ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΑΛΕΞΑΚΗ ΒΑΣΙΛΙΚΗ Α.Μ :1097464

ΕΡΓΑΣΙΑ 2: Φίλτρο μηνυμάτων EMAIL

Σκοπός της εργασίας είναι η ανάπτυξη και αξιολόγηση συστημάτων ταξινόμησης emails στις κατηγορίες **spam** και **non-spam**, χρησιμοποιώντας στατιστικές μεθόδους μηχανικής μάθησης.

Το σύνολο δεδομένων **emails.csv** διαχωρίστηκε σε σύνολο εκπαίδευσης, επικύρωσης και ελέγχου. Εφαρμόστηκαν διάφορες τεχνικές αναπαράστασης κειμένου, όπως TF-IDF και embeddings και διάφοροι ταξινομητές όπως Naive Bayes, k-NN, SVM και Logistic Regression.

Η απόδοση των μοντέλων αξιολογήθηκε με χρήση μετρικών όπως Precision, Recall, F1-score, confusion matrix και ROC-AUC, με στόχο την σύγκριση για εύρεση βέλτιστης λύσης.

ΕΡΩΤΗΜΑ 1&2 : Καθαρισμός κειμένου & Naïve Bayes - E2_NaïveBayes.py

Πριν την εκπαίδευση των μοντέλων, εφαρμόστηκε **διαδικασία καθαρισμού κειμένου** η οποία περιλαμβάνει: μετατροπή όλων των χαρακτήρων σε πεζά, αφαίρεση URLs, και emails, αφαίρεση αριθμών και σημείων στίξης και περιττών κενών χαρακτήρων.

Επιπλέον, πριν τον διαχωρισμό των δεδομένων εφαρμόστηκε **τυχαία αναδιάταξη shuffle**, ώστε να διασφαλιστεί η ομοιόμορφη κατανομή των κλάσεων στα σύνολα της εκπαίδευσης, επικύρωσης και ελέγχου. Απαραίτητο βήμα για την σωστή εκπαίδευση των μοντέλων.

Για την αρχική προσέγγιση του προβλήματος χρησιμοποιήθηκε ο ταξινομητής **Naive Bayes**, σε συνδυασμό με αναπαράσταση **TF-IDF** των email μηνυμάτων.

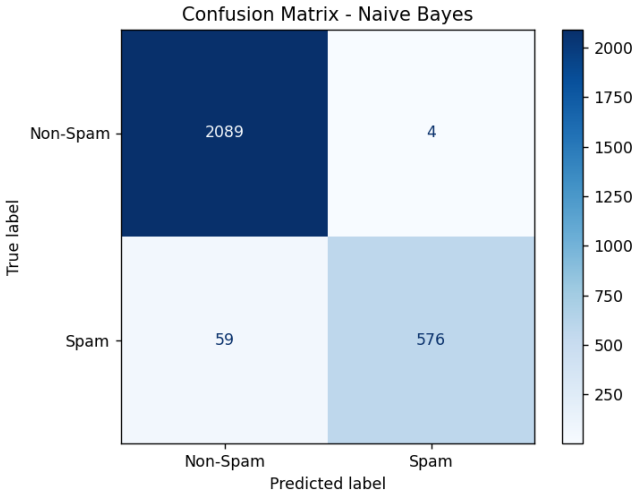
Η αξιολόγηση στο **σύνολο ελέγχου** έδωσε τα ακόλουθα αποτελέσματα:

```
Naive Bayes Test Set
Accuracy : 0.9769061583577713
Precision: 0.993103448275862
Recall   : 0.9070866141732283
F1-score : 0.9481481481481482
AUC      : 0.9985500976257566
```

Τα αποτελέσματα δείχνουν ότι το μοντέλο διακρίνει ικανοποιητικά τα spam και non-spam emails, με υψηλή ακρίβεια και ισορροπία μεταξύ precision και recall.

