ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

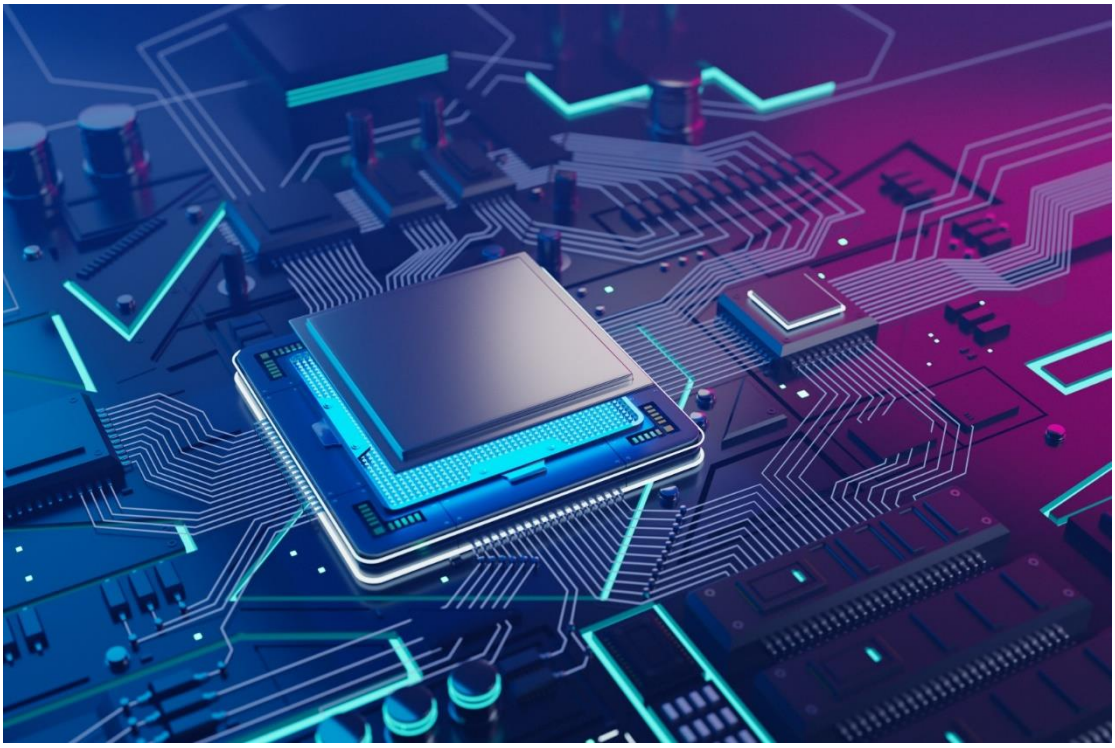
ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

