Υπολογιστική Νοημοσύνη - Στατιστική μάθηση
Πρώτη Εργασία

Κωστινούδης Ευάγγελος
ΑΕΜ: 112

30 Νοεμβρίου 2022

1 Περιγραφή προβλήματος που επιλέχτηκε

Για την εργασία αυτή επιλέχτηκε το πρόβλημα του διαχωρισμού κλάσεων και τα δεδομένα προέρχονται από τις βάσεις:

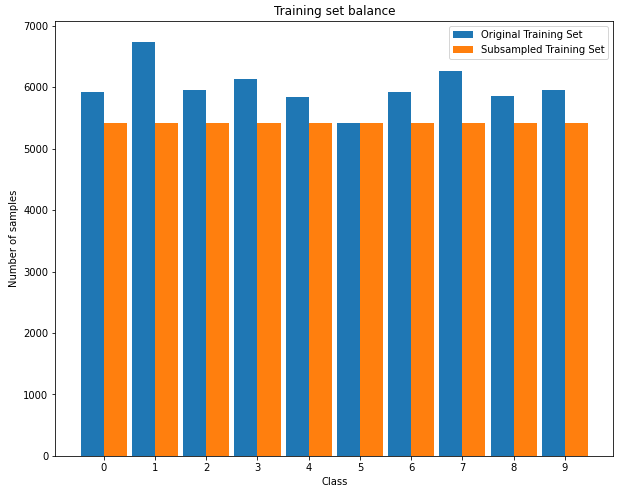
1. MNIST
2. Cifar-10

2 Υλοποίηση

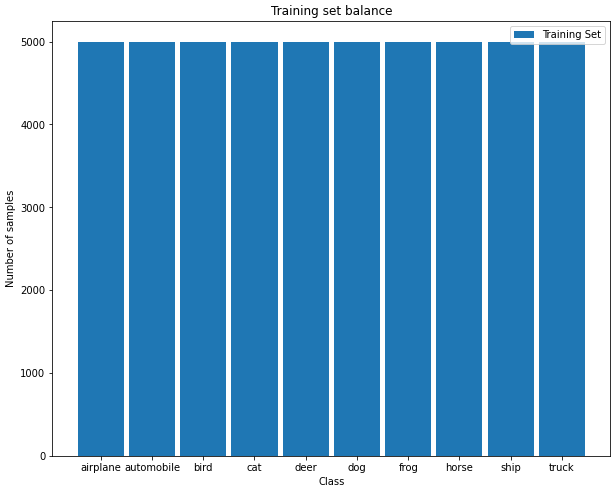
Για την εκπαίδευση των μοντέλων χρησιμοποιούνται τα δεδομένα εκπαίδευσης που δίνονται από τις δύο βάσεις που αναφέρονται παραπάνω. Αντίστοιχα χρησιμοποιούνται τα δεδομένα ελέγχουν για τον έλεγχο των μοντέλων.

2.1 Επιλογή δειγμάτων

Από το Σχήμα 1 παρατηρείται ότι στο σύνολο εκπαίδευσης κάθε κλάση έχει τον ίδιο αριθμό δειγμάτων για την βάση Cifar-10, ενώ το αντίστοιχο δεν ισχύει για την βάση MNIST. Γι' αυτό τον λόγο, υποδειγματοληπτείται το σύνολο εκπαίδευσης της βάσης MSIST, ούτως ώστε όλες οι κλάσεις να έχουν τον ίδιο αριθμό δειγμάτων όπως φαίνεται στο Σχήμα 1I. Το πρόβλημα αυτό θα μπορούσε να λυθεί με πολλές μεθόδους όπως π.χ. να χρησιμοποιηθούν βάρη για κάθε κλάση. Ο λόγος που επιλέχτηκε η υποδειγματοληψία είναι ο μεγάλος όγκος δειγμάτων που οδηγεί σε μεγάλο χρόνο εκτέλεσης των πειραμάτων.



(I) Ιστόγραμμα κλάσεων πριν και μετά την υποδειγματοληψία για τη βάση MNIST



(II) Ιστόγραμμα κλάσεων για τη βάση Cifar-10

Σχήμα 1: Ιστογράμματα κλάσεων για τις βάσεις δεδομένων

2.2 Προεπεξεργασία δεδομένων

Σε αυτό το στάδιο εφαρμόζεται η παρακάτω διαδικασία και για τις δύο βάσεις:

1. Μετασχηματισμός των δεδομένων στο διάστημα $[0, 1]$ για κάθε χαρακτηριστικό των δεδομένων.
2. Εφαρμογή της μεθόδου PCA, κρατώντας τουλάχιστον 90% της πληροφορίας.
3. Μετασχηματισμός των δεδομένων στο διάστημα $[0, 1]$ για κάθε χαρακτηριστικό των δεδομένων ξανά.

Με τη χρήση της μεθόδου PCA μειώνεται κατά μεγάλο βαθμό ο αριθμό των χαρακτηριστικών των δεδομένων.

Ο πρώτος μετασχηματισμός είναι περιττός αφού τα δεδομένα είναι 8-bit εικόνες, που σημαίνει ότι όλα τα χαρακτηριστικά βρίσκονται στο ίδιο διάστημα ($[0 - 255]$).

2.3 Επιλογή παραμέτρων

Τα μοντέλα που επιλέχτηκαν είναι:

1. Γραμμικό SVM
2. SVM με πολυωνυμικό πυρήνα
3. SVM με RBF πυρήνα
4. SVM με σιγμοειδή πυρήνα

Επειδή τα δεδομένα έχουν περισσότερες από δύο κλάσεις (έχουν δέκα), χρησιμοποιήθηκε η μέθοδος 1vs1 όπου εκπαιδεύονται $\frac{n(n-1)}{2}$ δυαδικοί ταξινομητές, όπου n ο αριθμός των κλάσεων (δηλαδή 45 δυαδικοί ταξινομητές για κάθε μοντέλο). Με αυτό τον τρόπο δημιουργούνται όλοι οι δυνατοί συνδυασμοί δυαδικών ταξινομητών. Οι ταξινομητές αυτοί "ψηφίζουν" ανάμεσα σε δύο κλάσεις και η τελική κλάση είναι αυτή με τις περισσότερες ψήφους.

Επιπλέον, χρησιμοποιήθηκαν οι μέθοδοι πλησιέστερων γειτόνων (Nearest Neighbors) και πλησιέστερου κέντρου κλάσης (Nearest Class Centroid), ώστε να συγκριθούν με τα μοντέλα SVM.