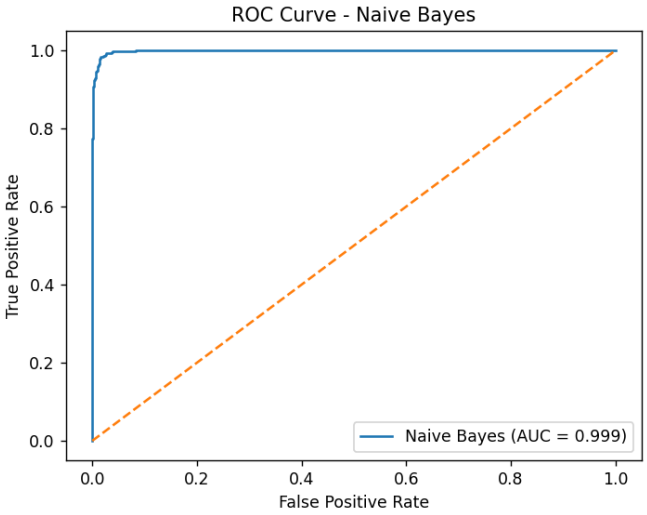
```
Classification Report:
              precision    recall  f1-score   support

    Non-Spam       0.97      1.00      0.99      2093
        Spam       0.99      0.91      0.95       635

    accuracy                           0.98      2728
   macro avg       0.98      0.95      0.97      2728
weighted avg       0.98      0.98      0.98      2728
```

Το μοντέλο είναι ιδιαίτερα αποτελεσματικό στον εντοπισμό spam emails, ενώ παράλληλα διατηρεί καλή απόδοση στη σωστή αναγνώριση των non-spam.



Confusion Matrix

Από τα 2093 non-spam emails του συνόλου ελέγχου, μόνο **4** χαρακτηρίστηκαν λανθασμένα ως spam (false positives), γεγονός που δείχνει ότι το μοντέλο έχει εξαιρετικά χαμηλό ποσοστό λανθασμένων συναγερμών. Αντίθετα, από τα **635** spam emails, τα **576** ανιχνεύθηκαν σωστά, ενώ **59** ταξινομήθηκαν λανθασμένα ως non-spam (false negatives).



ROC Curve

Η καμπύλη ROC του Naive Bayes βρίσκεται πολύ κοντά στο άνω αριστερό άκρο του διαγράμματος, με **AUC = 0.999**, γεγονός που υποδηλώνει σχεδόν ιδανική ταξινόμηση. Αυτό επιβεβαιώνει την πολύ υψηλή διακριτική ικανότητα του μοντέλου ανεξάρτητα από το κατώφλι απόφασης.

ΕΡΩΤΗΜΑ 3 : Sentence transformers και Embeddings - E3_Embeddings.py

Τα καθαρισμένα **email** μηνύματα μετατράπηκαν σε **embeddings** με χρήση του μοντέλου **paraphrase-multilingual-MiniLM-L12-v2** της βιβλιοθήκης **Sentence Transformers**. Κάθε μήνυμα αναπαραστάθηκε ως διάνυσμα **384** διαστάσεων, με συνολικά **2000** δείγματα στο σύνολο εκπαίδευσης, **1000** στο σύνολο επικύρωσης και **2728** στο σύνολο ελέγχου. Τα παραγόμενα embeddings χρησιμοποιήθηκαν ως είσοδος για τους επόμενους ταξινομητές.

```
Embeddings shapes:
Train: (2000, 384) Val: (1000, 384) Test: (2728, 384)
```

Ο υπολογισμός πραγματοποιήθηκε τμηματικά σε **batches**, ώστε να είναι εφικτή η επεξεργασία μεγάλου πλήθους κειμένων χωρίς υπερβολική κατανάλωση μνήμης.

```
Batches: 100%|                                                                                | 63/63 [01:59<00:00,  1.89s/it]
Batches: 100%|                                                                                | 32/32 [00:59<00:00,  1.86s/it]
Batches: 100%|                                                                                | 86/86 [02:41<00:00,  1.88s/it]
```