ΧΕΙΜΕΡΙΝΟ ΕΞΑΜΗΝΟ 2023-24

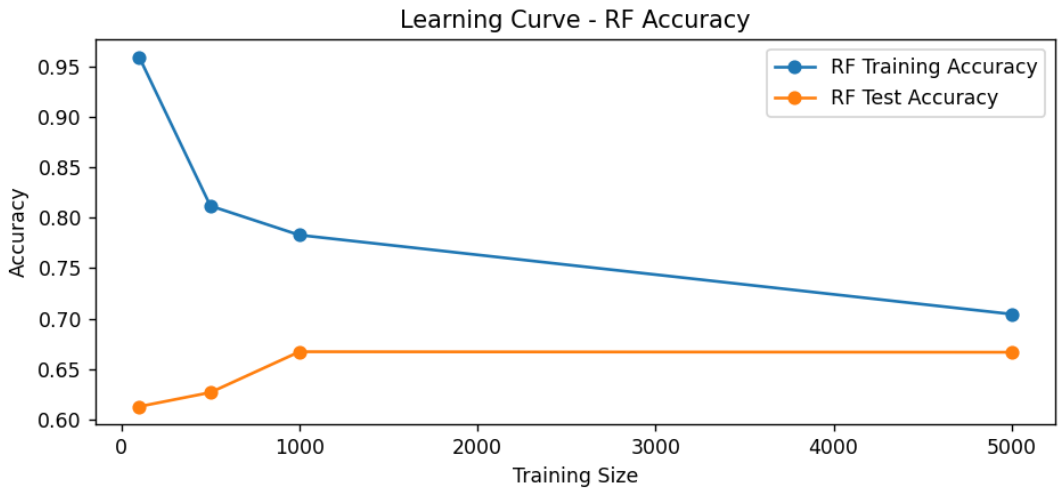
ΓΙΩΡΓΟΣ ΚΟΥΡΟΣ-3190095

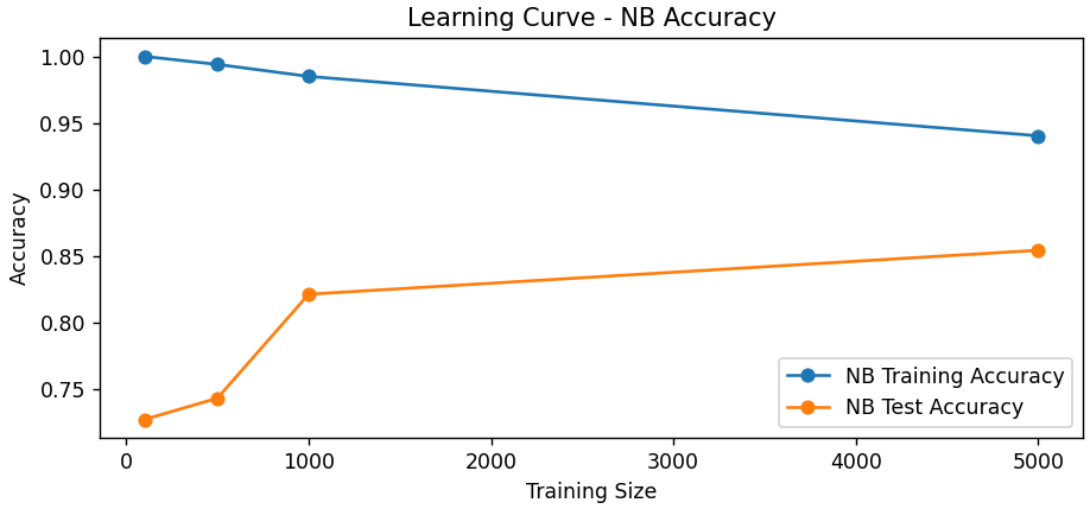
ΚΩΝΣΤΑΝΤΙΝΟΣ ΑΝΔΡΙΝΟΠΟΥΛΟΣ-3190009

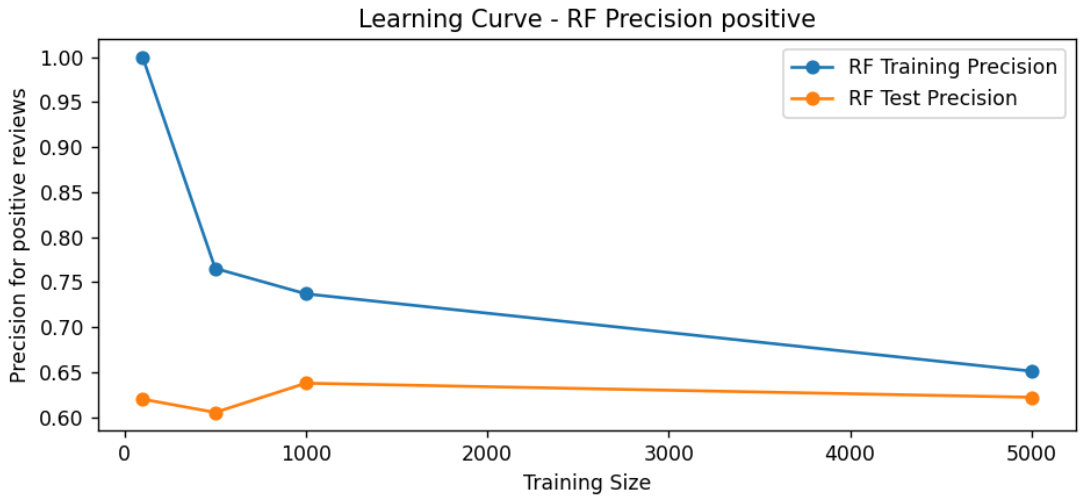
ΕΡΓΑΣΙΑ #2 – RANDOM FOREST ID3 & NAIVE BAYES

