**Α ΜΕΡΟΣ**

**Αφελής ταξινομητής Bayes**

Για την υλοποίηση του αλγορίθμου Naive Bayes, χρησιμοποιήθηκε η υπερπαράμετρος alpha = 1.0 και προστέθηκε ώστε να αποφευχθεί η πιθανότητα μηδενικής συχνότητας χαρακτηριστικών. Η τιμή αυτή επελέγη βάσει της βιβλιογραφίας, καθώς η προσθήκη του 1 στη συχνότητα εμφάνισης λέξεων βελτιώνει την απόδοση σε περιπτώσεις όπου υπάρχουν σπάνια χαρακτηριστικά.

***Καμπύλες Μάθησης***

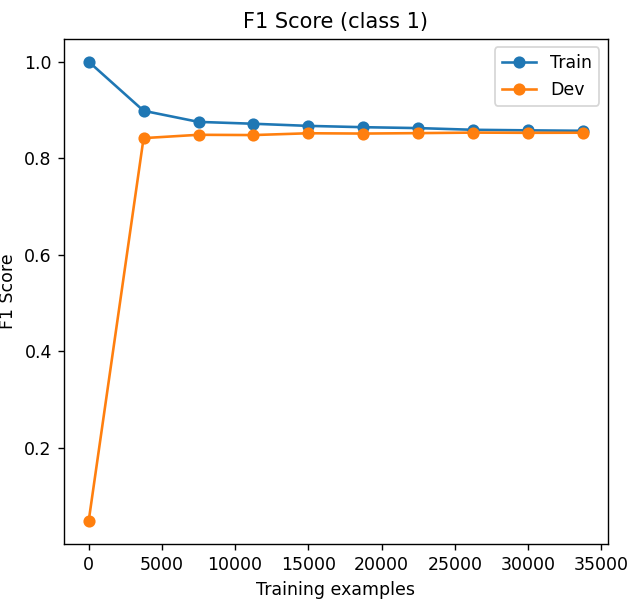
### Για την αξιολόγηση του μοντέλου, υπολογίστηκαν οι καμπύλες μάθησης χρησιμοποιώντας τα training data και τα development data. Το πλήθος των παραδειγμάτων εκπαίδευσης αυξήθηκε σταδιακά σε 10 βήματα, από 10 έως το συνολικό πλήθος των παραδειγμάτων εκπαίδευσης. Σε κάθε επανάληψη, εκπαιδεύτηκε το μοντέλο και μετρήθηκαν τα Precision, Recall και F1 Score για την κατηγορία 1, τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο ανάπτυξης. Επιλέξαμε την πρώτη κατηγορία για την αξιολόγηση, καθώς συνήθως σε προβλήματα δυαδικής ταξινόμησης, η θετική κλάση είναι αυτή που παρουσιάζει μεγαλύτερο ενδιαφέρον. Οι καμπύλες μάθησης που προέκυψαν παρουσιάζονται στα παρακάτω γραφήματα:

### F1-score: Το training F1 ξεκινά από 1.0 και σταθεροποιείται γύρω στο 0.85. Το dev F1 ανεβαίνει γρήγορα και φτάνει επίσης το 0.85. Αυτό δείχνει ότι το μοντέλο μαθαίνει σωστά και αποφεύγει το overfitting.

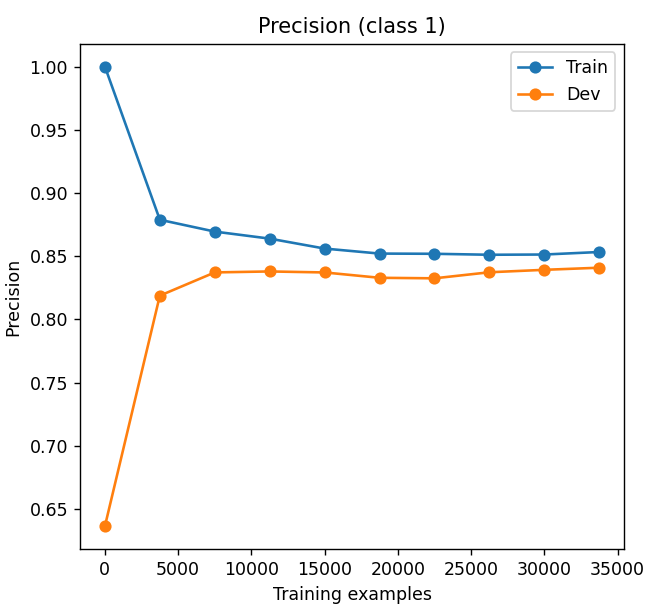
Precision: Το precision του training set ξεκινά από 1.0 και πέφτει σταδιακά γύρω στο 0.85. Το dev precision ανεβαίνει από ~0.65 σε 0.84, δείχνοντας βελτίωση με περισσότερα δεδομένα. Η μικρή διαφορά μεταξύ των δύο σημαίνει ότι το μοντέλο γενικεύει καλά.