Για κάθε μοντέλο υλοποιήθηκε η μέθοδος αναζήτησης πλέγματος (grid search) για την εύρεση των καλύτερων παραμέτρων εκτός του μοντέλου πλησιέστερου κέντρου κλάσης, όπου χρησιμοποιήθηκε η ευκλείδεια απόσταση.

Συγκεκριμένα για τη βάση MNIST χρησιμοποιήθηκε 3-fold cross validation για τις παραμέτρους:

1. Γραμμικό SVM: $C : (0.1, 1, 10)$
2. SVM με πολυωνυμικό πυρήνα: $C : (0.1, 1, 10), d : (2, 3, 4), \gamma : (0.1, 1, 10)$
3. SVM με RBF πυρήνα: $C : (1, 10, 50), \gamma : (0.01, 0.1, 1, 10, 100)$
4. SVM με σιγμοειδή πυρήνα: $C : (10, 100, 1000, 10000), \gamma : (0.0001, 0.001, 0.01, 0.1)$

Για τη βάση Cifar-10 χρησιμοποιήθηκε 2-fold cross validation για τις παραμέτρους:

1. Γραμμικό SVM: $C : (0.1, 1, 10)$
2. SVM με πολυωνυμικό πυρήνα: $C : (0.1, 1, 10), d : (2, 3), \gamma : (0.1, 1)$
3. SVM με RBF πυρήνα: $C : (1, 10, 50), \gamma : (0.1, 1, 10)$
4. SVM με σιγμοειδή πυρήνα: $C : (10, 100, 1000), \gamma : (0.0001, 0.001, 0.01)$

Όπου C είναι η παράμετρος του σφάλματος από της εξίσωση:

$$\begin{aligned} &\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_i \xi_i \\ &\text{subject to } y_i(\boldsymbol{w^T}\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \\ &\xi_i \geq 0 \end{aligned}$$

Επίσης οι υπόλοιπες παράμετροι για κάθε πυρήνα είναι:

1. SVM με πολυωνυμικό πυρήνα: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma \boldsymbol{x}_i^T \boldsymbol{x}_j)^d$
2. SVM με RBF πυρήνα: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2)$
3. SVM με σιγμοειδή πυρήνα: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = tanh(\gamma \boldsymbol{x}_i^T \boldsymbol{x}_j)$

Για το μοντέλο πλησιέστερων γειτόνων και για τις δύο βάσεις χρησιμοποιήθηκε 3-fold cross validation για τις παραμέτρους $n\_neighbors : (1, 3, 5, 10, 30, 100, 200), weights : (uniform, distance)$ όπου $n\_neighbors$ ο αριθμός των γειτόνων και $weights$ η χρήση βαρών ($distance$ υπολογισμός βαρών μέσω της ευκλείδειας απόστασης και $uniform$ ομοιόμορφα βάρη).

Η μετρική για την επιλογή των παραμέτρων είναι το macro F1-score δηλαδή ο μέσος όρος των F1-scores για κάθε κλάση.

Ο λόγος που επιλέχτηκαν παραπάνω παράμετροι για τη βάση MNIST αλλά και περισσότερες τιμές στο cross validation είναι ο χρόνος εκτέλεσης των πειραμάτων. Με τις παραμέτρους αυτές ο χρόνος εκτέλεσης των πειραμάτων για τις δύο βάσεις είναι παρόμοιος. Η διαφορά μεταξύ 2-fold και 3-fold είναι πολύ μεγαλύτερη από τα 2/3 του χρόνου εκτέλεσης γιατί τα πειράματα που τρέχουν έχουν και λιγότερα δείγματα και ο αλγόριθμος εκπαίδευσης (QP solver της libsvm) έχει πολυπλοκότητα μεταξύ $O(n_{features} \times n_{samples}^2)$ και $O(n_{features} \times n_{samples}^2)$.

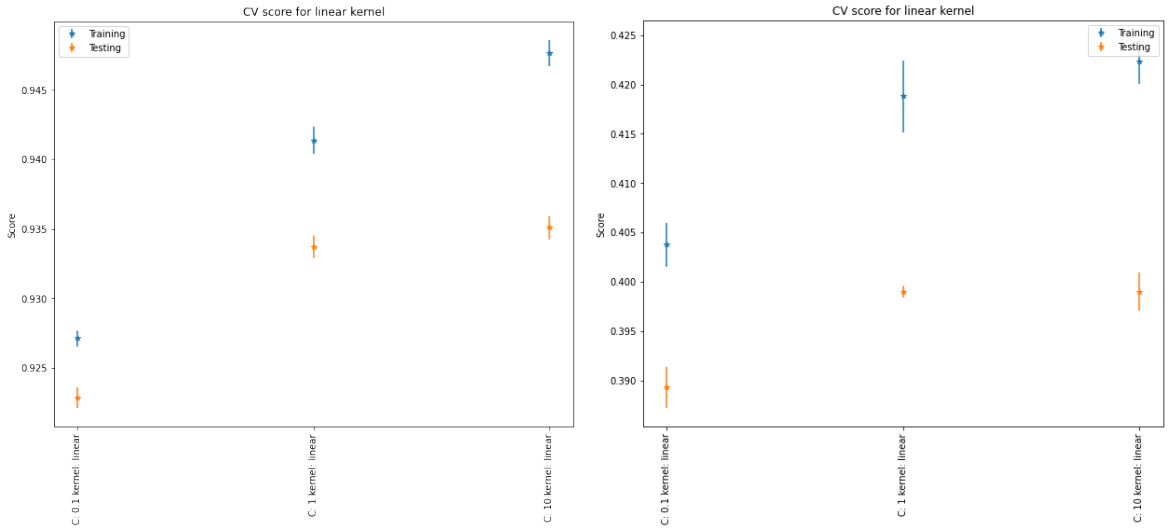
3 Αποτελέσματα

Τα πειράματα εκτελέστηκαν στο περιβάλλον του Google Colab.

3.1 Επιλογή παραμέτρων

3.1.1 Γραμμικό SVM

Από το Σχήμα 2 παρατηρείται ότι για το μοντέλο του γραμμικού SVM η καλύτερη παράμετρος του C για τη βάση MNIST είναι **10** ενώ για την βάση Cifar-10 είναι **1**.