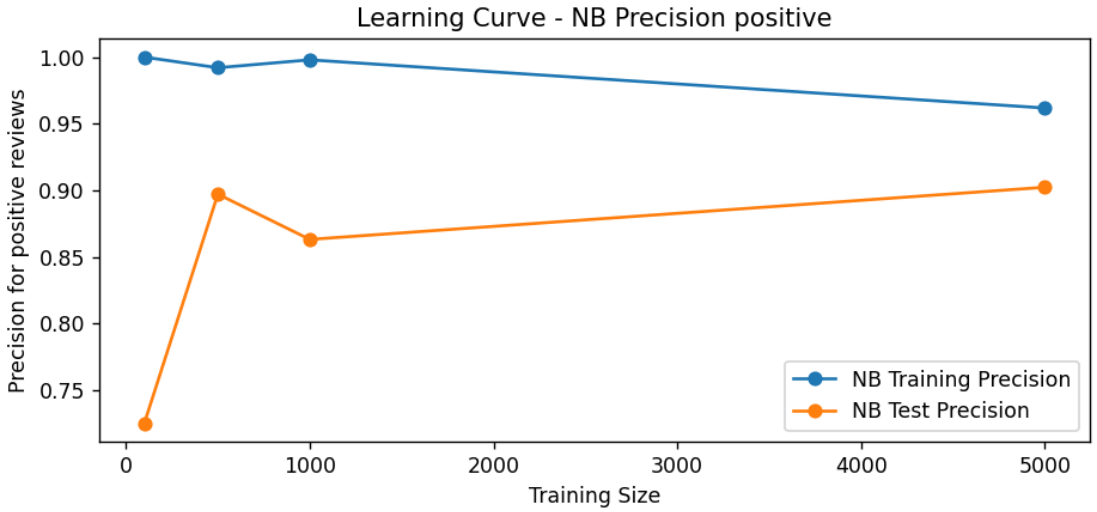


Α) Το πρόγραμμά μας χρησιμοποιεί τους αλγορίθμους Random Forest ID3 με maximum tree depth και αριθμό δένδρων σαν υπερπαραμέτρους και Naive Bayes με πολυμεταβλητή Bernoulli. Τα παρακάτω αποτελέσματα είναι για 4 μεγέθη δεδομένων εκπαίδευσης 100,500,1000 και 5000 (100 σημαίνει 50 θετικές και 50 αρνητικές κριτικές κ.ο.κ.), ο Random Forest κατασκευάζει 400 δέντρα με μέγιστο βάθος 15 και τα m,n,k έχουν τιμές 10000,50,100 αντίστοιχα. Η επιλογή των τιμών έγινε μετά από πειραματισμό και καταλήξαμε σε αυτές διότι μας εξασφάλιζαν το υψηλότερο ποσοστό για τις μετρικές μας (accuracy, precision, recall, F1). Όσο αυξάναμε τον αριθμό δέντρων και βάθους τόσο καλύτερα αποτελέσματα παίρναμε, αλλά από τα 400 δέντρα με 15 μέγιστο βάθος και μετά, τα αποτελέσματα άρχισαν να αλλοιώνονται, γι'αυτό το αφήσαμε σε αυτά τα νούμερα (με μικρό αριθμό δένδρων παρατηρήσαμε υποπροσαρμογή). Το ίδιο συνέβη και με τις παραμέτρους m,n,k όπου μετά από τις 10000 λέξεις υπήρχε υπερπροσαρμογή οπότε στον συνδυασμό 10000,50,100 είχαμε μια καλή εκτίμηση. Τα μεγέθη των δεδομένων εκπαίδευσης είναι ενδεικτικά, 2 μικρά μεγέθη (100,500) και 2 σχετικά μεγάλα (1000,5000). Ωστόσο θα μπορούσαν να χρησιμοποιηθούν όλα τα δεδομένα εκπαίδευσης (και τα 25000) αλλά επιλέξαμε αυτά τα μεγέθη δεδομένων εκπαίδευσης ώστε να γίνει αισθητή η σημαντικότητα του πλήθους των training data, πόσο σημαντικό ρόλο έχει για τις μεταβλητές accuracy, precision, recall και F1. Παρακάτω μπορούμε να δούμε τα αποτελέσματα (καμπύλες μάθησης και πίνακες) για τις συγκεκριμένες τιμές των παραμέτρων που χρησιμοποιήσαμε. (RF = Random Forest ID3, NB = Naive Bayes). Ακρίβεια, ανάκληση και F1 αφορά τις θετικές κριτικές.







