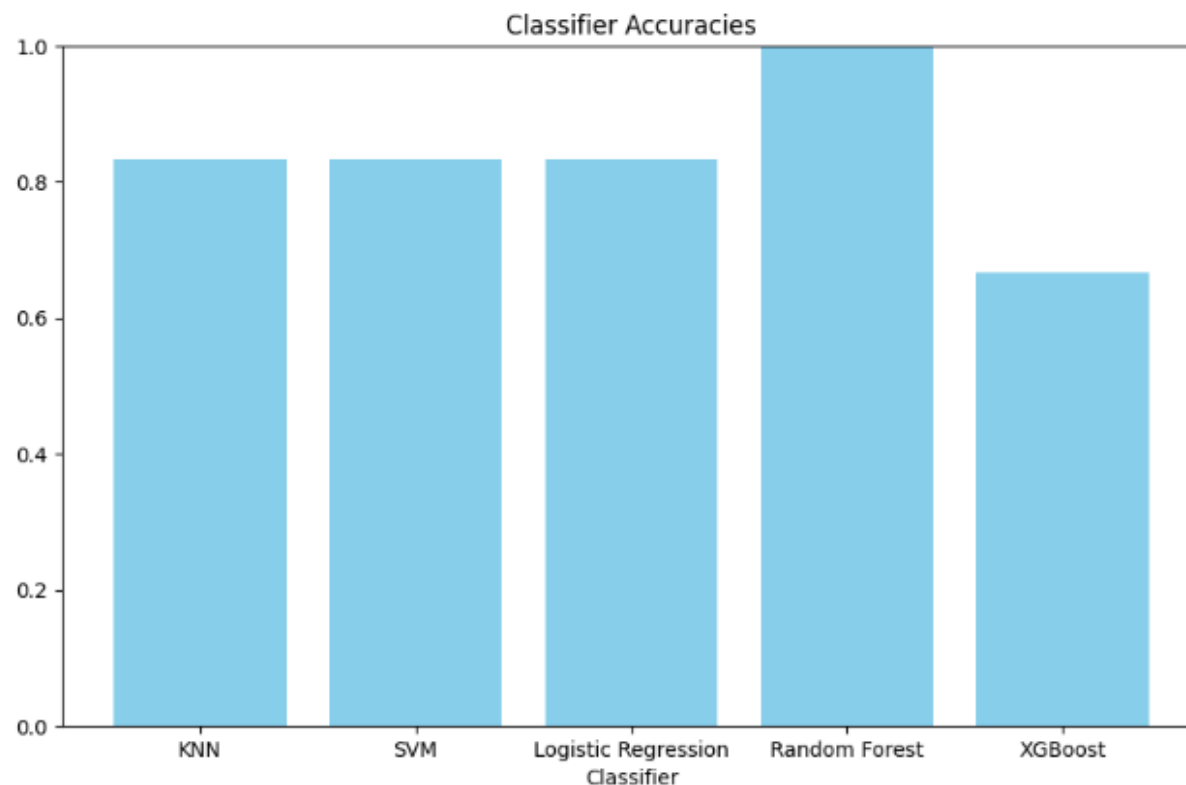
1. **Ακρίβεια:** Ποσοστό επιτυχών προβλέψεων έναντι του συνόλου των προβλέψεων
2. **ROC Καμπύλη:** Είναι μια καμπύλη που δείχνει την απόδοση ενός ταξινομητή σε διάφορα κατώφλια απόφασης
3. **AUC:** Είναι το **εμβαδόν κάτω από την καμπύλη ROC** και μετράει πόσο καλά το μοντέλο διαχωρίζει τις δύο κατηγορίες
 - **AUC = 1** → Τέλεια ταξινόμηση
 - **AUC = 0.5** → Τυχαία ταξινόμηση (50-50)
 - **AUC < 0.5** → Χειρότερα από την τύχη (πιθανό λάθος)
4. **Πίνακας Αλήθειας:** Παραθέτουν πληροφορίες για την επιτυχία πρόβλεψης σε συγκεκριμένες καταστάσεις.



ΜΕΤΡΗΣΗ ΑΠΟΔΟΤΙΚΟΤΗΤΑΣ - ΑΚΡΙΒΕΙΑ

Στο διπλανό ιστόγραμμα παρουσιάζεται η αξιολόγηση των ποσοστών ακρίβειας 5 ταξινομητών.