(I) MNIST (II) Cifar-10

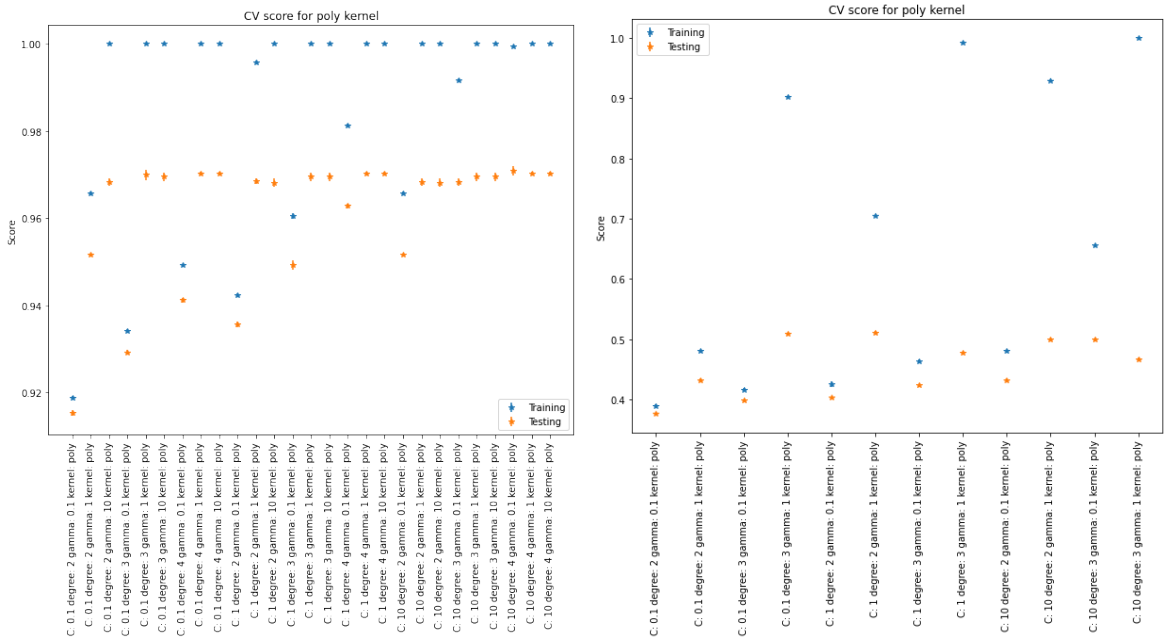
Σχήμα 2: Αποτελέσματα αναζήτησης πλέγματος για το γραμμικό SVM

3.1.2 SVM με πολυωνυμικό πυρήνα

Από το Σχήμα 3 παρατηρείται ότι για το μοντέλο του SVM με πολυωνυμικό πυρήνα οι καλύτερες παράμετροι για κάθε βάση δίνονται στο Πίνακας 1.

	MNIST	Cifar-10
C	10	1
γ	0.1	1
d	4	2

Πίνακας 1: Καλύτεροι παράμετροι SVM με πολυωνυμικό πυρήνα



(I) MNIST (II) Cifar-10

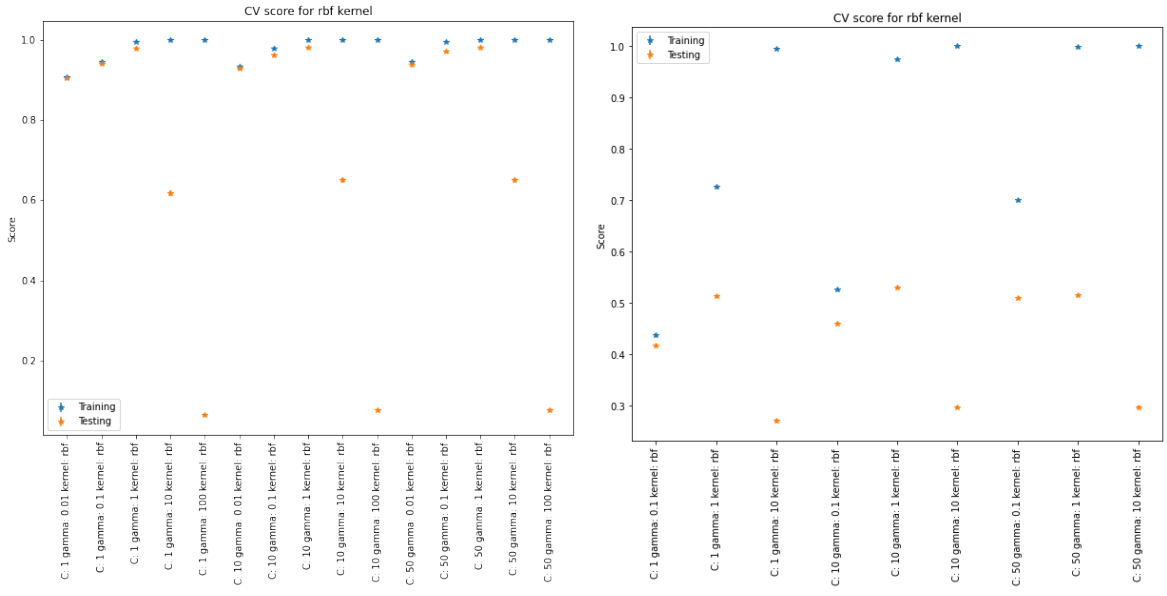
Σχήμα 3: Αποτελέσματα αναζήτησης πλέγματος για το SVM με πολυωνυμικό πυρήνα

3.1.3 SVM με RBF πυρήνα

Από το Σχήμα 4 παρατηρείται ότι για το μοντέλο του SVM με RBF πυρήνα οι καλύτερες παράμετροι για κάθε βάση δίνονται στο Πίνακας 2.