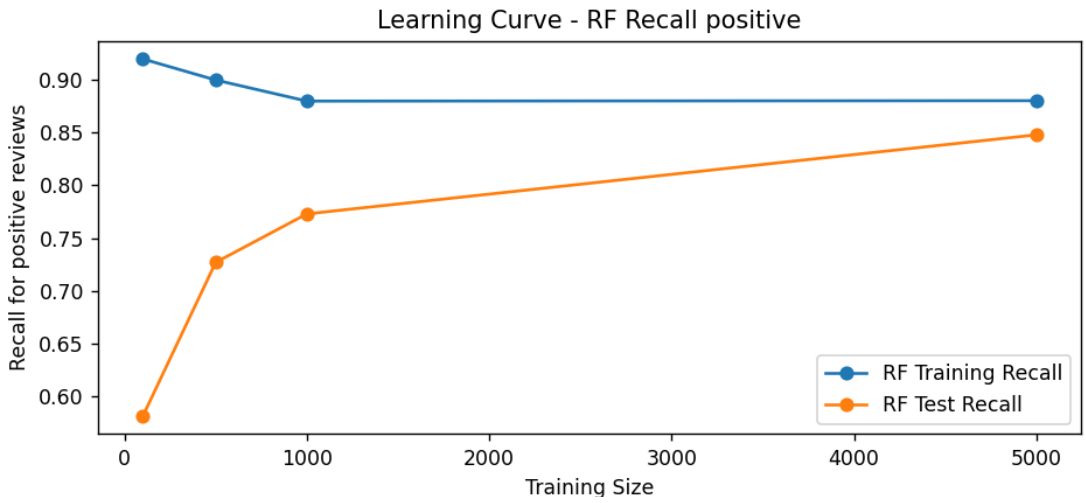


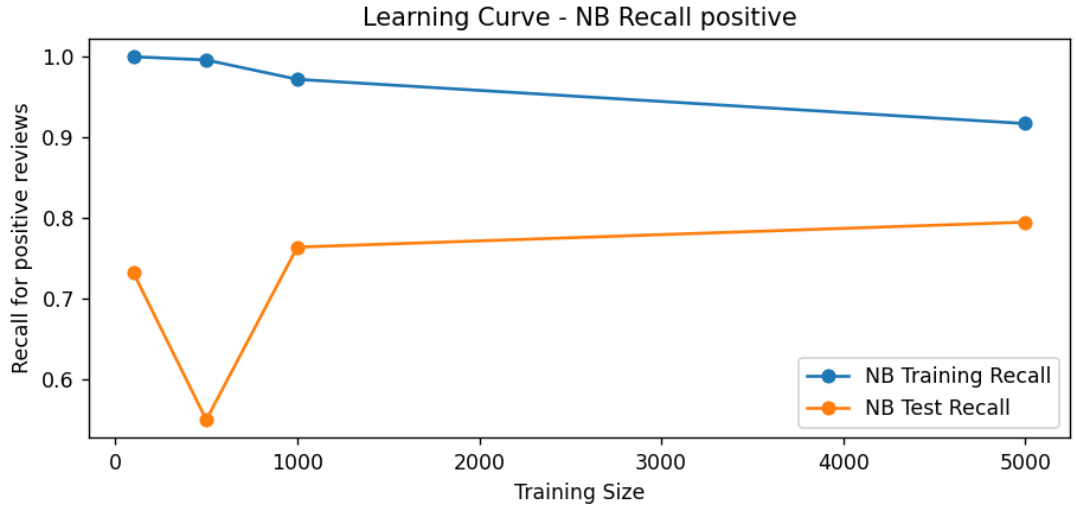

```
RESULTS FOR TRAINING DATA SIZE = 100

Random Forest Values:

Training Data:
Accuracy: 96.00%
Precision for positive reviews: 100.00%
Recall for positive reviews: 92.00%
F1 Score for positive reviews: 95.83%

Test Data:
Accuracy: 61.25%
Precision for positive reviews: 62.01%
Recall for positive reviews: 58.10%
F1 Score for positive reviews: 59.99%
```

```
Naive Bayes Values:

Training Data:
Accuracy: 100.00%
Precision for positive reviews: 100.00%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 100.00%

Test Data:
Accuracy: 72.75%
Precision for positive reviews: 72.55%
Recall for positive reviews: 73.20%
F1 Score for positive reviews: 72.87%
```

```
RESULTS FOR TRAINING DATA SIZE = 500

Random Forest Values:

Training Data:
Accuracy: 81.20%
Precision for positive reviews: 76.53%
Recall for positive reviews: 90.00%
F1 Score for positive reviews: 82.72%

Test Data:
Accuracy: 62.65%
Precision for positive reviews: 60.53%
Recall for positive reviews: 72.70%
F1 Score for positive reviews: 66.06%
```

```
Naive Bayes Values:

Training Data:
Accuracy: 99.40%
Precision for positive reviews: 99.20%
Recall for positive reviews: 99.60%
F1 Score for positive reviews: 99.40%

Test Data:
Accuracy: 74.35%
Precision for positive reviews: 89.72%
Recall for positive reviews: 55.00%
F1 Score for positive reviews: 68.20%
```



