UNIVERSITY OF PIRAEUS
DEPARTMENT OF DIGITAL SYSTEMS – BIG DATA AND ANALYTICS

[ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ: ΜΕΘΟΔΟΙ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ]

ΜΕΤΑΠΤΥΧΙΑΚΗ ΑΠΑΛΛΑΚΤΙΚΗ ΕΡΓΑΣΙΑ ΦΕΒ. 2024-25

Καθηγητής: Ο. Τελέλης

ΜΕΝΥΧΤΑ ΑΙΚΑΤΕΡΙΝΗ

ΜΕ2421

Table of Contents

1. Εισαγωγή 3
   - Σκοπός της Εργασίας 3
   - Αναλυτικοί Στόχοι Εργασίας 4
2. Συστήματα και Μέθοδοι 5
   - Περιγραφή του Συνόλου Δεδομένων του eBay 5
   - Διαχωρισμός των επιμέρους προβλημάτων 7
3. Εφαρμογή Μεθόδων Πρόβλεψης Στο Σύνολο Δεδομένων Από Δημοπρασίες του eBay 9
   - Πειράματα για την Πρόβλεψη Τιμής (Price) 9
   - Πειράματα για την Πρόβλεψη Πωλήσεων (QuantitySold) 11
   - Πειράματα για την Πρόβλεψη της Σχέσης Τιμής και Μέσης Τιμής Πώλησης (Price vs AvgPrice) 13
4. Εφαρμογή Μοντέλων Κατηγοριοποίησης στο CIFAR-10 17
   - Περιγραφή του Συνόλου Δεδομένων CIFAR-10 17
   - Διαχωρισμός του Συνόλου Δεδομένων 19
5. Συμπεράσματα 22
   - Συμπεράσματα και Σύγκριση των Μοντέλων Πρόβλεψης eBay 22
   - Συμπεράσματα και Συγκριτική Ανάλυση Αποτελεσμάτων για το CIFAR-10 24
   - Προτάσεις για Μελλοντική Έρευνα 26
6. Βιβλιογραφία 27

1. Εισαγωγή

Σκοπός της Εργασίας

Η παρούσα εργασία έχει ως σκοπό την εφαρμογή μεθόδων μηχανικής μάθησης για την ανάλυση και την πρόβλεψη δημοπρασιών στην πλατφόρμα eBay και τη χρήση αλγορίθμων για την κατηγοριοποίηση εικόνων μέσω του συνόλου δεδομένων CIFAR-10. Στην πρώτη ενότητα της εργασίας, αναλύουμε και εκπαιδεύουμε μοντέλα πρόβλεψης χρησιμοποιώντας δεδομένα από τις δημοπρασίες του eBay, με στόχο την πρόβλεψη της τιμής πώλησης (Price), της πιθανότητας πώλησης (QuantitySold) και της σύγκρισης τιμής με τη μέση τιμή του αντικειμένου (Price vs AvgPrice). Στη δεύτερη ενότητα, επικεντρωνόμαστε στο σύνολο δεδομένων CIFAR-10, που χρησιμοποιείται για την κατηγοριοποίηση εικόνας.

Στο πρώτο μέρος της εργασίας, το σύνολο δεδομένων του eBay περιλαμβάνει ιστορικά δεδομένα δημοπρασιών, όπου κάθε καταγραφή αναφέρεται σε δημοπρασίες που είτε κατέληξαν σε πώληση είτε όχι. Το σύνολο δεδομένων αυτό περιλαμβάνει χαρακτηριστικά όπως η τελική τιμή της δημοπρασίας, ο αριθμός των αντικειμένων που έχουν πουληθεί, καθώς και άλλες πληροφορίες που αφορούν τον πωλητή και τα προϊόντα. Η ανάλυση επικεντρώνεται στη δημιουργία μοντέλων πρόβλεψης για τρεις βασικούς στόχους, χρησιμοποιώντας τη γλώσσα Python και τις βιβλιοθήκες NumPy, SciKit-learn και Matplotlib.

Ο στόχος του πρώτου μέρους είναι να αξιολογηθούν και να συγκριθούν διάφορες τεχνικές μηχανικής μάθησης που εφαρμόστηκαν και περιλαμβάνουν τα μοντέλα Random Forest, Linear Regression και Logistic Regression, ενώ παράλληλα θα αναλυθούν τα αποτελέσματα μέσω γραφημάτων και πινάκων.

Στο δεύτερο μέρος της εργασίας, η ανάλυση επικεντρώνεται στη χρήση του συνόλου δεδομένων CIFAR-10, το οποίο περιλαμβάνει 60.000 έγχρωμες εικόνες σε 10

κατηγορίες, με σκοπό την κατηγοριοποίησή τους. Το CIFAR-10 είναι ένα σύνολο δεδομένων, το οποίο χρησιμοποιείται για την εκπαίδευση και αξιολόγηση αλγορίθμων κατηγοριοποίησης εικόνας. Η εργασία αυτή στοχεύει στο να ταξινομήσει τις εικόνες σε προκαθορισμένες κατηγορίες, όπως για παράδειγμα ζώα, οχήματα, κ.λπ, οπότε το πρόβλημα που καλούμαστε να λύσουμε αποτελεί πρόβλημα κατηγοριοποίησης έγχρωμων εικόνων.

Αναλυτικοί Στόχοι Εργασίας

Πιο αναλυτικά, στο πρώτο μέρος της εργασίας καλούμαστε να λύσουμε τα κάτωθι προβλήματα:

Πρόβλεψη Τιμής Πώλησης (Price): Η πρόβλεψη της τελικής τιμής μιας δημοπρασίας για προϊόντα που καταλήγουν σε πώληση, χρησιμοποιώντας αλγορίθμους όπως το Random Forest και η Γραμμική Παλινδρόμηση (Linear Regression). Η πρόβλεψη αυτή αποτελεί πρόβλημα παλινδρόμησης, καθώς η τιμή είναι μια συνεχής μεταβλητή.

Πρόβλεψη Πιθανότητας Πώλησης (QuantitySold): Η ταξινόμηση των δημοπρασιών σε εκείνες που κατέληξαν σε πώληση (1) και εκείνες που δεν κατέληξαν σε πώληση (0). Αυτό είναι ένα πρόβλημα ταξινόμησης, που επιλύεται μέσω ταξινομητών όπως ο Random Forest Classifier και η Λογιστική Παλινδρόμηση (Logistic Regression).

Σχέση Τιμής και Μέσης Τιμής Πώλησης (Price vs AvgPrice): Πρόβλεψη αν η τιμή στην οποία έκλεισε η δημοπρασία είναι μεγαλύτερη ή μικρότερη από τη μέση τιμή του αντικειμένου, χρησιμοποιώντας μοντέλα ταξινόμησης. Και αυτό αποτελεί πρόβλημα ταξινόμησης.

Η ανάλυση του πρώτου μέρους της εργασίας επικεντρώνεται στην εκπαίδευση αυτών των μοντέλων και την αξιολόγησή τους, συγκρίνοντας τις επιδόσεις τους μέσω διαγραμμάτων και πινάκων.

Στο δεύτερο μέρος της εργασίας, ο στόχος είναι να εφαρμοστούν μοντέλα κατηγοριοποίησης στο σύνολο δεδομένων CIFAR-10, χρησιμοποιώντας αλγορίθμους όπως το K-Nearest Neighbors

(KNN) και τα Neural Networks. Εδώ, η εργασία επικεντρώνεται στην ανάλυση των αποτελεσμάτων της κατηγοριοποίησης εικόνας και στην αξιολόγηση της απόδοσης των μοντέλων μέσω γραφημάτων και πινάκων.