Recall: Το training recall ξεκινά από 1.0 και σταθεροποιείται γύρω στο 0.86. Το dev recall ανεβαίνει από 0.0 σε 0.86, πράγμα που δείχνει ότι το μοντέλο γίνεται καλύτερο στον εντοπισμό θετικών περιπτώσεων.

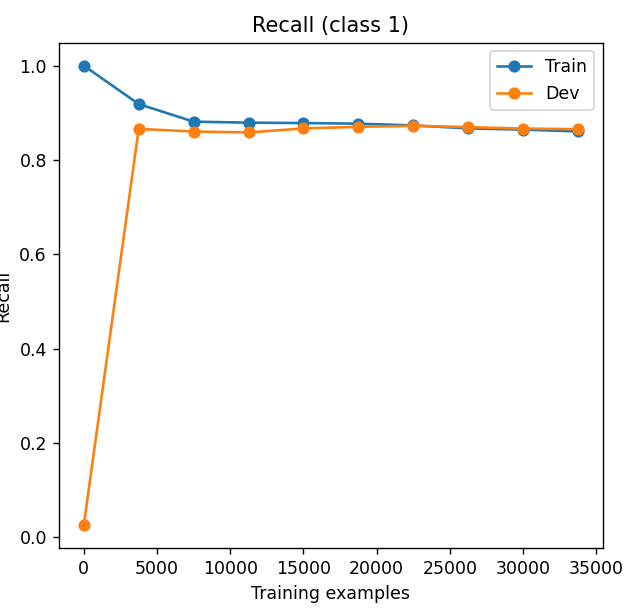
F1 Score (class 1) σε σχέση με το πλήθος δεδομένων εκπαίδευσης



Precision (class 1) σε σχέση με το πλήθος δεδομένων εκπαίδευσης

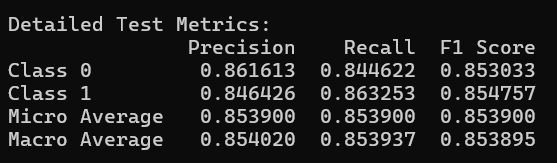


Recall (class 1) σε σχέση με το πλήθος δεδομένων εκπαίδευσης



***Αποτελέσματα στο Test Set***

Μετά την εκπαίδευση του αλγορίθμου Naive Bayes σε όλα τα δεδομένα εκπαίδευσης, το μοντέλο αξιολογήθηκε στο σύνολο test data και τα αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα:



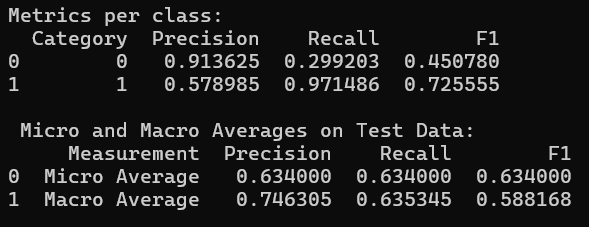
**Τυχαίο Δάσος (Random Forest)**

### Υπερπαράμετροι και Επιλογή τους: στην υλοποίηση του **ID3 Decision Tree** και του **RandomForestID3**, χρησιμοποιήθηκαν οι εξής υπερπαράμετροι προκειμένου να επιτευχθεί μια ισορροπία μεταξύ αποφυγής υπερπροσαρμογής (overfitting) και διατήρησης ικανοποιητικής απόδοσης του ταξινομητή.

* **Μέγιστο βάθος δέντρου (max\_depth): 4**
* **Ελάχιστος αριθμός δειγμάτων για διαχωρισμό (min\_samples\_split): 10**
* **Αριθμός δέντρων στο τυχαίο δάσος (n\_trees): 10**

### *Αποτελέσματα Αξιολόγησης*

Μετά την εκπαίδευση του **RandomForestID3** ταξινομητή, πραγματοποιήθηκε αξιολόγηση στο **development set** και στο **test set**. Παρακάτω παρουσιάζονται τα αποτελέσματα.