ΕΡΩΤΗΜΑ 4 : Ταξινόμηση k-NN - E4_kNN.py

Εφαρμόστηκε ο ταξινομητής **k-Nearest Neighbors** χρησιμοποιώντας ως χαρακτηριστικά τα **embeddings** των email μηνυμάτων. Η επιλογή της παραμέτρου **k** πραγματοποιήθηκε πειραματικά με βάση την απόδοση στο σύνολο επικύρωσης, δοκιμάζοντας τιμές από 1 έως 15.

```
  Validation results
k= 1 | F1=0.9260 | AUC=0.9377
k= 3 | F1=0.9301 | AUC=0.9772
k= 5 | F1=0.9175 | AUC=0.9835
k= 7 | F1=0.9175 | AUC=0.9837
k= 9 | F1=0.9073 | AUC=0.9856
k=11 | F1=0.8938 | AUC=0.9875
k=15 | F1=0.8878 | AUC=0.9897
Best k on validation (by F1): 3 (F1=0.9301)
```

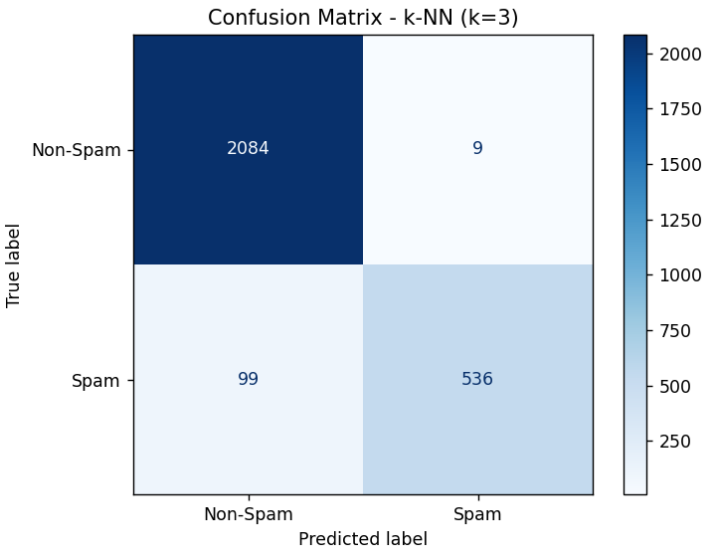
Τα αποτελέσματα στο **validation set** έδειξαν ότι η καλύτερη απόδοση επιτεύχθηκε για **k = 3**, με **F1-score = 0.9301** και **AUC = 0.9772**, τιμές υψηλότερες σε σύγκριση με τις υπόλοιπες επιλογές του k. Για μεγαλύτερες τιμές του k παρατηρείται σταδιακή μείωση του F1-score, άρα και απώλεια διακριτικής ικανότητας λόγω υπερβολικής εξομάλυνσης.

Η διαδικασία υπολογισμού των **embeddings** και της ταξινόμησης ολοκληρώθηκε σε **batches:**

```
Batches: 100%|                                                                                | 63/63 [02:00<00:00,  1.91s/it]
Batches: 100%|                                                                                | 32/32 [00:59<00:00,  1.86s/it]
Batches: 100%|                                                                                | 86/86 [03:25<00:00,  2.39s/it]
```