Στις επόμενες ενότητες της εργασίας, θα αναλύσουμε λεπτομερώς τα πειράματα και τις μεθόδους που χρησιμοποιήθηκαν για την εκπαιδευτική διαδικασία και αξιολόγηση των μοντέλων μηχανικής μάθησης. Στην ενότητα 2, θα περιγράψουμε τη διαδικασία εκπαίδευσης των μοντέλων για την πρόβλεψη της τιμής πώλησης και της κατηγοριοποίησης των δημοπρασιών στο eBay, ενώ θα εξετάσουμε την αξιολόγηση και τη σύγκριση της απόδοσης των επιλεγμένων μεθόδων. Στην ενότητα 3, θα επικεντρωθούμε στην ανάλυση των αποτελεσμάτων της κατηγοριοποίησης εικόνας με το σύνολο δεδομένων CIFAR-10, περιλαμβάνοντας την εφαρμογή αλγορίθμων όπως το K-Nearest Neighbors και τα Neural Networks, και θα συγκρίνουμε την απόδοση των μοντέλων μέσω των αντίστοιχων μετρικών. Τέλος, θα καταλήξουμε σε συμπεράσματα και προτάσεις για μελλοντική έρευνα που θα μπορούσε να οδηγήσει σε βελτιώσεις ή νέες κατευθύνσεις για την ανάλυση δεδομένων του eBay και της κατηγοριοποίησης εικόνας.

2. Συστήματα και Μέθοδοι

Περιγραφή του Συνόλου Δεδομένων του eBay

Το σύνολο δεδομένων που χρησιμοποιείται στην εργασία περιλαμβάνει δεδομένα από δημοπρασίες στην πλατφόρμα eBay και περιλαμβάνει τα εξής αρχεία CSV:

TrainingSet.csv και TestSet.csv: Αυτά τα αρχεία περιλαμβάνουν όλο το σύνολο δεδομένων και χρησιμοποιούνται για την εκπαίδευση και δοκιμή των μοντέλων. Κάθε γραμμή του συνόλου δεδομένων αναφέρεται σε μία δημοπρασία και περιλαμβάνει χαρακτηριστικά όπως η τελική τιμή πώλησης (Price), η πιθανότητα πώλησης (QuantitySold), καθώς και πληροφορίες για τον πωλητή και τα προϊόντα.

TrainingSubset.csv και TestSubset.csv: Αυτά τα αρχεία περιλαμβάνουν υποσύνολα των δεδομένων, στα οποία οι δημοπρασίες έχουν καταλήξει σε πώληση. Το "TrainingSet.csv"

χρησιμοποιείται για την εκπαίδευση των μοντέλων, ενώ το "TestSet.csv" για τη δοκιμή τους Αυτά τα δεδομένα χρησιμοποιούνται κυρίως για την πρόβλεψη των τιμών πώλησης και της σύγκρισης με τη μέση τιμή του αντικειμένου (Price vs AvgPrice).

Τα κύρια χαρακτηριστικά του συνόλου δεδομένων περιλαμβάνουν τα εξής:

- Price: Η τελική τιμή στην οποία έκλεισε η δημοπρασία. Εάν η δημοπρασία δεν οδήγησε σε πώληση, η τιμή μπορεί να είναι μηδενική ή να παραμείνει ίση με την αρχική προσφορά (StartingBid).
- QuantitySold: Ο αριθμός των αντικειμένων που πωλήθηκαν κατά την δημοπρασία, με τιμές 0 (όχι πώληση) ή 1 (πώληση).
- StartingBid: Η αρχική τιμή εκκίνησης της δημοπρασίας.
- BidCount: Ο αριθμός των προσφορών που κατατέθηκαν κατά τη διάρκεια της δημοπρασίας.
- SellerClosePercent: Η συνολική απόδοση του πωλητή στις δημοπρασίες του.
- SellerName: Το όνομα του πωλητή (αντί για SellerCountry που δεν περιλαμβάνεται στα δεδομένα).
- AvgPrice: Η μέση τιμή πώλησης για το ίδιο προϊόν (SKU).
-

Επιπλέον, το σύνολο δεδομένων περιλαμβάνει παράγωγα χαρακτηριστικά που εξάγονται από τις πληροφορίες των δημοπρασιών:

- IsHOF: Δείχνει αν ο αθλητής είναι μέλος του Hall of Fame.
- Authenticated: Δείχνει αν το αντικείμενο έχει πιστοποίηση αυθεντικότητας από τρίτο φορέα.
- SellerSaleAvgPriceRatio: Ο λόγος της τιμής πώλησης του αντικειμένου από έναν συγκεκριμένο πωλητή προς τη μέση τιμή για το ίδιο προϊόν.

Στην παρούσα εργασία, θα αναλύσουμε τα διαθέσιμα χαρακτηριστικά των δημοπρασιών στο eBay και θα εφαρμόσουμε διάφορους αλγορίθμους μηχανικής μάθησης για την πρόβλεψη της τελικής τιμής πώλησης και της πιθανότητας πώλησης ενός αντικειμένου. Συγκεκριμένα, θα εξετάσουμε τη συμβολή κάθε χαρακτηριστικού στην επιτυχία μιας δημοπρασίας, θα εκπαιδεύσουμε και θα αξιολογήσουμε μοντέλα όπως η Γραμμική Παλινδρόμηση και το Random Forest, και θα συγκρίνουμε την απόδοσή τους. Μέσω αυτής της διαδικασίας, στοχεύουμε να κατανοήσουμε καλύτερα τους παράγοντες που επηρεάζουν τις πωλήσεις στο eBay και να αναπτύξουμε αξιόπιστα μοντέλα πρόβλεψης.

Διαχωρισμός των επιμέρους προβλημάτων

eBay Dataset

1. Πρόβλεψη Τιμής Πώλησης (Price):

Για την εκτίμηση της τελικής τιμής πώλησης των δημοπρασιών, αξιοποιούμε τα υποσύνολα δεδομένων TrainingSubset.csv και TestSubset.csv, τα οποία περιλαμβάνουν μόνο τις δημοπρασίες που ολοκληρώθηκαν με επιτυχία (δηλαδή, όπου το QuantitySold είναι 1). Επιλέγουμε χαρακτηριστικά όπως η μέση τιμή του προϊόντος (AvgPrice), το ποσοστό επιτυχίας του πωλητή (SellerClosePercent), η αρχική τιμή προσφοράς (StartingBid) και το ποσοστό της τιμής πώλησης σε σχέση με την αρχική τιμή (PricePercent). Για την ανάπτυξη των μοντέλων πρόβλεψης, χρησιμοποιούμε αλγορίθμους όπως ο Random Forest και η Γραμμική Παλινδρόμηση (Linear Regression), οι οποίοι είναι κατάλληλοι για προβλήματα παλινδρόμησης.

2. Πρόβλεψη Πιθανότητας Πώλησης (QuantitySold):

Στόχος μας είναι να προβλέψουμε αν μια δημοπρασία θα οδηγήσει σε πώληση ή όχι. Για τον σκοπό αυτό, χρησιμοποιούμε το πλήρες σύνολο δεδομένων από τα αρχεία TrainingSet.csv και TestSet.csv, τα οποία περιλαμβάνουν όλες τις δημοπρασίες, ανεξαρτήτως έκβασης. Τα επιλεγμένα χαρακτηριστικά για την εκπαίδευση των μοντέλων περιλαμβάνουν τις τιμές Price, AvgPrice, SellerClosePercent, StartingBid, PricePercent και StartingBidPercent. Εφαρμόζουμε

αλγορίθμους όπως ο Random Forest και η Λογιστική Παλινδρόμηση (Logistic Regression), οι οποίοι είναι κατάλληλοι για προβλήματα κατηγοριοποίησης.

3. Πρόβλεψη της Σχέσης Τιμής και Μέσης Τιμής Πώλησης (Price vs AvgPrice):

Σε αυτή την ανάλυση, επιδιώκουμε να προβλέψουμε αν η τελική τιμή πώλησης μιας δημοπρασίας είναι μεγαλύτερη ή μικρότερη από τη μέση τιμή του αντίστοιχου προϊόντος. Χρησιμοποιούμε τα υποσύνολα δεδομένων TrainingSubset.csv και TestSubset.csv, φιλτραρισμένα για δημοπρασίες που κατέληξαν σε πώληση. Δημιουργούμε έναν δυαδικό στόχο, όπου η τιμή 1 αντιπροσωπεύει ότι η τελική τιμή είναι μεγαλύτερη ή ίση με τη μέση τιμή, ενώ η τιμή 0 υποδηλώνει ότι είναι μικρότερη. Τα χαρακτηριστικά που χρησιμοποιούνται περιλαμβάνουν τις μεταβλητές Price, AvgPrice, SellerClosePercent και StartingBid. Για την κατηγοριοποίηση, εφαρμόζουμε μοντέλα όπως ο Random Forest Classifier και η Λογιστική Παλινδρόμηση (Logistic Regression).