	MNIST	Cifar-10
C	10	10
γ	1	1

Πίνακας 2: Καλύτεροι παράμετροι SVM με RBF πυρήνα



(I) MNIST (II) Cifar-10

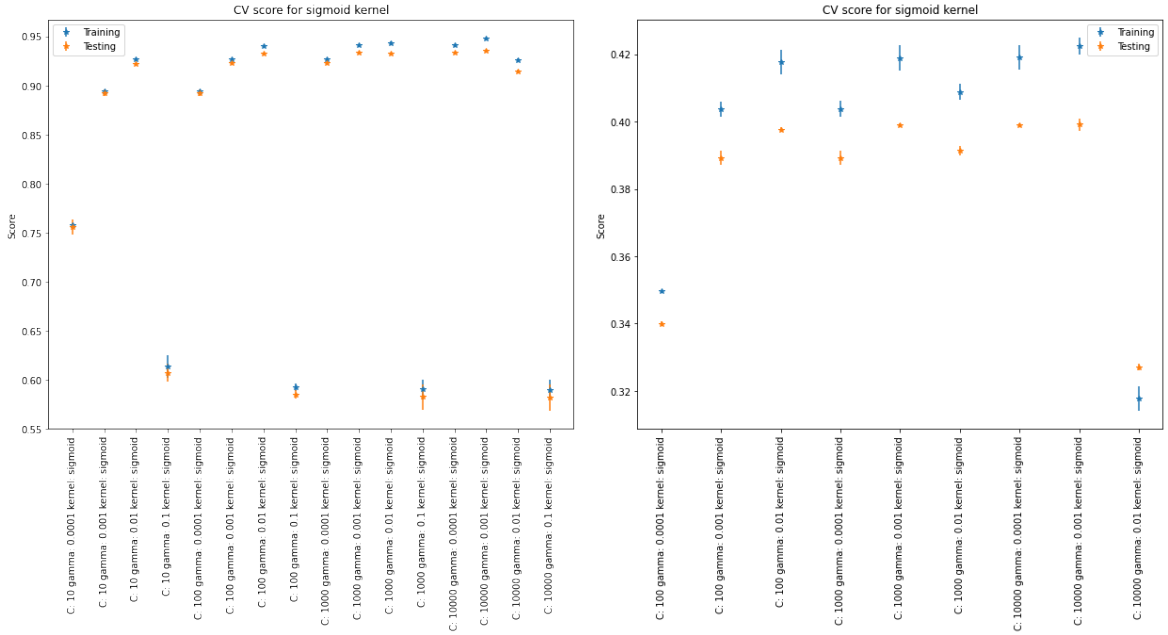
Σχήμα 4: Αποτελέσματα αναζήτησης πλέγματος για το SVM με RBF πυρήνα

3.1.4 SVM με σιγμοειδή πυρήνα

Από το Σχήμα 5 παρατηρείται ότι για το μοντέλο του SVM με σιγμοειδή πυρήνα οι καλύτερες παράμετροι για κάθε βάση δίνονται στο Πίνακας 3.

	MNIST	Cifar-10
C	10000	10000
γ	0.001	0.001

Πίνακας 3: Καλύτεροι παράμετροι SVM με σιγμοειδή πυρήνα



(I) MNIST (II) Cifar-10

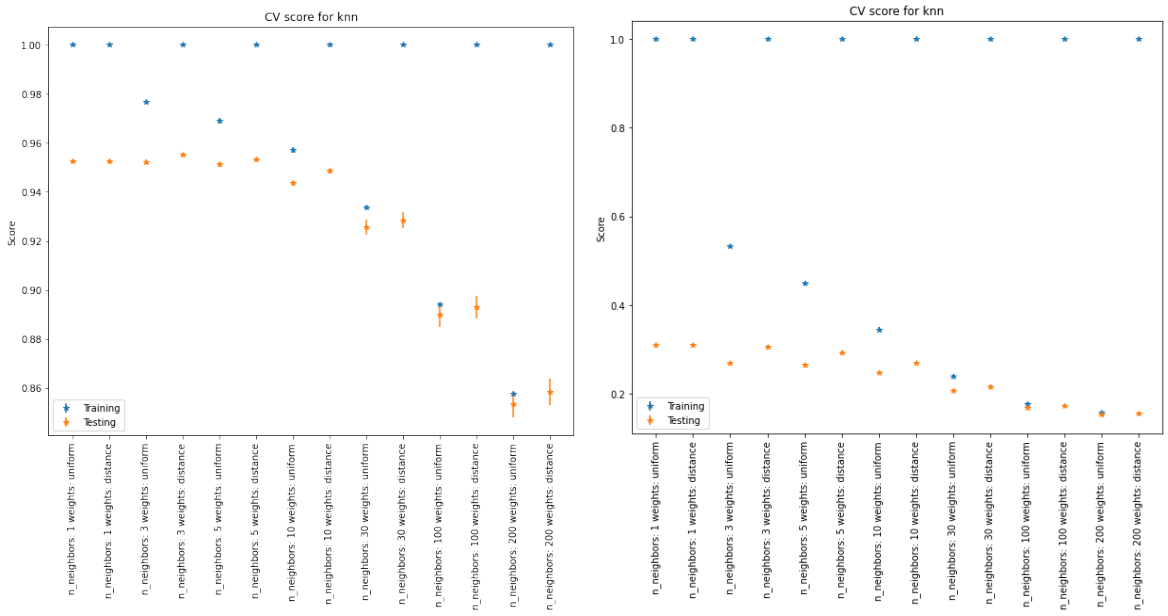
Σχήμα 5: Αποτελέσματα αναζήτησης πλέγματος για το SVM με σιγμοειδή πυρήνα

3.1.5 πλησιέστερων γειτόνων

Από το Σχήμα 6 παρατηρείται ότι για το μοντέλο των πλησιέστερων γειτόνων οι καλύτερες παράμετροι για κάθε βάση δίνονται στο Πίνακας 4.

	MNIST	Cifar-10
$n\_neighbors$	3	1
$weights$	distance	uniform

Πίνακας 4: Καλύτεροι παράμετροι πλησιέστερων γειτόνων



(I) MNIST (II) Cifar-10

Σχήμα 6: Αποτελέσματα αναζήτησης πλέγματος για το μοντέλο πλησιέστερων γειτόνων

3.2 Απόδοση μοντέλων

Αφού έχουν επιλεγεί οι καλύτερες παράμετροι για κάθε μοντέλο και έχουν εκπαιδευτεί όλα τα μοντέλα στο σύνολο εκπαίδευσης για κάθε βάση, θα συγκριθεί η απόδοση τους.

3.2.1 MNIST

Στο Σχήμα 7 και στους Πίνακας 5 και Πίνακας 6 φαίνονται τα αποτελέσματα των μετρικών για όλα τα μοντέλα στο σύνολο εκπαίδευσης και ελέγχου για τις καλύτερες παραμέτρους στη βάση MNIST.

Το μοντέλο του SVM με RBF πυρήνα έχει το καλύτερο αποτέλεσμα για όλες τις μετρικές στο σύνολο ελέγχου. Το μοντέλο των πλησιέστερων γειτόνων έχει καλύτερα αποτελέσματα από αυτά των SVM με γραμμικό και σιγμοειδή πυρήνα και το μοντέλο του πλησιέστερου κέντρου κλάσης έχει τα χειρότερα αποτελέσματα.