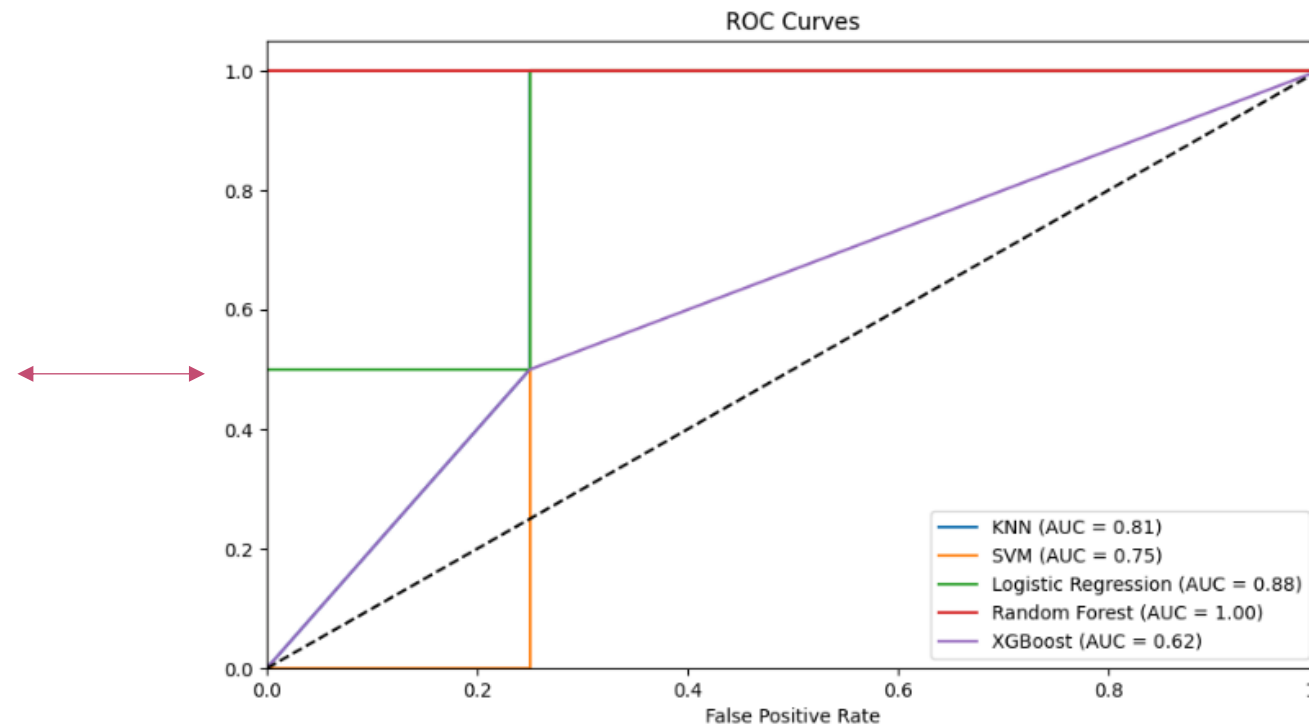
- KNN → 83.3%
- SVM → 83.3%
- Logistic Regression → 83.3%
- XGBoost → 66.7%
- Random Forest → 100%