4. Κλίμακα και Διαχωρισμός Δεδομένων για Εκπαίδευση και Δοκιμή:

Πριν από την εκπαίδευση των μοντέλων, πραγματοποιούμε έλεγχο για τυχόν κενές τιμές στα δεδομένα. Για τα αριθμητικά χαρακτηριστικά, οι κενές τιμές αντικαθίστανται με τον μέσο όρο της αντίστοιχης στήλης, ενώ για τα κατηγορικά χαρακτηριστικά χρησιμοποιείται η πιο συχνή τιμή (mode). Επιπλέον, εφαρμόζουμε κλιμάκωση των δεδομένων χρησιμοποιώντας τον StandardScaler από τη βιβλιοθήκη SciKit-learn, ώστε να διασφαλίσουμε την ορθή εκπαίδευση των μοντέλων.

Με την παραπάνω μεθοδολογία, στοχεύουμε στην ανάπτυξη αξιόπιστων μοντέλων που θα συμβάλλουν στην κατανόηση των παραγόντων που επηρεάζουν τις πωλήσεις στο eBay και θα παρέχουν ακριβείς προβλέψεις για μελλοντικές δημοπρασίες.

3. Εφαρμογή Μεθόδων Πρόβλεψης Στο Σύνολο Δεδομένων Από Δημοπρασίες του eBay

Πειράματα για την Πρόβλεψη Τιμής (Price)

Στην ενότητα αυτή, επικεντρωνόμαστε στην πρόβλεψη της τελικής τιμής πώλησης των δημοπρασιών στο eBay, εφαρμόζοντας και αξιολογώντας δύο μοντέλα μηχανικής μάθησης: τον Random Forest Regressor και τη Γραμμική Παλινδρόμηση. Η διαδικασία περιλαμβάνει τα εξής στάδια:

1. Φόρτωση και Προεπεξεργασία Δεδομένων

Χρησιμοποιούμε τα αρχεία TrainingSubset.csv και TestSubset.csv, τα οποία περιέχουν δημοπρασίες που ολοκληρώθηκαν με πώληση. Αρχικά, αντιμετωπίζουμε τυχόν ελλιπή δεδομένα, αντικαθιστώντας τις αριθμητικές ελλείψεις με τη μέση τιμή κάθε στήλης και τις κατηγορικές με τη συχνότερη τιμή. Στη συνέχεια, εφαρμόζουμε τον StandardScaler για την κλιμάκωση των χαρακτηριστικών, διασφαλίζοντας ομοιομορφία στην κλίμακα των δεδομένων.

2. Επιλογή Χαρακτηριστικών και Εκπαίδευση Μοντέλων

Τα χαρακτηριστικά που χρησιμοποιούνται για την εκπαίδευση των μοντέλων είναι τα εξής:

- **AvgPrice**: Μέση τιμή του αντικειμένου.
- **SellerClosePercent**: Ποσοστό επιτυχημένων δημοπρασιών του πωλητή.
- **StartingBid**: Αρχική τιμή εκκίνησης της δημοπρασίας.
- **PricePercent**: Ποσοστό της τελικής τιμής σε σχέση με την αρχική.

Εκπαιδεύουμε δύο μοντέλα:

1) Random Forest Regressor: Ένα σύνολο απόφασης δέντρων που συνδυάζονται για την πρόβλεψη της τιμής.
2) Γραμμική Παλινδρόμηση: Μοντέλο που υποθέτει γραμμική σχέση μεταξύ χαρακτηριστικών και τιμής.Διαχωρισμός Δεδομένων σε Εκπαίδευση και Δοκιμή

3. Διαχωρισμός Δεδομένων και Αξιολόγηση Μοντέλων

Χωρίζουμε τα δεδομένα σε εκπαιδευτικό και δοκιμαστικό σύνολο χρησιμοποιώντας τη μέθοδο train_test_split (80% εκπαίδευση, 20% δοκιμή). Επιπλέον, εφαρμόζουμε διασταυρούμενη επικύρωση (cross-validation) για την αξιολόγηση των μοντέλων. Οι μετρικές που χρησιμοποιούμε περιλαμβάνουν:

- Mean Squared Error (MSE): Μέσο τετραγωνικό σφάλμα.
- Root Mean Squared Error (RMSE): Τετραγωνική ρίζα του MSE.
- R^2 (Συντελεστής Προσδιορισμού): Δείκτης που μετρά την ικανότητα του μοντέλου να εξηγεί τη διακύμανση των δεδομένων.

5. Οπτικοποίηση Αποτελεσμάτων

Δημιουργούμε γραφήματα που συγκρίνουν τις πραγματικές με τις προβλεπόμενες τιμές, καθώς και ιστογράμματα των σφαλμάτων πρόβλεψης, για να αξιολογήσουμε την απόδοση των μοντέλων.

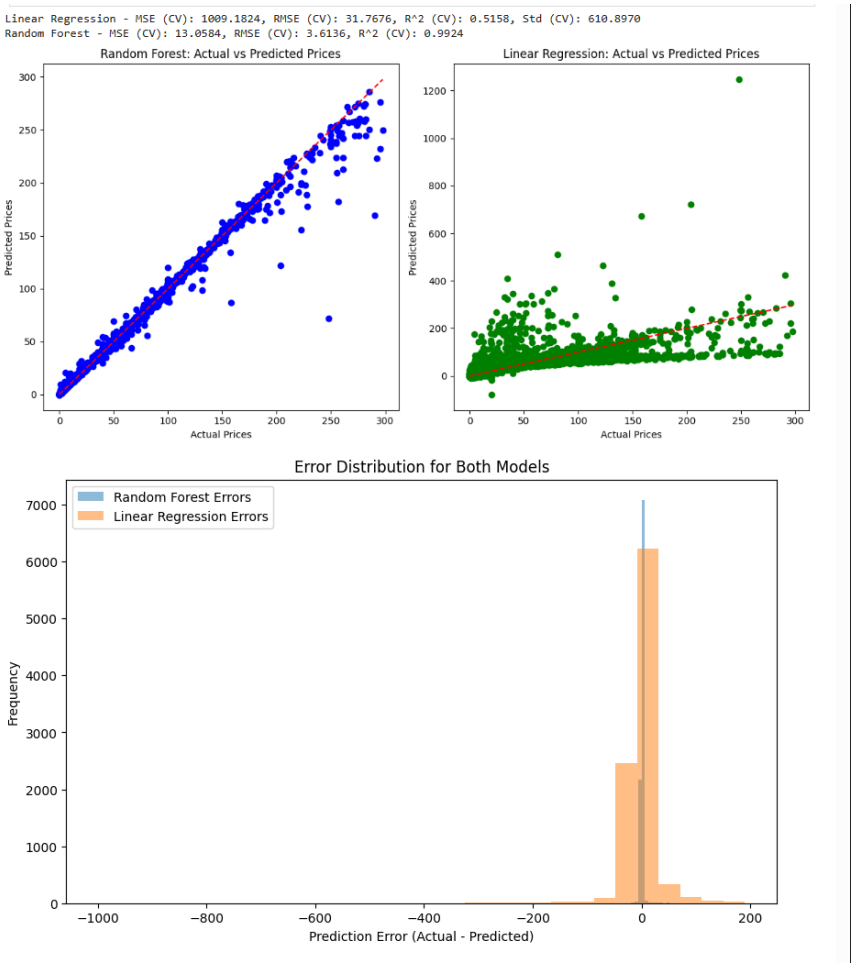

Figure 1 Πραγματικά και Προβλεπόμενα Αποτελέσματα για τα δύο μοντέλα

Στα συμπεράσματα (5), θα παρουσιάσουμε λεπτομερή ανάλυση των επιδόσεων των μοντέλων, συγκρίνοντας τις μετρικές και προτείνοντας βελτιώσεις για μελλοντική εργασία.