**Β ΜΕΡΟΣ**

**NaiveBayesSK**

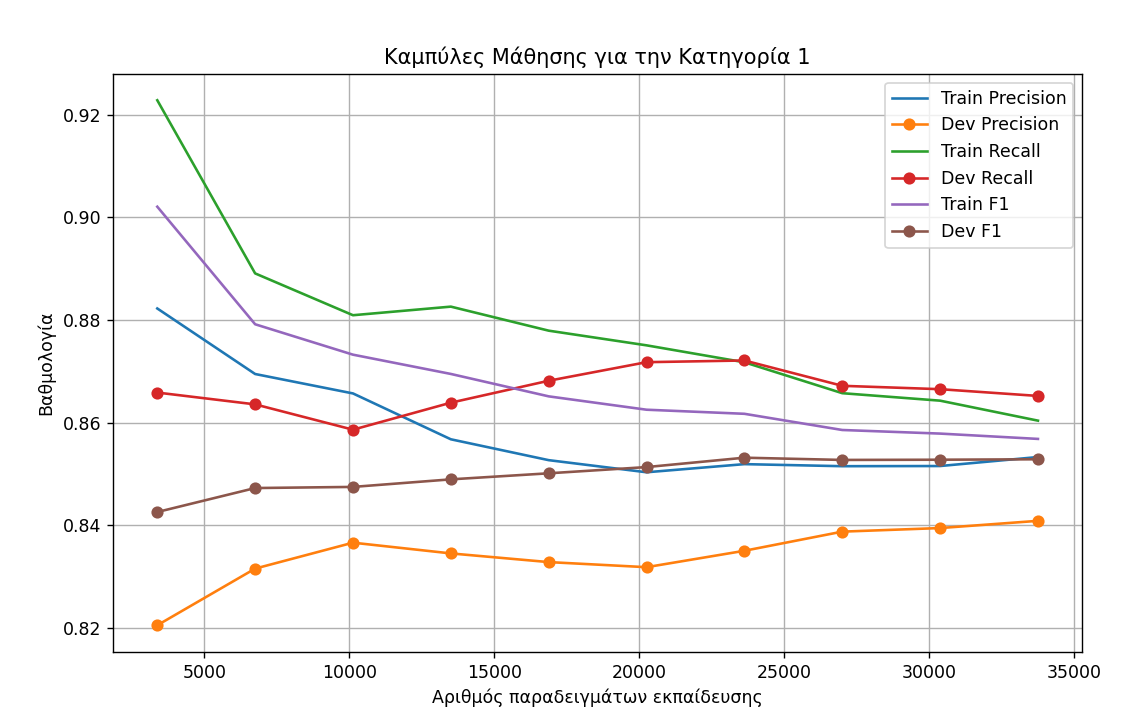
**Επιλέξαμε την κατηγορία 1 στις καμπύλες μάθησης** γιατί είναι η **θετική κατηγορία**, που συνήθως είναι η πιο σημαντική σε ένα classification πρόβλημα και δείχνει **πώς το μοντέλο μαθαίνει να εντοπίζει τις σημαντικές περιπτώσεις** καθώς αυξάνονται τα δεδομένα.  
Επίσης, οι μετρικές της (precision, recall, F1-score) είναι **πολύ κοντά στις συνολικές επιδόσεις** του μοντέλου, καθιστώντας την **αντιπροσωπευτική για την απόδοσή του**.

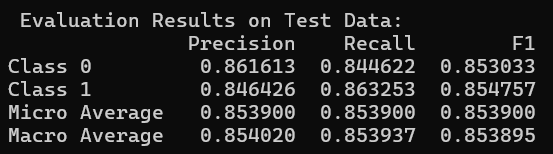
*Train Recall (Πράσινη γραμμή):* Ξεκινά πολύ υψηλά (~0.92) και μειώνεται σταδιακά καθώς αυξάνονται τα δεδομένα. Αυτό δείχνει ότι το μοντέλο προσαρμόζεται καλύτερα σε μεγαλύτερα σύνολα δεδομένων και αποφεύγει το overfitting.

*Dev Recall (Κόκκινη γραμμή):* Παραμένει σταθερό και τείνει να προσεγγίσει την απόδοση του training recall, δείχνοντας καλή γενίκευση.

*Train & Dev Precision (Μπλε & Πορτοκαλί γραμμή):* Το training precision ξεκινά από υψηλότερη τιμή και σταδιακά μειώνεται. Το dev precision είναι χαμηλότερο αλλά παραμένει σταθερό, κάτι που σημαίνει ότι το μοντέλο δεν υπερπροσαρμόζεται στα δεδομένα εκπαίδευσης.