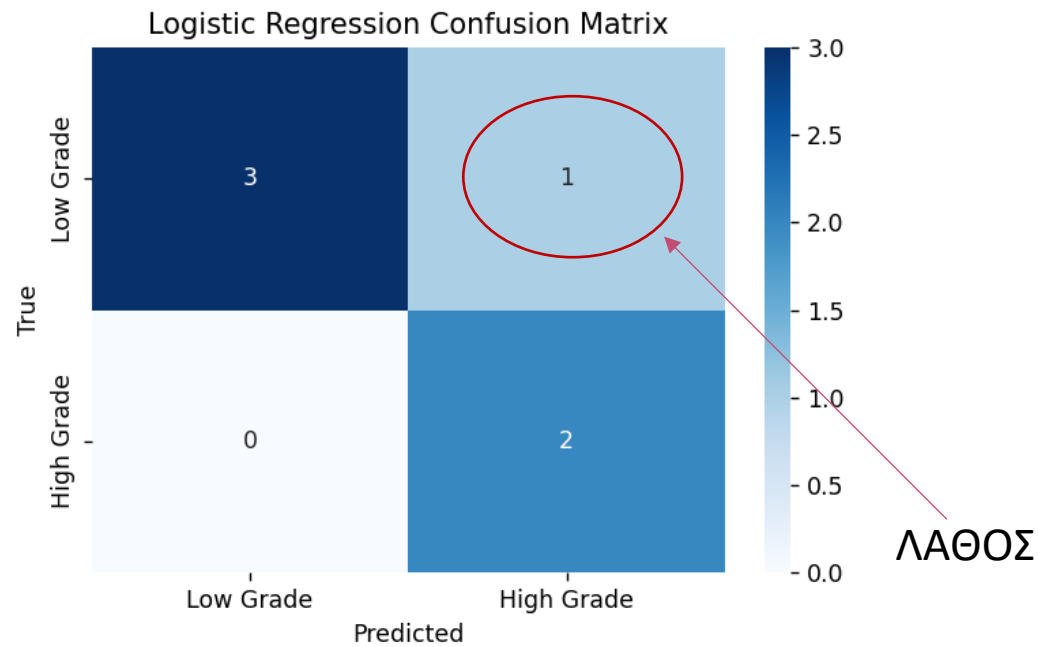
ΜΕΤΡΗΣΗ ΑΠΟΔΟΤΙΚΟΤΗΤΑΣ - ROC CURVE

ΠΙΝΑΚΑΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

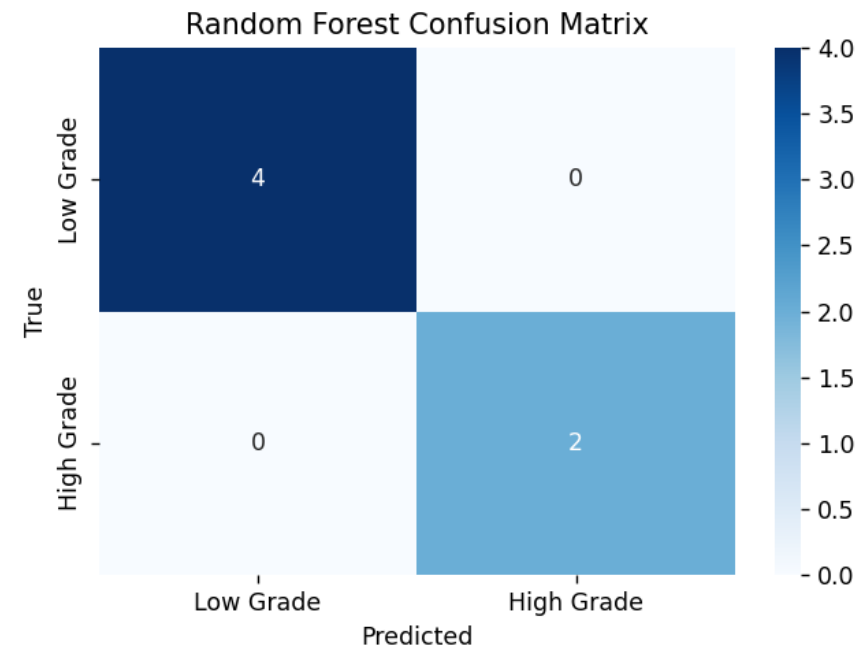
Model	AUC
KNN	81%
Logistic Regression	88%
SVM	75%
Random Forest	100%
XGBoost	62%