Πειράματα για την Πρόβλεψη Πωλήσεων (QuantitySold)

Η πρόβλεψη της πιθανότητας μια δημοπρασία στο eBay να οδηγήσει σε πώληση αποτελεί σημαντικό στόχο της ανάλυσής μας. Για την επίτευξη αυτού του στόχου, εφαρμόσαμε δύο αλγορίθμους μηχανικής μάθησης: τον Random Forest Classifier και τη Λογιστική Παλινδρόμηση. Η διαδικασία περιλάμβανε την προετοιμασία των δεδομένων, την εκπαίδευση των μοντέλων και την αξιολόγησή τους.

1. Φόρτωση και Προεπεξεργασία Δεδομένων

Χρησιμοποιήσαμε τα σύνολα δεδομένων TrainingSet.csv και TestSet.csv, τα οποία περιλαμβάνουν όλες τις δημοπρασίες, ανεξαρτήτως αν κατέληξαν σε πώληση ή όχι. Αρχικά, διαχωρίσαμε τα αριθμητικά και κατηγορικά χαρακτηριστικά. Τα κενά στις αριθμητικές στήλες συμπληρώθηκαν με τη μέση τιμή κάθε στήλης, ενώ στις κατηγορικές στήλες με τη συχνότερη τιμή (mode). Στη συνέχεια, εφαρμόσαμε τον StandardScaler για την κλιμάκωση των δεδομένων, διασφαλίζοντας ότι όλα τα χαρακτηριστικά βρίσκονται στην ίδια κλίμακα, βελτιώνοντας έτσι την απόδοση των μοντέλων.

2. Εκπαίδευση Μοντέλων

Για την εκπαίδευση των μοντέλων, επιλέξαμε τα εξής χαρακτηριστικά:

- Price: Η τελική τιμή πώλησης.
- AvgPrice: Η μέση τιμή πώλησης για το ίδιο προϊόν.
- SellerClosePercent: Το ποσοστό των δημοπρασιών του πωλητή που κατέληξαν σε πώληση.
- StartingBid: Η αρχική τιμή εκκίνησης της δημοπρασίας.
- PricePercent: Το ποσοστό της τελικής τιμής πώλησης σε σχέση με την αρχική τιμή.

Αυτά τα χαρακτηριστικά χρησιμοποιήθηκαν για την εκπαίδευση των ακόλουθων μοντέλων:

- Random Forest Classifier: Ένας αλγόριθμος που χρησιμοποιεί πολλαπλά δέντρα απόφασης για την ταξινόμηση των δεδομένων.
- Λογιστική Παλινδρόμηση: Ένα μοντέλο που εκτιμά την πιθανότητα ενός δυαδικού αποτελέσματος.

3. Διαχωρισμός Δεδομένων σε Εκπαίδευση και Δοκιμή

Χωρίσαμε τα δεδομένα σε εκπαιδευτικό και δοκιμαστικό σύνολο χρησιμοποιώντας τη μέθοδο train_test_split με αναλογία 80% για εκπαίδευση και 20% για δοκιμή. Για τον Random Forest Classifier, πραγματοποιήσαμε υπερπαραμετρική αναζήτηση (GridSearchCV) με διασταυρούμενη επικύρωση (cross-validation) για τον εντοπισμό των βέλτιστων παραμέτρων.

4. Αξιολόγηση των Μοντέλων

Η αξιολόγηση των μοντέλων έγινε με τις εξής μετρικές:

- Ακρίβεια (Accuracy): Το ποσοστό των σωστών προβλέψεων.
- Μήτρα Σύγχυσης (Confusion Matrix): Πίνακας που παρουσιάζει τις πραγματικές έναντι των προβλεπόμενων κατηγοριών.
- Αναφορά Ταξινόμησης (Classification Report): Παρέχει μετρικές όπως ακρίβεια, ανάκληση και F1-score.
- Καμπύλη ROC και AUC: Αξιολογούν την ικανότητα του μοντέλου να διαχωρίζει τις κατηγορίες.

Με βάση αυτές τις μετρικές, συγκρίναμε τις επιδόσεις των μοντέλων για να επιλέξουμε το καταλληλότερο.

5. Οπτικοποίηση Αποτελεσμάτων

Κατά την αξιολόγηση των μοντέλων, δημιουργήσαμε καμπύλες ROC και μήτρες σύγχυσης για καθένα από αυτά, προκειμένου να συγκρίνουμε τις επιδόσεις τους.

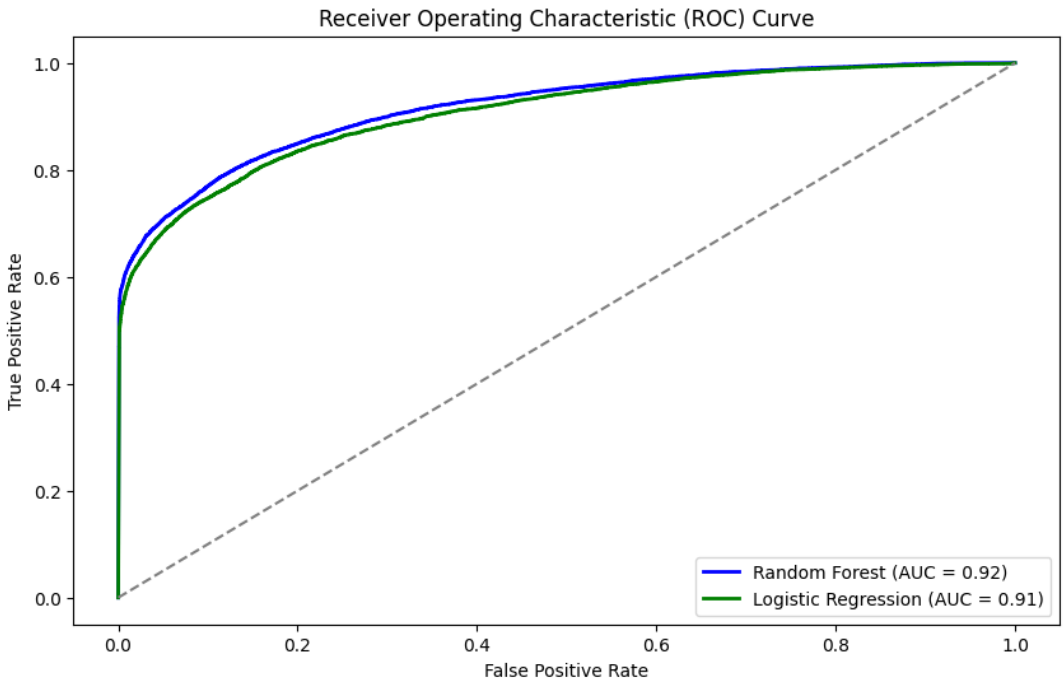

Figure 2 Καμπύλη ROC για τα δύο μοντέλα

```
Confusion Matrix for Logistic Regression
[[26036  2025]
 [ 2669  6730]]
```

```
Confusion Matrix for Random Forest:
[[26634  1427]
 [ 2812  6587]]
```

Figure 3 Μήτρες Σύγχυσης για τα δύο μοντέλα

Στα συμπεράσματα της μελέτης, θα παρουσιαστούν λεπτομερώς τα αποτελέσματα και οι προτάσεις για τη βελτίωση των μοντέλων.

Πειράματα για την Πρόβλεψη της Σχέσης Τιμής και Μέσης Τιμής Πώλησης (Price vs AvgPrice)

Στην παρούσα ενότητα, επικεντρωνόμαστε στην πρόβλεψη του αν η τελική τιμή πώλησης μιας δημοπρασίας στο eBay είναι μεγαλύτερη ή μικρότερη από τη μέση τιμή του αντίστοιχου αντικειμένου (AvgPrice). Αυτό συνιστά ένα δυαδικό πρόβλημα κατηγοριοποίησης, όπου η τιμή στόχος ορίζεται ως 1 αν η Price είναι μεγαλύτερη ή ίση με την AvgPrice, και 0 αν είναι μικρότερη. Για την επίλυση αυτού του προβλήματος, εφαρμόζουμε δύο αλγορίθμους μηχανικής μάθησης: τον Random Forest Classifier και τη Λογιστική Παλινδρόμηση (Logistic Regression).