*Train & Dev F1-score (Μοβ & Καφέ γραμμή):* Οι καμπύλες F1-score δείχνουν σταθερότητα καθώς αυξάνονται τα δεδομένα, επιβεβαιώνοντας ότι το μοντέλο μαθαίνει ομαλά χωρίς να παρουσιάζει υπερεκπαίδευση (overfitting).





## *****Γενική Απόδοση του Μοντέλου*****

**Test Accuracy: 85.39%** Η συνολική ακρίβεια του μοντέλου (micro avg) είναι **0.8539**, που σημαίνει ότι περίπου το **85.4% των προβλέψεων είναι σωστές**. **Macro Average-F1-score: 0.8539** Το μοντέλο διατηρεί μια **ισορροπημένη απόδοση** σε όλες τις κατηγορίες, καθώς το Macro Average είναι πολύ κοντά στο micro average. Συμπερασματικά, το μοντέλο επιτυγχάνει **πολύ ισορροπημένες τιμές precision και recall**, με μικρές διαφορές μεταξύ των δύο κλάσεων.

**RandomForestSK**

***Precision (Πρώτο Γράφημα)***

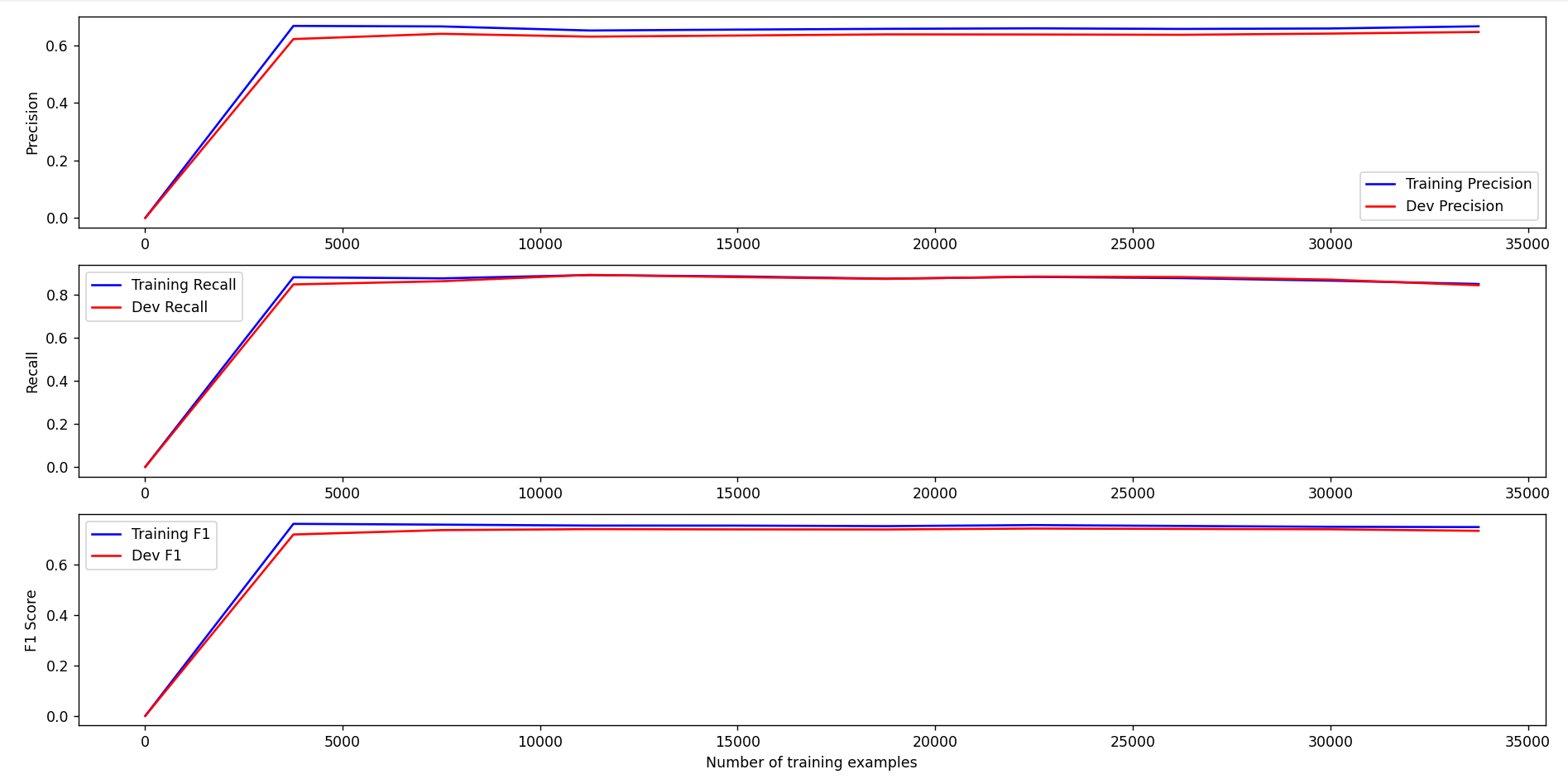
Αρχικά, το precision είναι χαμηλό επειδή το μοντέλο έχει λίγα δεδομένα και κάνει λάθος προβλέψεις. Όσο όμως αυξάνονται τα δεδομένα εκπαίδευσης, το precision βελτιώνεται και για το training και για το dev set. Η μπλε γραμμή (training) είναι ελαφρώς υψηλότερη από την κόκκινη (dev), κάτι που δείχνει μικρό overfitting αλλά όχι υπερβολικό. Σταθεροποιείται κοντά στο 0.65-0.7,