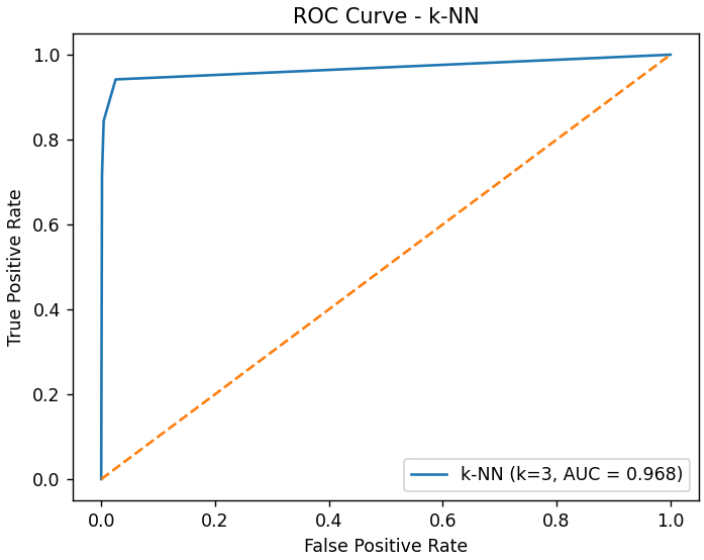
Ο ταξινομητής **k-NN** με **k = 3** αξιολογήθηκε στο **σύνολο ελέγχου**, όπου:

```
 Best k-NN
Precision: 0.9834862385321101
Recall   : 0.8440944881889764
F1       : 0.9084745762711864
AUC      : 0.9678030630786536
```



Confusion Matrix

Ο πίνακας δείχνει ότι από τα **2093** non-spam emails, τα **2084** ταξινομήθηκαν σωστά, ενώ μόνο **9** χαρακτηρίστηκαν λανθασμένα ως spam (false positives). Αντίστοιχα, από τα **635** spam emails, τα **536** ανιχνεύθηκαν σωστά, ενώ **99** ταξινομήθηκαν λανθασμένα ως non-spam (false negatives). Τα αποτελέσματα δείχνουν ότι ο **k-NN** παρουσιάζει καλή ακρίβεια, αλλά χαμηλότερη ανάκληση στην κατηγορία spam.



Η καμπύλη **ROC** του ταξινομητή **3-NN** παρουσιάζει καλή διακριτική ικανότητα, με τιμή **AUC = 0.968**. Η καμπύλη βρίσκεται σημαντικά πάνω από τη διαγώνιο τυχαίας ταξινόμησης, δείχνοντας ότι το μοντέλο μπορεί να διαχωρίσει αποτελεσματικά τα spam από τα non-spam emails.

ΕΡΩΤΗΜΑ 5 : Ταξινόμηση SVM - E5_SVM.py

Στο σύνολο επικύρωσης δοκιμάστηκαν τρεις διαφορετικοί πυρήνες του SVM : **linear**, **polynomial** και **RBF**. Τα αποτελέσματα έδειξαν ότι ο RBF πυρήνας παρουσίασε την καλύτερη απόδοση με F1-score = 0.9569 και AUC = 0.9962, ξεπερνώντας τόσο τον γραμμικό όσο και τον πολυωνυμικό πυρήνα. Έτσι, ο **RBF** επιλέχθηκε ως ο **βέλτιστος** πυρήνας για την τελική αξιολόγηση του μοντέλου.

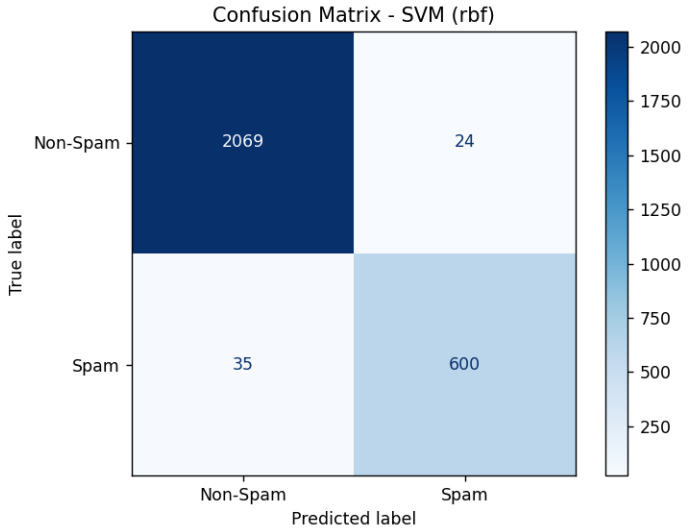
```
 Validation results
kernel=linear | F1=0.9018 | AUC=0.9898
kernel=poly   | F1=0.9565 | AUC=0.9952
kernel=rbf    | F1=0.9569 | AUC=0.9962
```

```
Best kernel on validation (by F1): rbf (F1=0.9569)
```