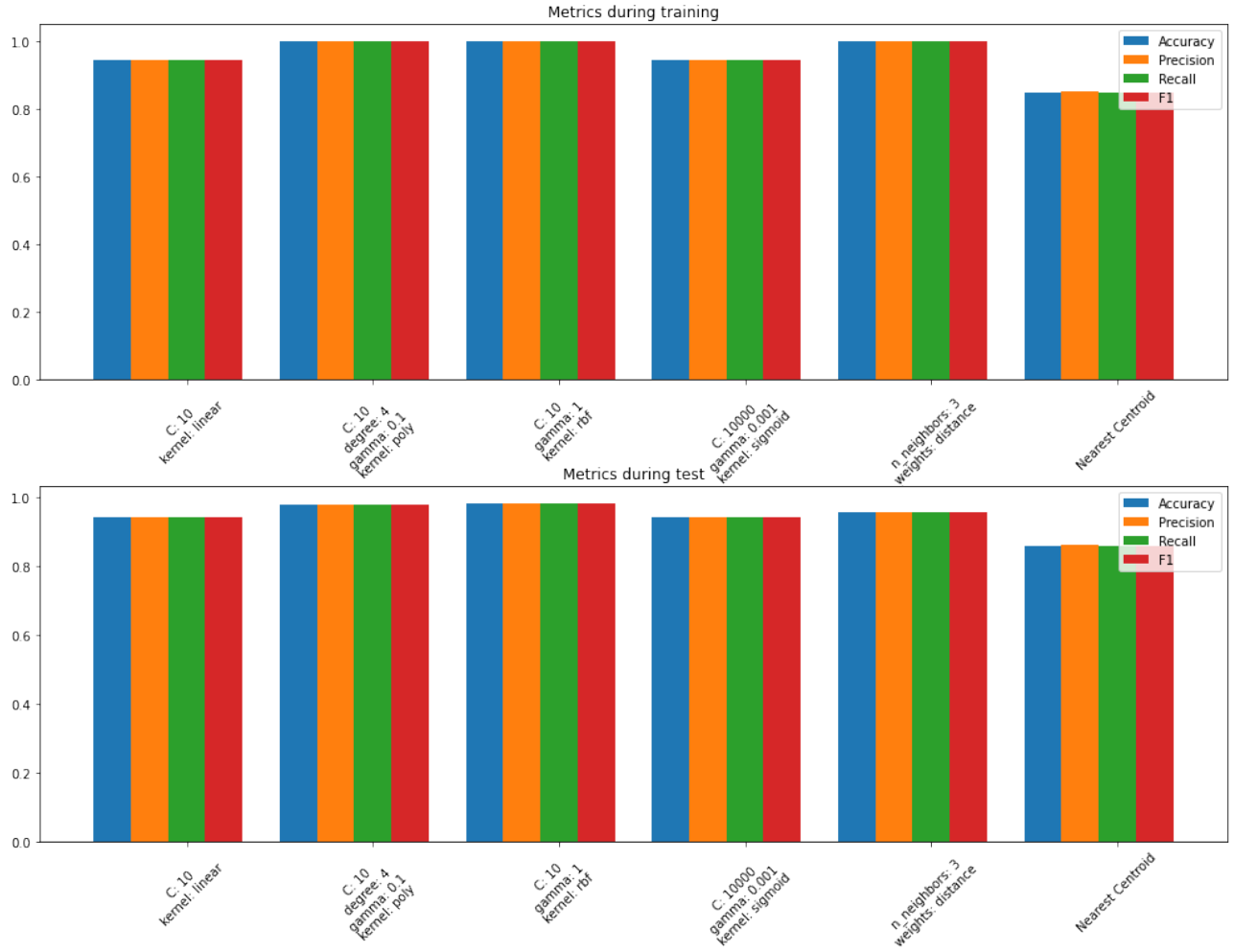
Από το Σχήμα 8 και τον Πίνακας 5 φαίνεται ο χρόνος εκπαίδευσης των μοντέλων. Το μοντέλο του SVM με RBF πυρήνα που έχει τα καλύτερα αποτελέσματα έχει το μεγαλύτερο χρόνο εκπαίδευσης. Επίσης, τα μοντέλα των πλησιέστερων γειτόνων και του πλησιέστερου κέντρου κλάσης έχουν πολύ πιο μικρό χρόνο εκπαίδευσης σε σχέση με τα μοντέλα SVM.

Model	Accuracy	Precision	Recall	F1	Training Time (seconds)
C: 10 kernel: linear	0,9461	0,946	0,9461	0,946	57,5676
C: 10 degree: 4 gamma: 0.1 kernel: poly	0,9993	0,9993	0,9993	0,9993	54,2103
C: 10 gamma: 1 kernel: rbf	1	1	1	1	125,2382
C: 10000 gamma: 0.001 kernel: sigmoid	0,946	0,9459	0,946	0,9459	63,1962
n_neighbors: 3 weights: distance	1	1	1	1	0,0161
Nearest Centroid	0,8495	0,852	0,8495	0,8496	0,0508

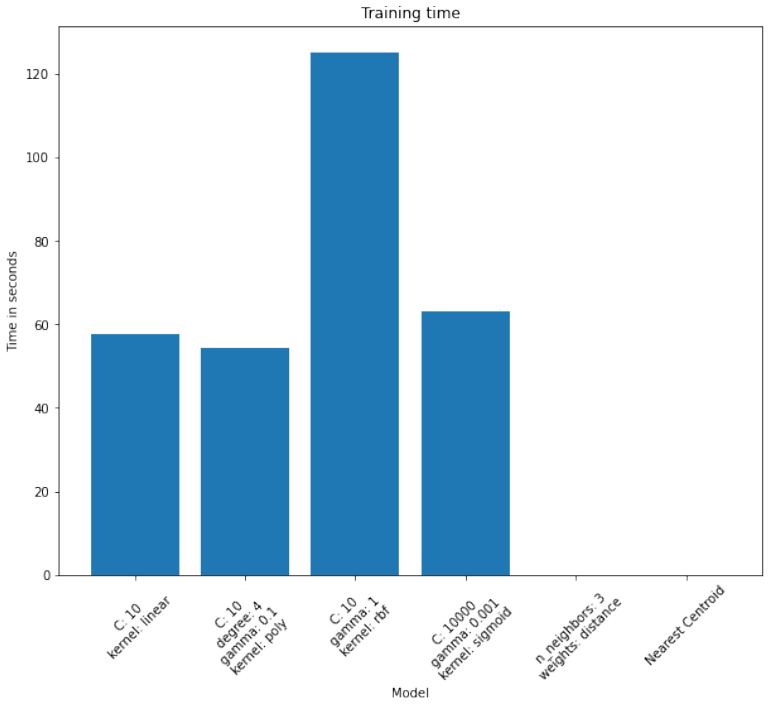
Πίνακας 5: Μετρικές αποτελεσμάτων στο σύνολο εκπαίδευσης και χρόνος εκπαίδευσης για τη βάση MNIST