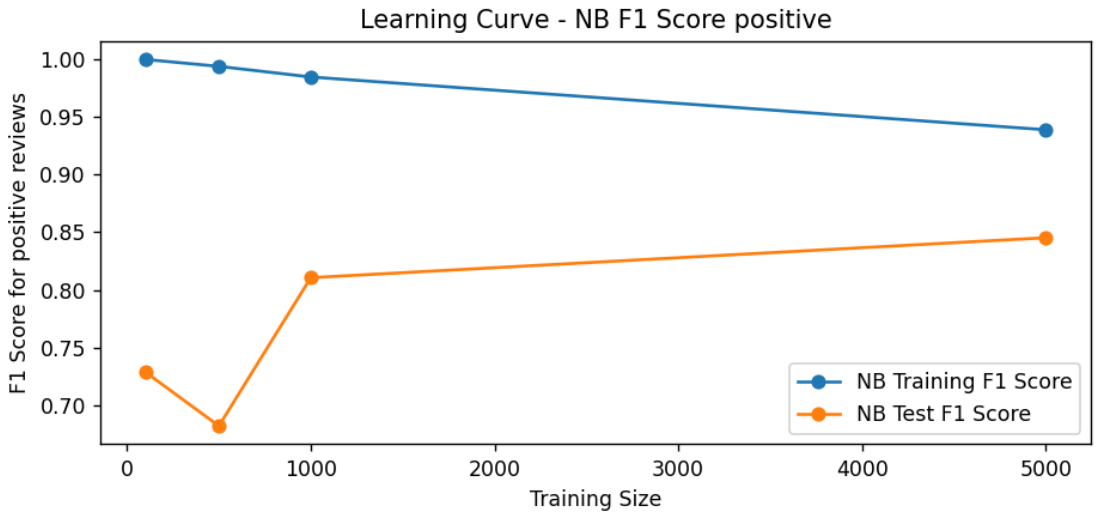


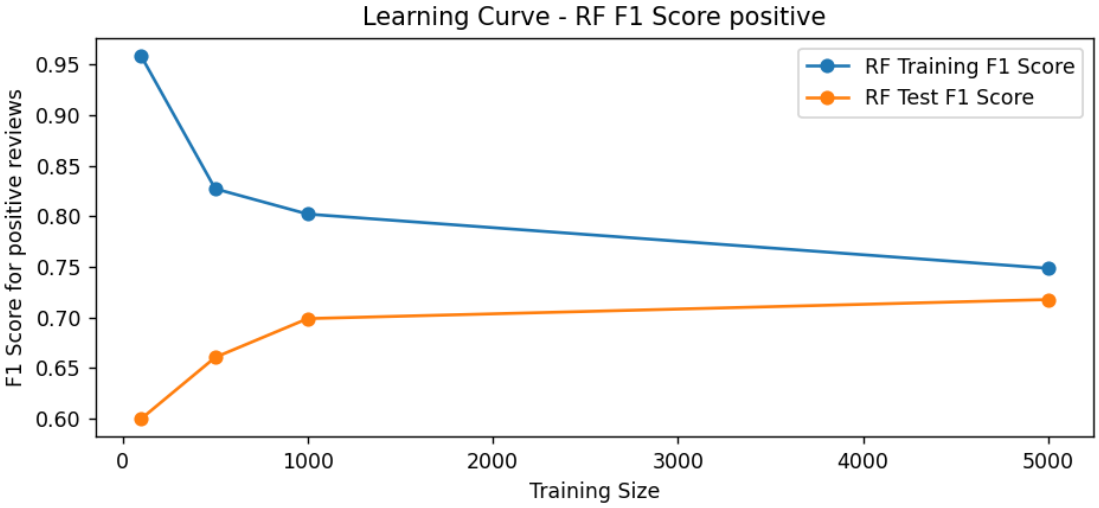

```
RESULTS FOR TRAINING DATA SIZE = 1000

Random Forest Values:

Training Data:
Accuracy: 78.30%
Precision for positive reviews: 73.70%
Recall for positive reviews: 88.00%
F1 Score for positive reviews: 80.22%

Test Data:
Accuracy: 66.70%
Precision for positive reviews: 63.78%
Recall for positive reviews: 77.30%
F1 Score for positive reviews: 69.89%
```

```
Naive Bayes Values:

Training Data:
Accuracy: 98.50%
Precision for positive reviews: 99.79%
Recall for positive reviews: 97.20%
F1 Score for positive reviews: 98.48%

Test Data:
Accuracy: 82.15%
Precision for positive reviews: 86.33%
Recall for positive reviews: 76.40%
F1 Score for positive reviews: 81.06%
```