Στο σύνολο ελέγχου, το **SVM με RBF** πυρήνα πέτυχε:

```
 Test results
Precision: 0.9615384615384616
Recall   : 0.9448818897637795
F1       : 0.9531374106433678
AUC      : 0.9976140942248439
```

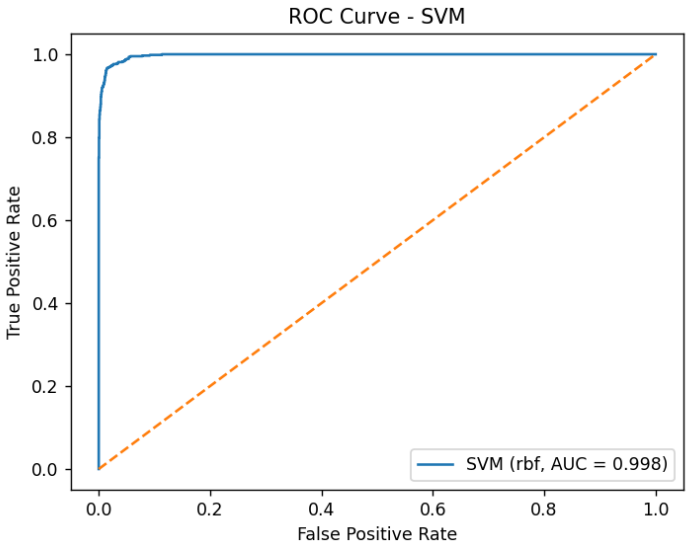
Τα αποτελέσματα δείχνουν **καλή ισορροπία** μεταξύ **precision** και **recall**, γεγονός που υποδηλώνει ότι το μοντέλο εντοπίζει αποτελεσματικά τα spam emails χωρίς να αυξάνει σημαντικά τα false positives.



Confusion Matrix

Από τα **2093** non-spam emails, τα **2069** ταξινομήθηκαν σωστά, ενώ **24** χαρακτηρίστηκαν λανθασμένα ως spam.

Από τα **635** spam emails, τα **600** ανιχνεύθηκαν σωστά, ενώ μόνο **35** ταξινομήθηκαν λανθασμένα ως non-spam.



Η καμπύλη **ROC** του **SVM** βρίσκεται πολύ κοντά στο άνω αριστερό άκρο του διαγράμματος, με **AUC = 0.998**, γεγονός που υποδηλώνει σχεδόν ιδανική διακριτική ικανότητα. Αυτό σημαίνει ότι το μοντέλο διατηρεί υψηλή απόδοση ανεξάρτητα από το κατώφλι απόφασης.

ΕΡΩΤΗΜΑ 6 : Μείωση διάστασης με PCA - E6_PCA.py

Εδώ εξετάζεται η επίδραση της μείωσης διάστασης μέσω **Ανάλυσης Κύριων Συνιστωσών** στην απόδοση του ταξινομητή **SVM** με **RBF** πυρήνα, ο οποίος είχε επιλεγεί ως ο καλύτερος στο ερώτημα 5.

Το **PCA** εφαρμόστηκε με στόχο τη διατήρηση του **90%**, **95%** και **99%** της συνολικής μεταβλητότητας των **embeddings**.