ΜΕΤΡΗΣΗ ΑΠΟΔΟΤΙΚΟΤΗΤΑΣ – ΠΙΝΑΚΑΣ ΑΛΗΘΕΙΑΣ



83.3% ΕΠΙΤΥΧΙΑ ΠΡΟΒΛΕΨΕΩΝ

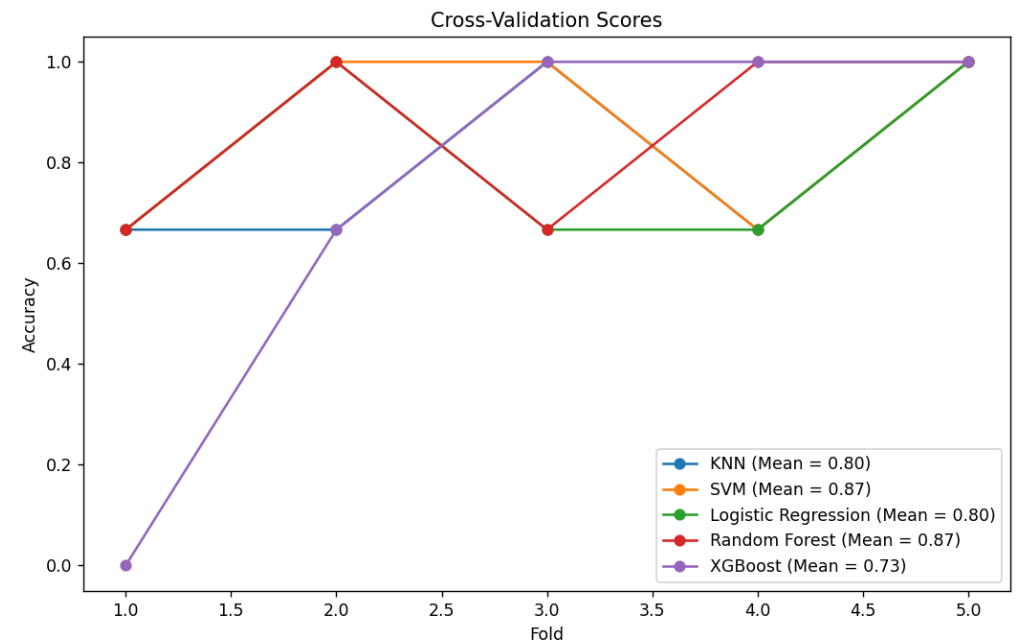


100% ΕΠΙΤΥΧΙΑ ΠΡΟΒΛΕΨΕΩΝ

ΕΓΚΥΡΟΤΗΤΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

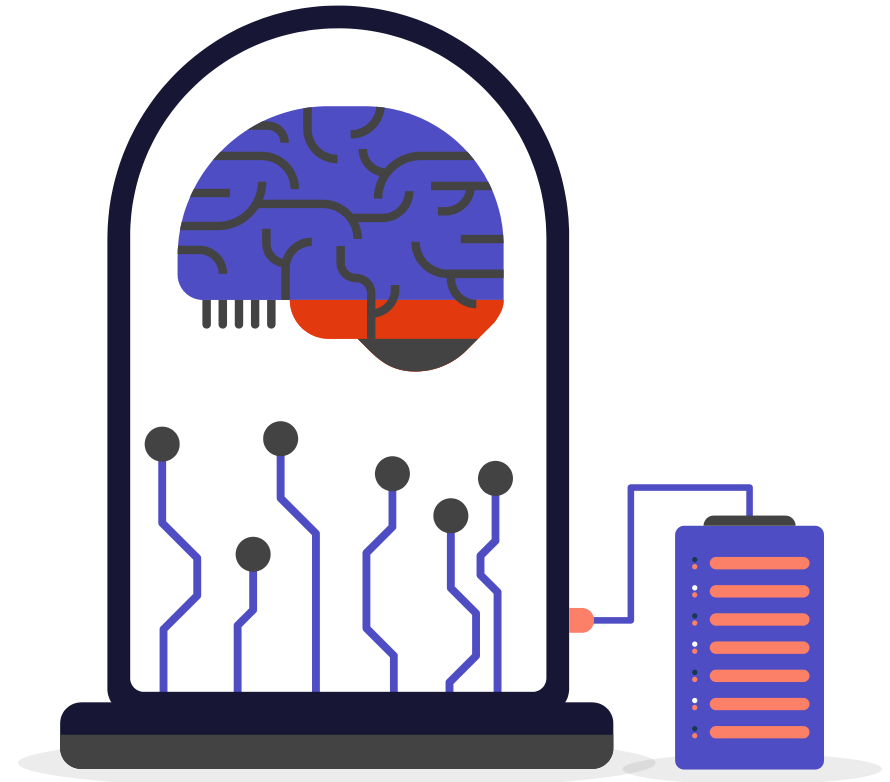
- K-Fold Cross Validation (GreadSearchCV)
- Χωρίζεται σε k αριθμό υποσυνόλων και εκπαιδεύεται σε όλα τα υποσύνολα εκτός από ένα ($k-1$) υποσύνολο που χρησιμοποιείται για την αξιολόγηση

Cross Validation Score
KNN= 80%
Logistic Regression = 80%
Random Forest= 87%
Random Forest= 87%
XGBoost= 73%



ΕΦΑΡΜΟΓΕΣ

- Ανακάλυψη νέων φαρμάκων
- Προσωποποιημένη θεραπεία
- Συστήματα υποβοήθησης στη διάγνωση
- Υποβοήθηση στην ακτινοθεραπεία



ΕΦΑΡΜΟΓΕΣ