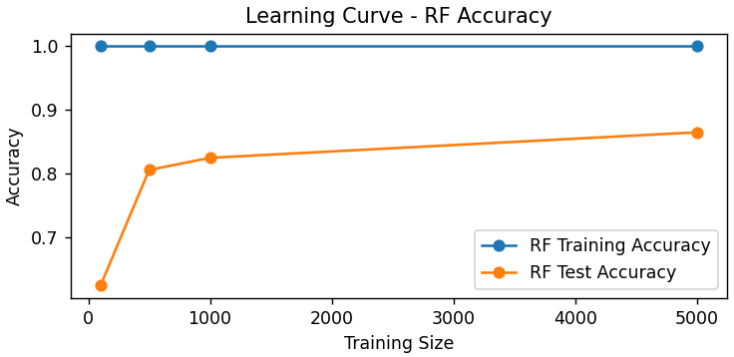
```
RESULTS FOR TRAINING DATA SIZE = 5000

Random Forest Values:

Training Data:
Accuracy: 70.44%
Precision for positive reviews: 65.12%
Recall for positive reviews: 88.04%
F1 Score for positive reviews: 74.86%

Test Data:
Accuracy: 66.65%
Precision for positive reviews: 62.22%
Recall for positive reviews: 84.80%
F1 Score for positive reviews: 71.77%
```

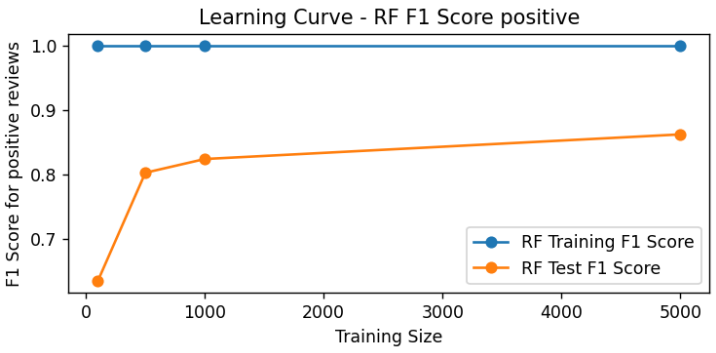
```
Naive Bayes Values:

Training Data:
Accuracy: 94.04%
Precision for positive reviews: 96.18%
Recall for positive reviews: 91.72%
F1 Score for positive reviews: 93.90%

Test Data:
Accuracy: 85.45%
Precision for positive reviews: 90.24%
Recall for positive reviews: 79.50%
F1 Score for positive reviews: 84.53%
```

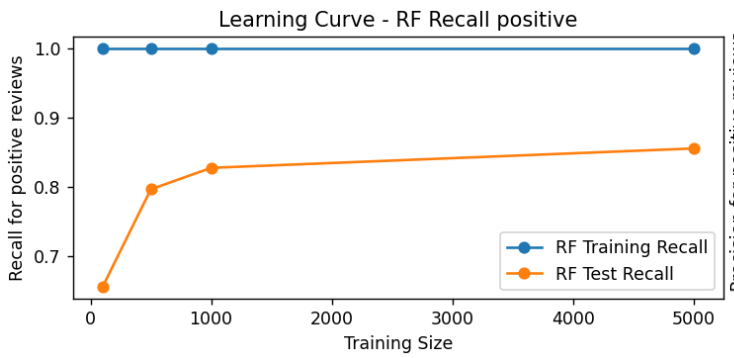


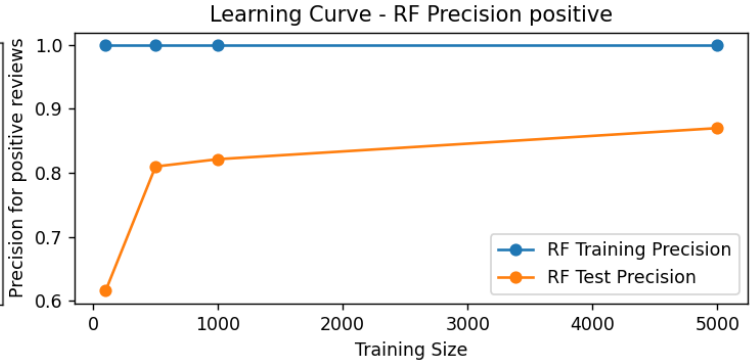


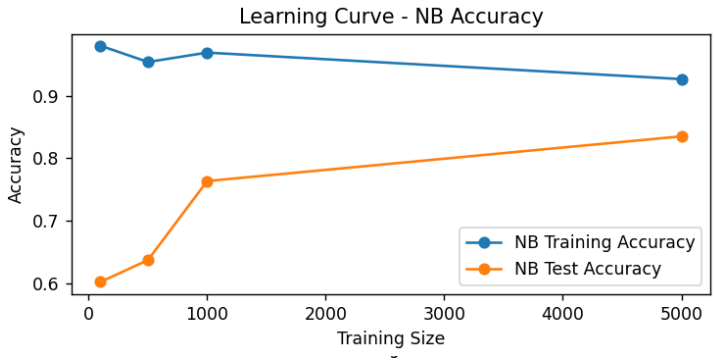
B) Αποτελέσματα για ίδιες τιμές των υπερπαραμέτρων (m,n,k,training sizes) χρησιμοποιώντας Random Forest και Bernoulli Naive Bayes από την βιβλιοθήκη sklearn. (RandomForestClasifier, BernoulliNB)