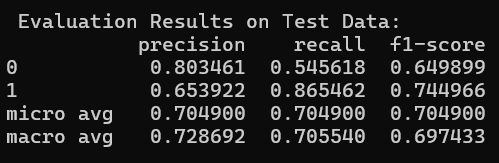
## *****Recall (Δεύτερο Γράφημα)*****

## **Ξεκινά χαμηλά και ανεβαίνει γρήγορα με τα πρώτα δεδομένα. Σταθεροποιείται κοντά στο 0.8, που σημαίνει ότι το μοντέλο μπορεί να ανιχνεύσει τη μεγάλη πλειοψηφία των θετικών περιπτώσεων. Η διαφορά μεταξύ training και dev recall είναι πολύ μικρή, που σημαίνει ότι το μοντέλο γενικεύει καλά.**

## *****F1-score (Τρίτο Γράφημα)*****

## **Σταθεροποιείται γύρω στο 0.7, που είναι ικανοποιητική τιμή. Όπως και με το recall, η διαφορά μεταξύ training και dev είναι πολύ μικρή, δείχνοντας καλή γενίκευση. Άρα, τ**ο F1-score είναι σταθερό και δείχνει ότι το μοντέλο έχει καλή ισορροπία μεταξύ ακρίβειας και ανάκλησης.





***Σύγκριση RandomForestClassifier (Scikit-learn) και RandomForestID3***

**Ο Scikit-learn RandomForest έχει σαφώς καλύτερη ακρίβεια από τον ID3. Ο Scikit-learn RandomForest υπερέχει σημαντικά στο Micro Average (~0.705) έναντι του ID3 (~0.634) και έχει πιο ισορροπημένη απόδοση και στις δύο κλάσεις.  
Ο ID3 έχει υψηλότερο macro precision (0.7463 έναντι 0.7287), αλλά χαμηλότερο recall και F1-score. Επίσης, έχει πολύ υψηλό precision στην κλάση 0 (0.9136), αλλά εξαιρετικά χαμηλό Recall (0.2992).** Αξίζει να σημειωθεί ότι χρησιμοποιούμε τιςίδιες μετρικές (Accuracy, Precision, Recall, F1-score) και για τις δύο υλοποιήσεις, ενώ, τα δύο μοντέλα έχουν παρόμοιες ρυθμίσεις για n =10, max\_depth = 4 και min\_samples\_split = 10.

***Σύγκριση Naive Bayes και Scikit-learn Naive Bayes***

**Τα δύο μοντέλα έχουν πολύ κοντινή ακρίβεια**, με το **Scikit-learn BernoulliNB** να έχει ελαφρώς καλύτερη απόδοση. Το **Scikit-learn Naive Bayes έχει λίγο καλύτερο F1-score και recall στην κλάση 1**, πράγμα που σημαίνει ότι **εντοπίζει περισσότερα θετικά δείγματα**. **Τα δύο μοντέλα έχουν σχεδόν ταυτόσημες τιμές Micro & Macro Average,**  
με το **Scikit-learn BernoulliNB να έχει μικρό προβάδισμα (~0.003 διαφορά)**, αλλά η διαφορά είναι πρακτικά αμελητέα.