1. Φόρτωση και Προεπεξεργασία Δεδομένων

Χρησιμοποιούμε τα σύνολα δεδομένων TrainingSubset.csv και TestSubset.csv, τα οποία περιλαμβάνουν μόνο τις δημοπρασίες που κατέληξαν σε πώληση. Για τη δημιουργία του δυαδικού στόχου, συγκρίνουμε την τιμή πώλησης (Price) με τη μέση τιμή του αντικειμένου (AvgPrice), ορίζοντας την τιμή στόχο ως 1 αν η Price είναι μεγαλύτερη ή ίση με την AvgPrice, και 0 αν είναι μικρότερη.

Τα χαρακτηριστικά που χρησιμοποιούνται για την εκπαίδευση των μοντέλων περιλαμβάνουν:

- Price: Η τελική τιμή στην οποία έκλεισε η δημοπρασία.
- AvgPrice: Η μέση τιμή του αντικειμένου.
- SellerClosePercent: Το ποσοστό των δημοπρασιών του πωλητή που κατέληξαν σε πώληση.
- StartingBid: Η αρχική τιμή εκκίνησης της δημοπρασίας.
- PricePercent: Το ποσοστό της τελικής τιμής πώλησης σε σχέση με την αρχική τιμή.
- StartingBidPercent: Το ποσοστό της αρχικής τιμής εκκίνησης σε σχέση με την τιμή του αντικειμένου.

2. Διαχωρισμός Δεδομένων σε Εκπαίδευση και Δοκιμή

Χωρίζουμε το σύνολο δεδομένων σε εκπαιδευτικό και δοκιμαστικό υποσύνολο χρησιμοποιώντας τη μέθοδο train_test_split, με αναλογία 80% για εκπαίδευση και 20% για δοκιμή. Επιπλέον, εφαρμόζουμε 5-fold cross-validation κατά την εκπαίδευση των μοντέλων για να διασφαλίσουμε την αξιοπιστία των αποτελεσμάτων.

3. Εκπαίδευση Μοντέλων

Εφαρμόζουμε τους εξής αλγορίθμους κατηγοριοποίησης:

- Random Forest Classifier: Ένας αλγόριθμος που βασίζεται στη δημιουργία πολλαπλών δέντρων απόφασης και καθορίζει την τελική κατηγορία μέσω της πλειοψηφίας των αποφάσεων των δέντρων.
- Λογιστική Παλινδρόμηση (Logistic Regression): Ένας αλγόριθμος κατηγοριοποίησης που χρησιμοποιεί μια γραμμική συνάρτηση για την εκτίμηση της πιθανότητας μιας δυαδικής έκβασης.

4. Αξιολόγηση των Μοντέλων

Η αξιολόγηση των μοντέλων πραγματοποιείται με τις εξής μετρικές:

- Ακρίβεια (Accuracy): Το ποσοστό των σωστών προβλέψεων σε σχέση με το σύνολο των παρατηρήσεων.
- Μήτρα Σύγχυσης (Confusion Matrix): Ένας πίνακας που επιτρέπει την ανάλυση των επιδόσεων του μοντέλου, παρουσιάζοντας τις πραγματικές έναντι των προβλεπόμενων κατηγοριών.
- Αναφορά Ταξινόμησης (Classification Report): Περιλαμβάνει μετρικές όπως precision, recall και F1-score για κάθε κατηγορία.
- Καμπύλη ROC και AUC: Η καμπύλη ROC απεικονίζει την απόδοση του μοντέλου σε διάφορα κατώφλια ταξινόμησης, ενώ το AUC (Area Under the Curve) παρέχει μια συνολική μέτρηση της ικανότητας του μοντέλου να διαχωρίζει τις κατηγορίες.

5. Οπτικοποίηση Αποτελεσμάτων

Για την αξιολόγηση των μοντέλων, δημιουργούμε καμπύλες ROC και υπολογίζουμε το AUC για καθένα από αυτά. Επιπλέον, παρουσιάζουμε τις μήτρες σύγχυσης και τις αναφορές ταξινόμησης, προκειμένου να αξιολογήσουμε την απόδοση των μοντέλων και να προσδιορίσουμε ποιο προσφέρει την καλύτερη ακρίβεια και ποιότητα ταξινόμησης για το συγκεκριμένο πρόβλημα.

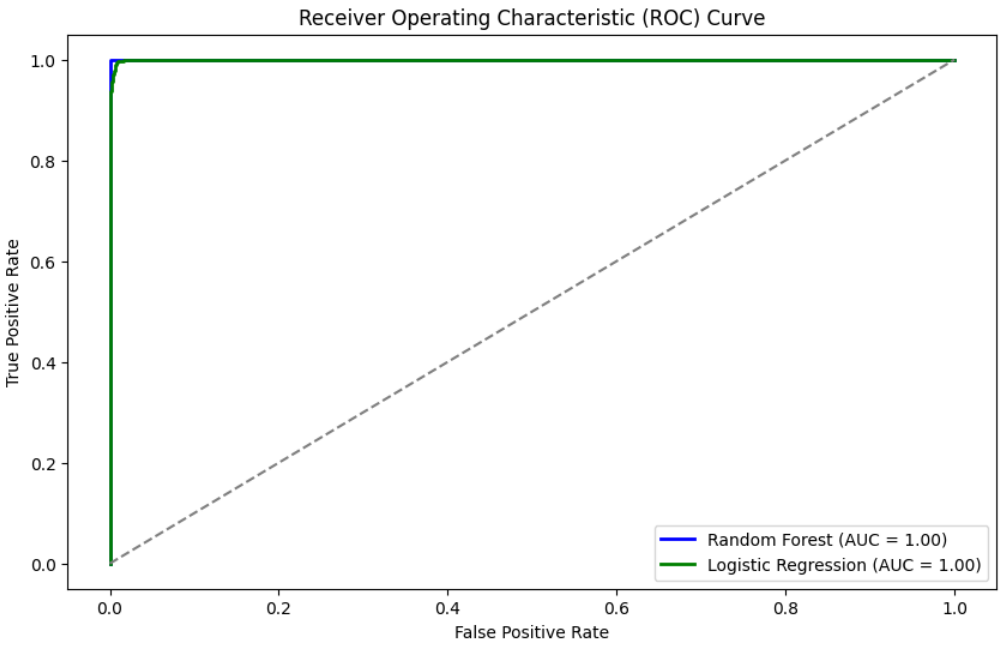

Figure 4 ROC & AUC για τα δύο μοντέλα

```
Confusion Matrix for Random Forest:
[[6700    0]
 [   0 2699]]
```

```
Confusion Matrix for Logistic Regression:
[[6484  216]
 [   3 2696]]
```

Figure 5 Μήτρες Σύγχυσης για τα δύο μοντέλα

Στα συμπεράσματα της μελέτης (Ενότητα 5), θα αναλύσουμε τις επιδόσεις των μοντέλων, θα συγκρίνουμε τις μετρικές αξιολόγησης και θα παρουσιάσουμε τις τελικές παρατηρήσεις μας.

4. Εφαρμογή Μοντέλων Κατηγοριοποίησης στο CIFAR-10

Περιγραφή του Συνόλου Δεδομένων CIFAR-10

Το σύνολο δεδομένων CIFAR-10 (Canadian Institute for Advanced Research) περιλαμβάνει 60.000 έγχρωμες εικόνες σε 10 διαφορετικές κατηγορίες, με 6.000 εικόνες ανά κατηγορία. Το σύνολο αυτό παρέχει μια πλούσια βάση για πειραματισμούς με αλγορίθμους μηχανικής μάθησης, καθώς περιλαμβάνει ένα ευρύ φάσμα εικόνων που αναπαριστούν καθημερινά αντικείμενα, γεγονός που το καθιστά ιδανικό για την εκπαίδευση μοντέλων κατηγοριοποίησης εικόνας [1].