Model	Accuracy	Precision	Recall	F1
C: 10 kernel: linear	0,9436	0,9429	0,9428	0,9427
C: 10 degree: 4 gamma: 0.1 kernel: poly	0,9777	0,9776	0,9775	0,9776
C: 10 gamma: 1 kernel: rbf	0,9841	0,984	0,9841	0,984
C: 10000 gamma: 0.001 kernel: sigmoid	0,9437	0,943	0,9429	0,9428
n_neighbors: 3 weights: distance	0,9579	0,9584	0,9575	0,9576
Nearest Centroid	0,8606	0,8614	0,859	0,8594

Πίνακας 6: Μετρικές αποτελεσμάτων στο σύνολο ελέγχο για τη βάση MNIST



Σχήμα 7: Μετρικές για τη βάση MNIST



Σχήμα 8: Χρόνος εκπαίδευσης για τη βάση MNIST

3.2.2 Cifar-10

Στο Σχήμα 9 και στους Πίνακας 7 και Πίνακας 8 φαίνονται τα αποτελέσματα των μετρικών για όλα τα μοντέλα στο σύνολο εκπαίδευσης και ελέγχου για τις καλύτερες παραμέτρους στη βάση Cifar-10.

Το μοντέλο του SVM με RBF πυρήνα έχει το καλύτερο αποτέλεσμα για όλες τις μετρικές στο σύνολο ελέγχου όπως και στη βάση MNIST. Ακόμα, τα SVM μοντέλα έχουν καλύτερα αποτελέσματα σε σχέση με τα μοντέλα των πλησιέστερων γειτόνων και του πλησιέστερου κέντρου κλάσης, εκτός από τη μετρική *precision* για το μοντέλο των κοντινότερων γειτόνων.

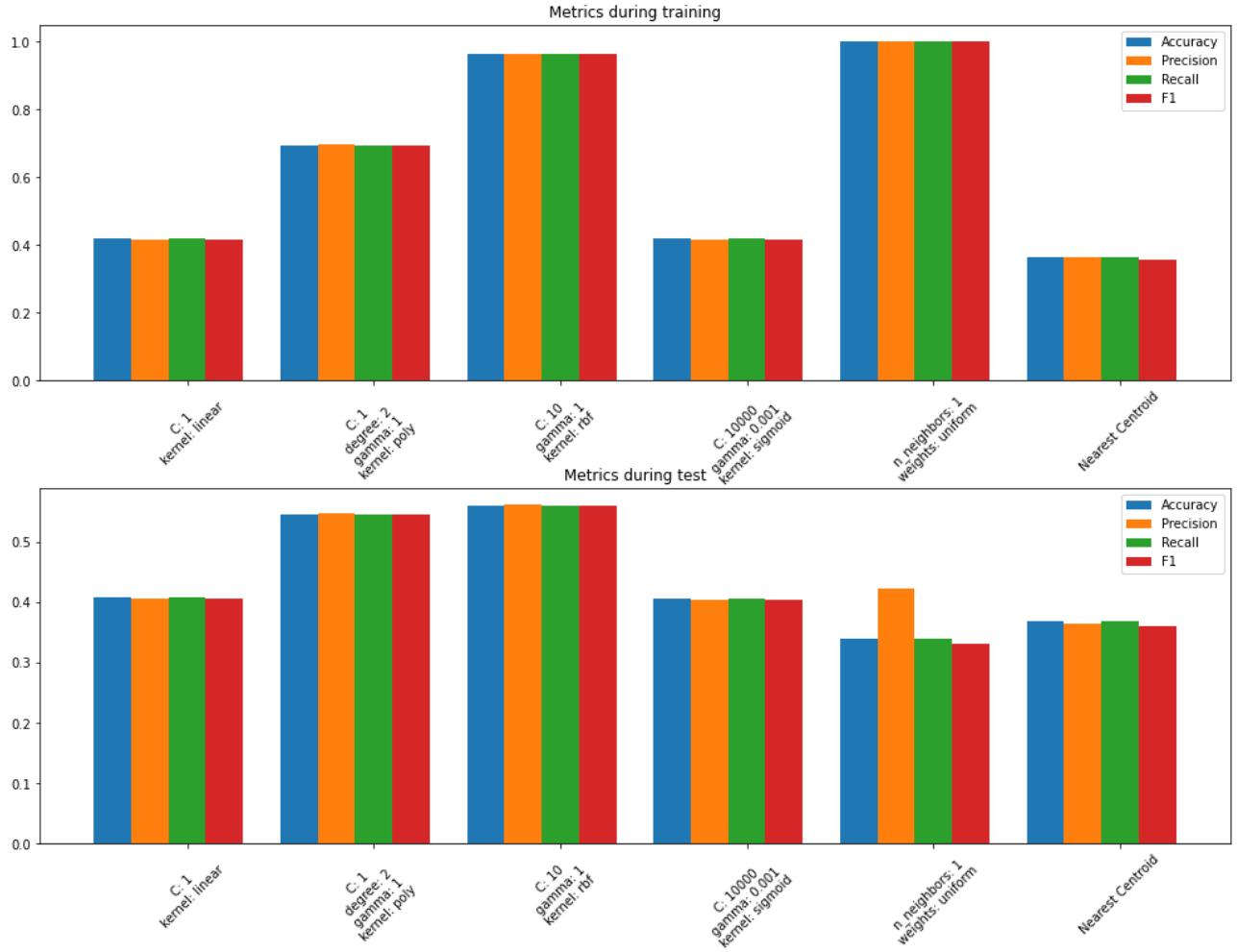
Από το Σχήμα 10 και τον Πίνακας 7 φαίνεται ο χρόνος εκπαίδευσης των μοντέλων. Το μοντέλο του SVM με πολυωνυμικό πυρήνα έχει το μεγαλύτερο χρόνο εκτέλεσης. Επίσης, όπως και στη βάση MNIST τα μοντέλα των πλησιέστερων γειτόνων και του πλησιέστερου κέντρου κλάσης έχουν πολύ πιο μικρό χρόνο εκπαίδευσης σε σχέση με τα μοντέλα SVM.