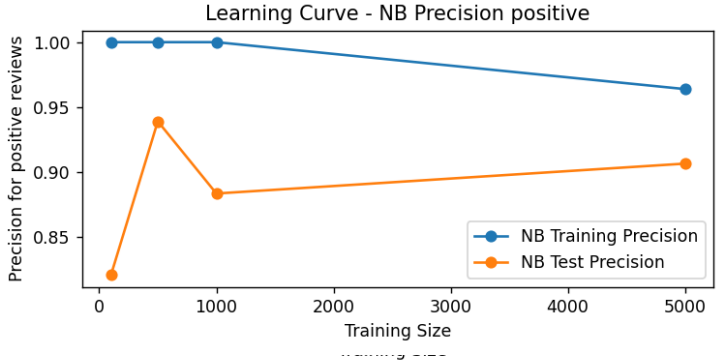


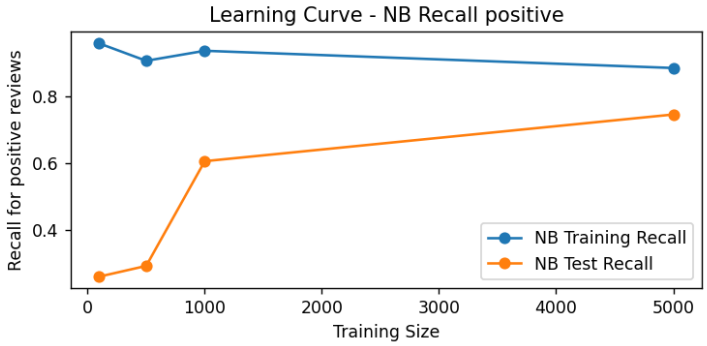


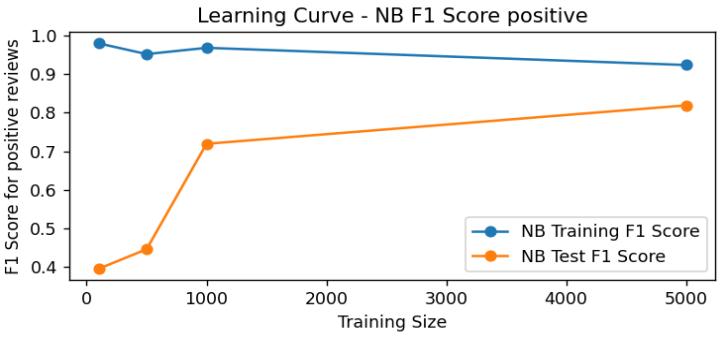













```
RESULTS FOR TRAINING DATA SIZE = 100

Random Forest Values:

Training Data:
Accuracy: 100.00%
Precision for positive reviews: 100.00%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 100.00%

Test Data:
Accuracy: 62.35%
Precision for positive reviews: 61.60%
Recall for positive reviews: 65.60%
F1 Score for positive reviews: 63.54%
```

```
Naive Bayes Values:

Training Data:
Accuracy: 98.00%
Precision for positive reviews: 100.00%
Recall for positive reviews: 96.00%
F1 Score for positive reviews: 97.96%

Test Data:
Accuracy: 60.20%
Precision for positive reviews: 82.08%
Recall for positive reviews: 26.10%
F1 Score for positive reviews: 39.61%
```

```
RESULTS FOR TRAINING DATA SIZE = 500

Random Forest Values:

Training Data:
Accuracy: 100.00%
Precision for positive reviews: 100.00%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 100.00%

Test Data:
Accuracy: 80.50%
Precision for positive reviews: 81.00%
Recall for positive reviews: 79.70%
F1 Score for positive reviews: 80.34%
```

```
Naive Bayes Values:

Training Data:
Accuracy: 95.40%
Precision for positive reviews: 100.00%
Recall for positive reviews: 90.80%
F1 Score for positive reviews: 95.18%

Test Data:
Accuracy: 63.70%
Precision for positive reviews: 93.91%
Recall for positive reviews: 29.30%
F1 Score for positive reviews: 44.66%
```