Δομή του Συνόλου Δεδομένων

Το CIFAR-10 χωρίζεται σε δύο βασικά υποσύνολα: το εκπαιδευτικό υποσύνολο (Training Set) και το δοκιμαστικό υποσύνολο (Test Set). Το Training Set περιλαμβάνει 50.000 εικόνες, ενώ το Test Set περιλαμβάνει 10.000 εικόνες. Κάθε εικόνα έχει διάσταση 32x32x3 (πλάτος x ύψος x κανάλια χρώματος RGB), κάτι που σημαίνει ότι κάθε εικόνα είναι έγχρωμη και έχει 3 κανάλια για τα χρώματα Red, Green και Blue [2].

Οι 10 κατηγορίες του CIFAR-10 είναι οι εξής:

- Αεροπλάνο (Airplane)
- Αυτοκίνητο (Automobile)
- Πουλί (Bird)
- Γάτα (Cat)
- Ελάφι (Deer)
- Σκύλος (Dog)
- Φρόιντ (Frog)
- Άλογο (Horse)
- Σκάφος (Ship)
- Τρένο (Truck)

Κάθε κατηγορία περιλαμβάνει 6.000 εικόνες, οι οποίες είναι χωρισμένες ισόποσα μεταξύ των εκπαιδευτικών και δοκιμαστικών δεδομένων. Έτσι, το σύνολο δεδομένων περιέχει συνολικά

60.000 εικόνες, που έχουν ποικιλία σε τύπους αντικειμένων και σκηνές, όπως οχήματα, ζώα, και άλλα καθημερινά αντικείμενα [3].

Κατηγορίες Εικόνας

Κάθε εικόνα του CIFAR-10 αντιστοιχεί σε μία από τις παραπάνω κατηγορίες. Οι εικόνες περιέχουν αντικείμενα με διαφορετικές κλίμακες, γωνίες και φωτισμό, και συχνά περιλαμβάνουν περισσότερα από ένα αντικείμενα στην ίδια εικόνα. Αυτή η ποικιλία καθιστά το σύνολο δεδομένων ιδανικό για την εκπαίδευση μοντέλων κατηγοριοποίησης εικόνας με πραγματικές συνθήκες [1].

Για παράδειγμα:

Αεροπλάνα και Αυτοκίνητα συχνά έχουν φόντο τον ουρανό ή τον δρόμο, ενώ τα Πουλιά μπορεί να εμφανίζονται σε φυσικά περιβάλλοντα ή να πετούν στον αέρα.

Οι Γάτες και τα Σκύλοι εμφανίζονται συχνά σε εσωτερικούς ή εξωτερικούς χώρους, ενώ τα Φρόιντ και Άλογα εμφανίζονται σε σκηνές της φύσης ή ζώων [3].

Χρησιμότητα και Σκοπός

Το CIFAR-10 είναι ένα σύνολο δεδομένων που έχει σχεδιαστεί για την εκπαίδευση και αξιολόγηση αλγορίθμων κατηγοριοποίησης εικόνας, και είναι ευρέως χρησιμοποιούμενο σε πειράματα για τη βελτίωση αλγορίθμων συμπίεσης εικόνας, σχεδιασμού νευρωνικών δικτύων και άλλων εφαρμογών μηχανικής μάθησης. Η ποικιλία στις κατηγορίες και το μέγεθος των εικόνων το καθιστούν κατάλληλο για την εκπαίδευση επιφανειακών αλγορίθμων όπως τα SVM, τα Random Forests και τα K-Nearest Neighbors (KNN), καθώς και για την εκπαίδευση πιο σύνθετων συνελικτικών νευρωνικών δικτύων (CNN) [4].

Χρησιμοποιούμενα Εργαλεία και Βιβλιοθήκες

Για την επεξεργασία και ανάλυση του συνόλου δεδομένων CIFAR-10, χρησιμοποιήθηκαν οι βιβλιοθήκες TensorFlow και SciKit-learn για τη φόρτωση του συνόλου δεδομένων και την εκπαίδευση των μοντέλων αντίστοιχα. Επιπλέον, χρησιμοποιήθηκαν τεχνικές όπως PCA (Principal Component Analysis) για μείωση της διάστασης των δεδομένων και Min-Max Scaling για κανονικοποίηση των εικόνων πριν από την εκπαίδευση των μοντέλων [6].

Διαχωρισμός του Συνόλου Δεδομένων

CIFAR-10 Dataset

Η ανάλυση που πραγματοποιήσαμε στο σύνολο δεδομένων CIFAR-10 επικεντρώνεται στην ταξινόμηση εικόνων, μελετώντας την απόδοση διαφορετικών αλγορίθμων σε υποσύνολα δεδομένων με διαφορετικά επίπεδα πολυπλοκότητας. Συγκεκριμένα, σχεδιάσαμε τρεις περιπτώσεις πειραμάτων, στις οποίες επιλέξαμε 2, 4 και 6 κατηγορίες εικόνων, αντί για το πλήρες σύνολο των 10 κατηγοριών. Αυτή η προσέγγιση μας επιτρέπει να αναλύσουμε πώς η αύξηση του αριθμού των κατηγοριών επηρεάζει την απόδοση των ταξινομητών και να προσαρμόσουμε ανάλογα την επιλογή των χαρακτηριστικών και των αλγορίθμων.

1 Επιλογή Υποσυνόλων Δεδομένων και Διαχωρισμός Δεδομένων

Το CIFAR-10 αποτελείται από 60.000 έγχρωμες εικόνες (32x32x3), κατανεμημένες ισομερώς σε 10 κατηγορίες.

Για να καταστήσουμε τη μελέτη πιο εστιασμένη, διαχωρίσαμε το σύνολο δεδομένων στις εξής περιπτώσεις:

- 2 Κατηγορίες → Επιλέξαμε δύο κατηγορίες με διακριτά χαρακτηριστικά.
- Κατηγορίες → Αυξήσαμε την πολυπλοκότητα εισάγοντας περισσότερες κατηγορίες από διαφορετικές κλάσεις.
- Κατηγορίες → Περιλαμβάνοντας περισσότερες κατηγορίες, εξετάσαμε πώς η αύξηση της πολυπλοκότητας επηρεάζει την ακρίβεια των ταξινομητών.

Η επιλογή των κατηγοριών έγινε με np.isin, ώστε να φιλτραριστούν οι εικόνες που ανήκουν στις επιλεγμένες κατηγορίες.

Διαχωρισμός Εκπαιδευτικών και Δοκιμαστικών Δεδομένων

Το σύνολο δεδομένων χωρίστηκε σε εκπαιδευτικό (training), επικυρωτικό (validation) και δοκιμαστικό (testing) υποσύνολο:

- 80% των δεδομένων χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων.
- Το υπόλοιπο 20% κρατήθηκε για δοκιμή.
- Ο διαχωρισμός πραγματοποιήθηκε με τη train_test_split της SciKit-learn, διασφαλίζοντας ότι τα δεδομένα είναι ισορροπημένα μεταξύ των κατηγοριών.

2 Προεπεξεργασία Δεδομένων

Για τη σωστή εκπαίδευση των μοντέλων, εφαρμόσαμε τις εξής τεχνικές προεπεξεργασίας:

Κανονικοποίηση Δεδομένων

- Τα δεδομένα κλιμακώθηκαν στο διάστημα [0,1] μέσω Min-Max Scaling, ώστε όλα τα χαρακτηριστικά να έχουν ενιαία κλίμακα, επιταχύνοντας την εκπαίδευση των μοντέλων.

Αναδιάταξη Δεδομένων (Reshaping & Flattening)

- Οι εικόνες (32x32x3) μετατράπηκαν σε μονοδιάστατα διανύσματα (flattening) για να είναι συμβατές με αλγορίθμους όπως SVM και Random Forest.

Μείωση Διάστασης (PCA)

- Για να μειώσουμε τον αριθμό των χαρακτηριστικών και να επιταχύνουμε την εκπαίδευση των μοντέλων, εφαρμόσαμε Principal Component Analysis (PCA).
- Διατηρήσαμε το 95% της διακύμανσης των δεδομένων, μειώνοντας έτσι την πολυπλοκότητα, χωρίς να χάνουμε κρίσιμες πληροφορίες.