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**Training Time (seconds)**
C: 1 kernel: linear	0,4189	0,416	0,4189	0,416	410,7339
C: 1 degree: 2 gamma: 1 kernel: poly	0,6943	0,6989	0,6943	0,695	823,6295
C: 10 gamma: 1 kernel: rbf	0,9631	0,964	0,9631	0,9634	569,9598
C: 10000 gamma: 0.001 kernel: sigmoid	0,4204	0,4175	0,4204	0,4175	603,1902
n_neighbors: 1 weights: uniform	1	1	1	1	0,0092
Nearest Centroid	0,3661	0,3626	0,3661	0,3575	0,0291

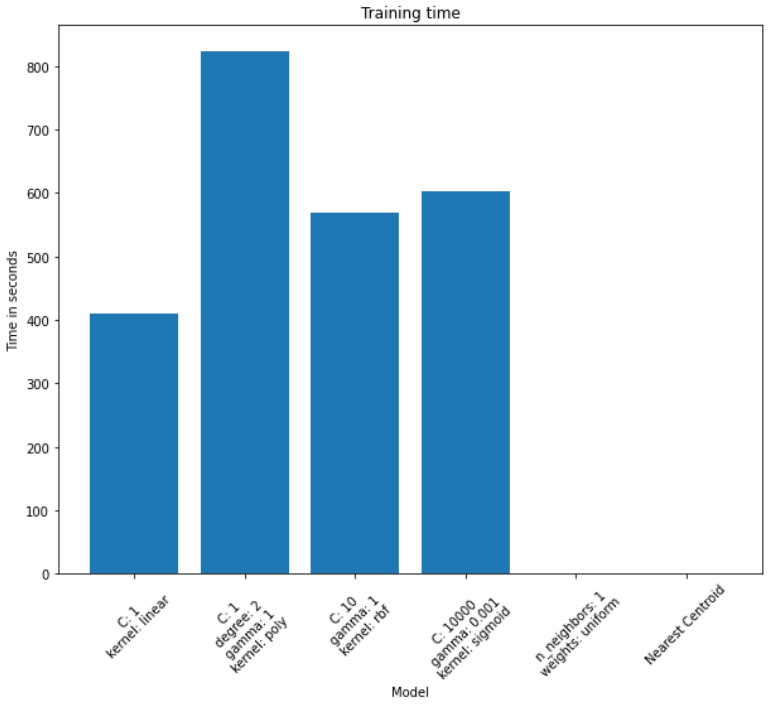
Πίνακας 7: Μετρικές αποτελεσμάτων στο σύνολο εκπαίδευσης και χρόνος εκπαίδευσης για τη βάση Cifar-10.

Model	Accuracy	Precision	Recall	F1
C: 1 kernel: linear	0,4092	0,4067	0,4092	0,4067
C: 1 degree: 2 gamma: 1 kernel: poly	0,5463	0,5485	0,5463	0,5457
C: 10 gamma: 1 kernel: rbf	0,56	0,5626	0,56	0,5607
C: 10000 gamma: 0.001 kernel: sigmoid	0,4074	0,4046	0,4074	0,4048
n_neighbors: 1 weights: uniform	0,3394	0,4238	0,3394	0,3307
Nearest Centroid	0,369	0,3647	0,369	0,3601

Πίνακας 8: Μετρικές αποτελεσμάτων στο σύνολο ελέγχο για τη βάση Cifar-10



Σχήμα 9: Μετρικές για τη βάση Cifar-10



Σχήμα 10: Χρόνος εκπαίδευσης για τη βάση Cifar-10

3.3 Απόδοση καλύτερων μοντέλων

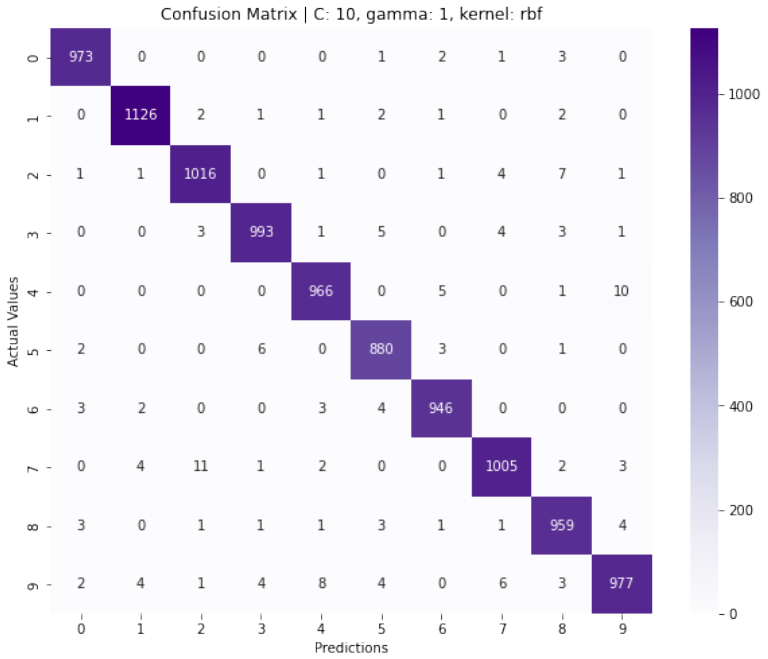
3.3.1 MNIST

Το καλύτερο μοντέλο για τη βάση MNIST είναι το SVM με RBF πυρήνα. Στο Σχήμα 11 φαίνεται το confusion matrix γι' αυτό το μοντέλο. Από το σχήμα αυτό παρατηρείται ότι οι περισσότερες εσφαλμένες ταξινομήσεις γίνονται όταν:

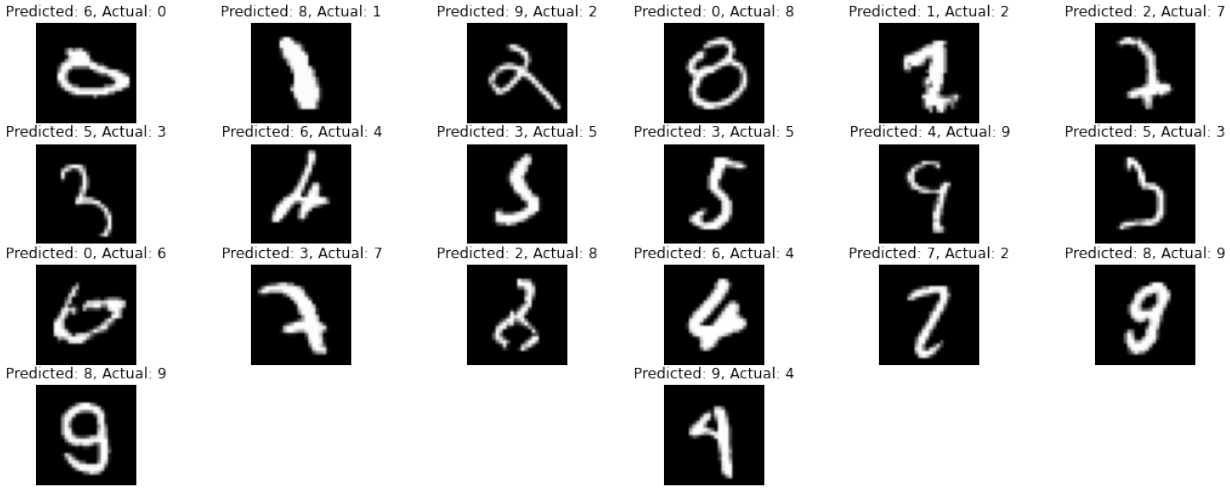
- Η πραγματική κλάση είναι **7** ενώ το μοντέλο την ταξινομεί στη **2**.
- Η πραγματική κλάση είναι **4** ενώ το μοντέλο την ταξινομεί στη **9**.
- Η πραγματική κλάση είναι **9** ενώ το μοντέλο την ταξινομεί στη **4**.
- Η πραγματική κλάση είναι **2** ενώ το μοντέλο την ταξινομεί στη **8**.

Το μοντέλο αυτό "μπερδεύει" κλάσεις που έχουν παρόμοια μορφολογία όπως αυτή του **7** και του **5** και του **4** και του **9**.

Από το Σχήμα 12 φαίνονται μερικές από τις λάθος ταξινομήσεις του μοντέλου. Από το σχήμα αυτό επιβεβαιώνεται η προηγούμενη παρατήρηση και επίσης παρατηρείται ότι μερικές εικόνες δεν είναι τόσο εύκολο να ταξινομηθούν ακόμα και από έναν άνθρωπο.



Σχήμα 11: Confusion matrix για τη βάση MNIST



Σχήμα 12: Λάθος ταξινομήσεις του καλύτερου μοντέλου για τη βάση MNIST

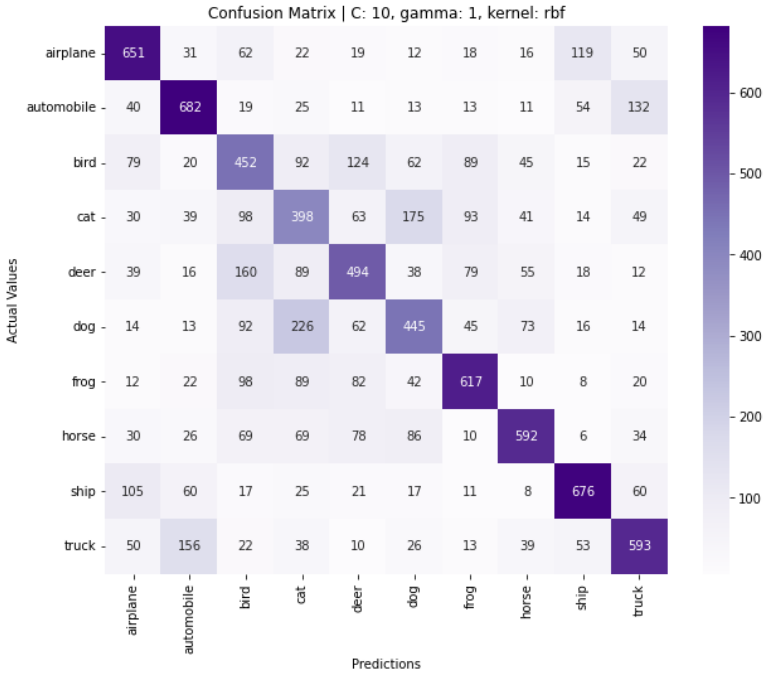
3.3.2 Cifar-10

Το καλύτερο μοντέλο για τη βάση Cifar-10 είναι το SVM με RBF πυρήνα. Στο Σχήμα 13 φαίνεται το confusion matrix γι' αυτό το μοντέλο. Από το σχήμα αυτό παρατηρείται ότι οι περισσότερες εσφαλμένες ταξινομήσεις γίνονται όταν:

- Η πραγματική κλάση είναι **σκύλος** ενώ το μοντέλο την ταξινομεί στη **γάτα**.
- Η πραγματική κλάση είναι **γάτα** ενώ το μοντέλο την ταξινομεί στη **σκύλος**.
- Η πραγματική κλάση είναι **ελάφι** ενώ το μοντέλο την ταξινομεί στη **πουλί**.
- Η πραγματική κλάση είναι **φορτηγό** ενώ το μοντέλο την ταξινομεί στη **αυτοκινητιστικό** (automobile).

Παρατηρείται ότι το μοντέλο πραγματοποιεί λάθος ταξινομήσεις όταν οι κλάσεις έχουν παρόμοια χαρακτηριστικά όπως π.χ. η κλάση του **σκύλος** και της **γάτας** και του **φορτηγού** και του **αυτοκινητιστικό**. Επίσης λάθη παρατηρούνται όταν το περιβάλλον της εικόνας μπορεί να είναι παρόμοιο όπως π.χ. η κλάση του **πλοίου** και του **αεροπλάνου** (έχουν μπλε παρασκήνιο). Τέλος, υπάρχουν λάθη που δεν είναι τόσο εξηγήσιμα όπως π.χ. η κλάση του **ελαφιού** και του **πουλιού**.

Στο Σχήμα 14 παρουσιάζονται μερικές λάθος ταξινομήσεις του μοντέλου.



Σχήμα 13: Confusion matrix για τη βάση Cifar-10

Σχήμα 14: Λάθος ταξινομήσεις του καλύτερου μοντέλου για τη βάση Cifar-10