Αποτελέσματα στο σύνολο επικύρωσης:

```
 PCA + SVM results
Variance=90% | PCA dims=113 | Val F1=0.9548 | Val AUC=0.9960
Variance=95% | PCA dims=149 | Val F1=0.9613 | Val AUC=0.9962
Variance=99% | PCA dims=231 | Val F1=0.9593 | Val AUC=0.9963
```

Στην περίπτωση του **90%**, η διάσταση μειώθηκε από **384** σε **113** χαρακτηριστικά, επιτυγχάνοντας F1-score ίσο με 0.9548 και AUC 0.9960.

Με **διατήρηση του 95%** της μεταβλητότητας, η διάσταση μειώθηκε σε **149** χαρακτηριστικά και το μοντέλο πέτυχε την καλύτερη συνολική απόδοση με F1-score 0.9613 και AUC 0.9962.

Τέλος, για **διατήρηση του 99%** της μεταβλητότητας, χρησιμοποιήθηκαν **231** χαρακτηριστικά, με F1-score 0.9593 και AUC 0.9963.

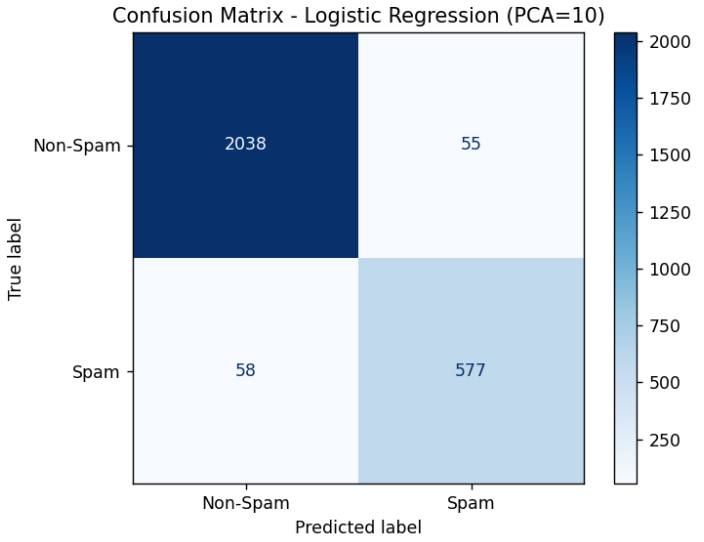
Τα αποτελέσματα δείχνουν ότι η διατήρηση του 95% της μεταβλητότητας προσφέρει την καλύτερη ισορροπία μεταξύ μείωσης διάστασης και απόδοσης.

ΕΡΩΤΗΜΑ 7 : Λογιστική Παλινδρόμηση μετά από PCA - E7_10PCA.py

Εδώ εφαρμόζεται **Logistic Regression** σε δεδομένα που έχουν μειωθεί σε **10 διαστάσεις** με χρήση **PCA**. Στόχος είναι να εξεταστεί αν ένα απλό μοντέλο μπορεί να επιτύχει ικανοποιητική απόδοση στην ταξινόμηση spam emails με πολύ μικρό αριθμό χαρακτηριστικών.

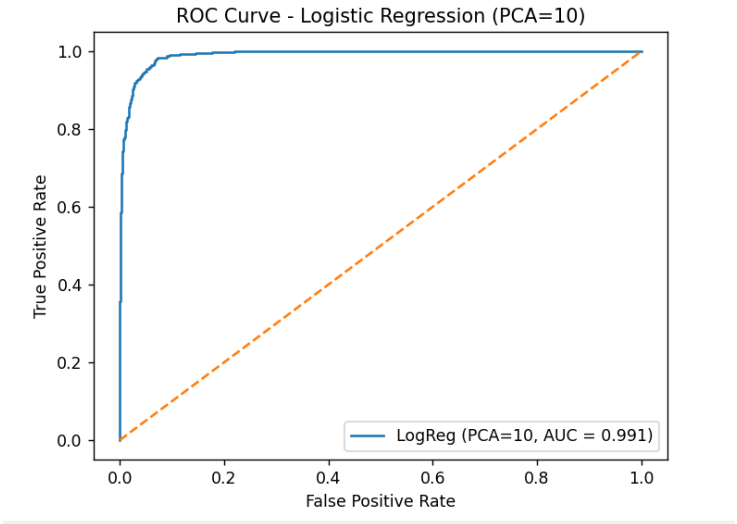
```
Logistic Regression με PCA=10 Test Set
Precision: 0.9129746835443038
Recall   : 0.9086614173228347
F1       : 0.9108129439621152
AUC      : 0.990727998465075
```