3 Επιλογή και Εκπαίδευση Μοντέλων

Για την ταξινόμηση των εικόνων, χρησιμοποιήσαμε τρεις διαφορετικούς αλγορίθμους μηχανικής μάθησης:

1) Support Vector Machine (SVM)
    - Εκπαιδεύτηκε με γραμμικό πυρήνα (linear kernel) και παραμέτρους C=1.

- Το SVM είναι ιδιαίτερα αποτελεσματικό για γραμμικά διαχωρίσιμα δεδομένα και λειτουργεί καλά με υψηλές διαστάσεις.

2) Random Forest Classifier
   - Χρησιμοποιεί 100 δέντρα απόφασης (trees) με μέγιστο βάθος 10.
   - Το Random Forest συνδυάζει πολλά δέντρα απόφασης για να μειώσει το πρόβλημα της υπερεκπαίδευσης και να βελτιώσει την ακρίβεια.
3) K-Nearest Neighbors (KNN)
   - Χρησιμοποιεί k=5 γειτονικά σημεία (neighbors) για την ταξινόμηση.

Το KNN είναι κατάλληλο για μη γραμμικά προβλήματα, αλλά μπορεί να είναι αργό για μεγάλα σύνολα δεδομένων.

4. Εκπαίδευση και Αξιολόγηση των Μοντέλων

Για την εκπαίδευση και αξιολόγηση των μοντέλων, ακολουθήσαμε τα εξής βήματα:

Εκπαίδευση

- Χρησιμοποιήσαμε PCA-transformed και Min-Max scaled δεδομένα για την εκπαίδευση κάθε μοντέλου.

Επικύρωση

- Οι επιδόσεις των μοντέλων αξιολογήθηκαν μέσω μετρικών όπως:
  - Accuracy (ποσοστό σωστών προβλέψεων).
  - Precision, Recall και F1-score (για κάθε κατηγορία εικόνας).
  - Confusion Matrix (ανάλυση των ταξινομήσεων και των σφαλμάτων).

Σταυρωτή Επικύρωση (Cross-Validation)

- Εφαρμόσαμε 5-fold cross-validation, έτσι ώστε το μοντέλο να αξιολογηθεί σε διαφορετικά υποσύνολα των δεδομένων, βελτιώνοντας την αξιοπιστία των αποτελεσμάτων.

5. Αξιολόγηση Απόδοσης των Μοντέλων

Η ακρίβεια των ταξινομητών μειώνεται όσο αυξάνεται ο αριθμός των κατηγοριών, καθώς το πρόβλημα γίνεται πιο σύνθετο.

Αναλυτικά:

- Με 2 κατηγορίες, τα περισσότερα μοντέλα πέτυχαν υψηλή ακρίβεια.

- Με 4 κατηγορίες, η ακρίβεια μειώθηκε, καθώς η διάκριση μεταξύ των κατηγοριών έγινε δυσκολότερη.
- Με 6 κατηγορίες, τα μοντέλα δυσκολεύτηκαν ακόμα περισσότερο, απαιτώντας πιο πολύπλοκες αρχιτεκτονικές.

Για την αξιολόγηση, χρησιμοποιήσαμε:

- Confusion Matrix για ανάλυση λαθών.
- ROC Curve & AUC για την εκτίμηση της ποιότητας των προβλέψεων.

5. Συμπεράσματα

Συμπεράσματα και Σύγκριση των Μοντέλων Πρόβλεψης eBay

Στην παρούσα μελέτη, εξετάσαμε την απόδοση διαφόρων μοντέλων μηχανικής μάθησης στην πρόβλεψη χαρακτηριστικών των δημοπρασιών του eBay. Συγκεκριμένα, εστιάσαμε στην πρόβλεψη της τελικής τιμής πώλησης (Price), στην πιθανότητα μιας δημοπρασίας να οδηγήσει σε πώληση (QuantitySold) και στο αν η τιμή πώλησης είναι μεγαλύτερη ή μικρότερη από τη μέση τιμή του αντικειμένου (Price vs AvgPrice). Για κάθε ένα από αυτά τα προβλήματα, χρησιμοποιήσαμε τα μοντέλα Random Forest, Linear Regression και Logistic Regression, και αξιολογήσαμε την απόδοσή τους με βάση μετρικές όπως το μέσο τετραγωνικό σφάλμα (MSE), η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), ο συντελεστής προσδιορισμού (R^2), η ακρίβεια (accuracy), η ανάκληση (recall) και το F1-Score.

Τα αποτελέσματα συνοψίζονται στους παρακάτω πίνακες:

Πίνακας 1: Αποτελέσματα Πρόβλεψης Τιμής Πώλησης (Price)

Μοντέλο	MSE (CV)	RMSE (CV)	R^2 (CV)	Std (CV)
Linear Regression	1009.1824	31.7676	0.5158	610.8970
Random Forest	13.0584	3.6136	0.9924	-

Συμπέρασμα: Το μοντέλο Random Forest παρουσιάζει σαφή υπεροχή στην πρόβλεψη της τιμής πώλησης, με σημαντικά χαμηλότερο MSE και υψηλότερο R^2 σε σύγκριση με το Linear Regression.

Πίνακας 2: Αποτελέσματα Πρόβλεψης Πιθανότητας Πώλησης (QuantitySold)

Μοντέλο	Ακρίβεια Επικύρωσης	Ακρίβεια Δοκιμής	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Random Forest	0.8866	0.8964	0.92	0.64	0.76
Logistic Regression	0.8766	0.8892	0.92	0.62	0.74

Συμπέρασμα: Το Random Forest Classifier υπερέχει του Logistic Regression στην πρόβλεψη της πιθανότητας πώλησης, παρουσιάζοντας υψηλότερη ακρίβεια, precision και F1-Score.

Πίνακας 3: Αποτελέσματα Πρόβλεψης Σχέσης Τιμής και Μέσης Τιμής Πώλησης (Price vs AvgPrice)

Μοντέλο	Ακρίβεια Επικύρωσης	Ακρίβεια Δοκιμής	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Random Forest	1.0000	1.0000	1.00	1.00	1.00
Logistic Regression	0.9767	0.9767	0.93	1.00	0.96

Συμπέρασμα: Στην πρόβλεψη της σχέσης τιμής και μέσης τιμής πώλησης, το Random Forest Classifier επιτυγχάνει άριστη απόδοση με ακρίβεια 100%, ενώ το Logistic Regression παρουσιάζει ελαφρώς χαμηλότερη απόδοση.

Αντιμετώπιση Υπερπροσαρμογής (Overfitting):

Για την αντιμετώπιση του φαινομένου της υπερπροσαρμογής, εφαρμόστηκαν οι εξής τεχνικές:

- Διαχωρισμός Δεδομένων: Χωρίσαμε τα δεδομένα σε εκπαιδευτικό και δοκιμαστικό σύνολο, διασφαλίζοντας ότι η αξιολόγηση των μοντέλων γίνεται σε μη ορατά δεδομένα κατά την εκπαίδευση.
- Διασταυρούμενη Επικύρωση (Cross-Validation): Χρησιμοποιήθηκε η τεχνική της διασταυρούμενης επικύρωσης (cross-validation) για την αξιόπιστη εκτίμηση της απόδοσης των μοντέλων και την αποφυγή υπερπροσαρμογής.
- Ρύθμιση Υπερπαραμέτρων (Hyperparameter Tuning): Για το Random Forest, πραγματοποιήθηκε ρύθμιση των υπερπαραμέτρων μέσω της μεθόδου GridSearchCV, βελτιστοποιώντας παραμέτρους όπως το μέγιστο βάθος των δέντρων, ο αριθμός των δέντρων και το ελάχιστο δείγμα φύλλων.

Συνολικά, τα αποτελέσματα υποδεικνύουν ότι το μοντέλο Random Forest παρουσιάζει ανώτερη απόδοση στις περισσότερες περιπτώσεις, καθιστώντας το κατάλληλο εργαλείο για την πρόβλεψη χαρακτηριστικών των δημοπρασιών του eBay. Ωστόσο, η επιλογή του κατάλληλου μοντέλου εξαρτάται από τη φύση των δεδομένων και τις συγκεκριμένες απαιτήσεις της εκάστοτε εφαρμογής.