Παρατηρείται καλή συνολική απόδοση παρά τη σημαντική μείωση της διάστασης των δεδομένων.



Confusion Matrix

Ο πίνακας δείχνει ότι από τα **2093** non-spam emails, τα **2038** ταξινομήθηκαν σωστά, ενώ **55** χαρακτηρίστηκαν λανθασμένα ως spam. Αντίστοιχα, από τα **635** spam emails, τα **577** ανιχνεύθηκαν σωστά και **58** ταξινομήθηκαν λανθασμένα ως non-spam. Τα αποτελέσματα δείχνουν καλή ισορροπία μεταξύ false positives και false negatives.



Η καμπύλη **ROC** βρίσκεται πολύ κοντά στο άνω αριστερό τμήμα του διαγράμματος, με **AUC = 0.991**, γεγονός που υποδηλώνει υψηλή διακριτική ικανότητα του μοντέλου. Αυτό δείχνει ότι ακόμη και με μόνο 10 διαστάσεις, η Logistic Regression μπορεί να διαχωρίσει αποτελεσματικά spam και non-spam emails.

Συμπέρασμα

Στην παρούσα εργασία μελετήθηκαν και συγκρίθηκαν διαφορετικές μέθοδοι ταξινόμησης για την ανίχνευση spam emails, χρησιμοποιώντας τόσο κλασικές αναπαραστάσεις κειμένου όσο και σημασιολογικά embeddings. Η αξιολόγηση των μοντέλων πραγματοποιήθηκε με μετρικές Precision, Recall, F1-score και AUC, καθώς και με confusion matrix και καμπύλες ROC.

Ο παρακάτω πίνακας συνοψίζει τα τελικά αποτελέσματα στο σύνολο ελέγχου για κάθε μοντέλο:

Μοντέλο	Precision	Recall	F1-score	AUC
Naive Bayes (TF-IDF)	0.993	0.907	0.948	0.999
k-NN (k=3, Embeddings)	0.983	0.844	0.908	0.968
SVM (RBF, Embeddings)	0.962	0.945	0.953	0.998
PCA 95% + SVM (RBF)	–	–	0.961	0.996
Logistic Regression + PCA=10	0.913	0.909	0.911	0.991

Από τα αποτελέσματα παρατηρείται ότι το **SVM με RBF πυρήνα**, και ειδικότερα σε συνδυασμό με **PCA στο 95% της μεταβλητότητας**, παρουσίασε την καλύτερη συνολική απόδοση, επιτυγχάνοντας το υψηλότερο **F1-score** με σημαντικά μειωμένη διάσταση χαρακτηριστικών.

Παράλληλα, η **Logistic Regression με PCA=10** έδειξε ότι ακόμη και απλά γραμμικά μοντέλα μπορούν να επιτύχουν ικανοποιητική απόδοση με πολύ χαμηλό υπολογιστικό κόστος.

Συνολικά, η χρήση **σημασιολογικών embeddings** σε συνδυασμό με κατάλληλες τεχνικές **μείωσης διάστασης** βελτιώνει σημαντικά την αποτελεσματικότητα των ταξινομητών στο πρόβλημα ανίχνευσης spam.