Συμπεράσματα και Συγκριτική Ανάλυση Αποτελεσμάτων για το CIFAR-10

Στην ενότητα αυτή, συγκρίνονται τα αποτελέσματα των πειραμάτων κατηγοριοποίησης εικόνας χρησιμοποιώντας το σύνολο δεδομένων CIFAR-10. Δοκιμάστηκαν τρεις αλγόριθμοι κατηγοριοποίησης: Support Vector Machines (SVM), Random Forest (RF) και K-Nearest Neighbors (KNN), σε τρία διαφορετικά σύνολα δεδομένων με 2, 4 και 6 κατηγορίες. Οι επιδόσεις τους αξιολογήθηκαν με βάση την ακρίβεια (accuracy), την F1-score, την μήτρα σύγχυσης και την υπερπροσαρμογή (overfitting) των μοντέλων.

Συγκριτική Ανάλυση των Μοντέλων

- Δύο κατηγορίες: Το Random Forest (RF) παρουσίασε την καλύτερη απόδοση, με accuracy 86%, ενώ το SVM ακολούθησε με 83%. Το KNN είχε χαμηλότερη απόδοση (72%), με πολύ υψηλή recall για τη μία κατηγορία αλλά χαμηλή recall για την άλλη, υποδηλώνοντας προβλήματα ισορροπίας στο μοντέλο.
- Τέσσερις κατηγορίες: Η απόδοση όλων των μοντέλων μειώθηκε, καθώς το πρόβλημα έγινε πιο περίπλοκο. Το RF παρέμεινε το ισχυρότερο μοντέλο (66%), ακολουθούμενο

από το SVM (63%) και το KNN (59%). Το RF είχε καλύτερη συνολική ισορροπία μεταξύ recall και precision, ενώ το KNN υπέφερε από χαμηλή γενίκευση.

- Έξι κατηγορίες: Η ακρίβεια όλων των μοντέλων μειώθηκε περαιτέρω, λόγω της αυξημένης πολυπλοκότητας. Το Random Forest διατήρησε προβάδισμα (52%), ακολουθούμενο από το SVM (49%), ενώ το KNN παρουσίασε τη χαμηλότερη επίδοση (45%). Η αλληλοεπικάλυψη χαρακτηριστικών μεταξύ των κατηγοριών μείωσε την ικανότητα των μοντέλων να διαχωρίσουν αποτελεσματικά τις εικόνες.

Συνολικά, το Random Forest αναδείχθηκε ως το πιο αποδοτικό μοντέλο σε όλα τα πειράματα, λόγω της ευελιξίας του στην αναγνώριση χαρακτηριστικών και της ικανότητάς του να αντιμετωπίζει υψηλές διαστάσεις δεδομένων. Το SVM ήταν ανταγωνιστικό, αλλά με περιορισμένη δυνατότητα κλιμάκωσης καθώς ο αριθμός των κατηγοριών αυξανόταν. Το KNN παρουσίασε τη χαμηλότερη απόδοση, ιδιαίτερα σε περιβάλλοντα με μεγαλύτερη πολυπλοκότητα, πιθανότατα λόγω της ευαισθησίας του στον θόρυβο και της υψηλής υπολογιστικής πολυπλοκότητας.

Επίδραση της Προεπεξεργασίας

Η χρήση Principal Component Analysis (PCA) συνέβαλε στη βελτίωση της απόδοσης των μοντέλων, καθώς μείωσε τη διάσταση των δεδομένων διατηρώντας το 95% της διακύμανσης. Αυτό αποδείχθηκε κρίσιμο για το SVM και το RF, καθώς επέτρεψε ταχύτερη εκπαίδευση και μείωση της υπερπροσαρμογής.

Η εφαρμογή Min-Max Scaling διασφάλισε την ομοιομορφία στις τιμές των pixel, συμβάλλοντας στην σταθερότητα των αλγορίθμων.

Έλεγχος Υπερπροσαρμογής

Για την ανίχνευση υπερπροσαρμογής, υπολογίσαμε την απόδοση των μοντέλων σε ένα τυχαίο 35% υποσύνολο των δεδομένων εκπαίδευσης. Παρατηρήθηκε ότι το Random Forest εμφάνισε την υψηλότερη απόκλιση μεταξύ των επιδόσεων εκπαίδευσης και δοκιμής, υποδηλώνοντας μια πιθανή τάση υπερπροσαρμογής. Το SVM διατήρησε μια πιο σταθερή επίδοση, ενώ το KNN εμφάνισε χαμηλή γενίκευση σε όλα τα υποσύνολα δεδομένων.

Προτάσεις για Μελλοντική Έρευνα

Για μελλοντική έρευνα, όσον αφορά το σύνολο δεδομένων με τις δημοπρασίες του eBay, προτείνεται η εξέταση νέων αλγορίθμων, όπως τα Gradient Boosting Machines (GBM) και τα Deep Neural Networks (DNN), για την καλύτερη ανάλυση μη γραμμικών σχέσεων μεταξύ των χαρακτηριστικών των δημοπρασιών. Επιπλέον, η ενσωμάτωση νέων χαρακτηριστικών, όπως παράγοντες που σχετίζονται με τη δημοτικότητα του αντικειμένου ή τη χρονική περίοδο της δημοπρασίας, θα μπορούσε να βελτιώσει την απόδοση των μοντέλων.

Η εφαρμογή αυτών των προτάσεων θα μπορούσε να οδηγήσει σε μεγαλύτερη ακριβή πρόβλεψη και καλύτερη αναγνώριση εικόνας σε πραγματικούς χρόνους ή σε εμπορικές εφαρμογές.

Για την κατηγοριοποίηση εικόνας στο CIFAR-10, θα ήταν χρήσιμο να διερευνηθούν τα εξής:

- Χρήση Convolutional Neural Networks (CNNs): Τα CNNs έχουν αποδειχθεί ιδιαίτερα αποδοτικά σε προβλήματα αναγνώρισης εικόνας. Η εφαρμογή τους μπορεί να βελτιώσει σημαντικά την απόδοση, καθώς μπορούν να μάθουν πολύπλοκα πρότυπα στις εικόνες που δεν μπορούν να ανιχνευθούν από τους παραδοσιακούς ταξινομητές.
- Hyperparameter Tuning: Η αναζήτηση υπερπαραμέτρων μπορεί να βοηθήσει στη βελτιστοποίηση των αλγορίθμων Random Forest και SVM, προσαρμόζοντας τα βάθη δέντρων, αριθμούς γειτόνων (KNN), τιμές C (SVM) και πυρήνες.

Συμπερασματικά, ενώ τα κλασικά μοντέλα όπως Random Forest και SVM απέδωσαν ικανοποιητικά, οι σύγχρονες προσεγγίσεις βαθιάς μάθησης είναι αναγκαίες για τη βελτίωση της ακρίβειας και της δυνατότητας γενίκευσης των μοντέλων.

6. Βιβλιογραφία

CIFAR-10 Image Classification:

[1] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Master's Thesis, University of Toronto, 2009. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[2] A. Krizhevsky, "CIFAR-10 Dataset," 2009. Available: https://www.cs.toronto.edu/~kriz/cifar.html

[3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014. Available: https://arxiv.org/abs/1409.1556

[4] C. Szegedy et al., "Going Deeper with Convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Szegedy_Going_Deeper_With_2015_CVPR_paper.pdf

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

eBay Data Analysis:

[6] J. Grossman, "Predicting eBay Auction Sales with Machine Learning," 2013. Available: http://jaygrossman.com/post/2013/06/10/Predicting-eBay-Auction-Sales-with-Machine-Learning.aspx

[7] S. Xu, "Study of Auction Theory in eBay Data," 2011. Available: https://www.stat.berkeley.edu/~aldous/Research/Ugrad/selene_xu.pdf

[8] I. Raykhel and D. Ventura, "Real-time Automatic Price Prediction for eBay Online Trading," in Proceedings of the 19th Conference on Innovative Applications of Artificial Intelligence, 2007. Available: https://www.aaai.org/Papers/IAAI/2007/IAAI07-019.pdf