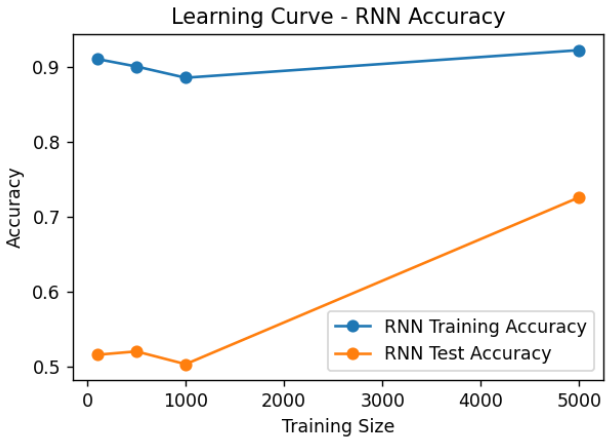
```
RESULTS FOR TRAINING DATA SIZE = 1000

Random Forest Values:

Training Data:
Accuracy: 100.00%
Precision for positive reviews: 100.00%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 100.00%

Test Data:
Accuracy: 82.40%
Precision for positive reviews: 82.14%
Recall for positive reviews: 82.80%
F1 Score for positive reviews: 82.47%
```

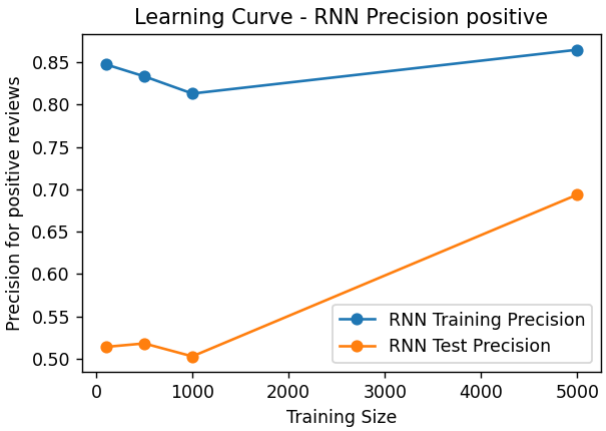
```
Naive Bayes Values:

Training Data:
Accuracy: 96.90%
Precision for positive reviews: 100.00%
Recall for positive reviews: 93.80%
F1 Score for positive reviews: 96.80%

Test Data:
Accuracy: 76.35%
Precision for positive reviews: 88.36%
Recall for positive reviews: 60.70%
F1 Score for positive reviews: 71.96%
```

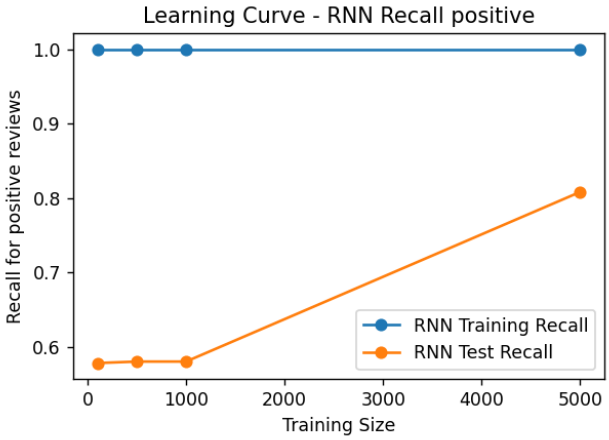
```
RESULTS FOR TRAINING DATA SIZE = 5000

Random Forest Values:

Training Data:
Accuracy: 100.00%
Precision for positive reviews: 100.00%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 100.00%

Test Data:
Accuracy: 86.40%
Precision for positive reviews: 86.99%
Recall for positive reviews: 85.60%
F1 Score for positive reviews: 86.29%
```

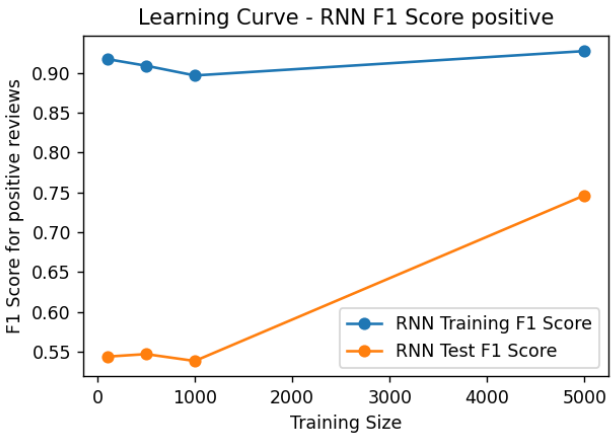
```
Naive Bayes Values:

Training Data:
Accuracy: 92.66%
Precision for positive reviews: 96.39%
Recall for positive reviews: 88.64%
F1 Score for positive reviews: 92.35%

Test Data:
Accuracy: 83.50%
Precision for positive reviews: 90.66%
Recall for positive reviews: 74.70%
F1 Score for positive reviews: 81.91%
```

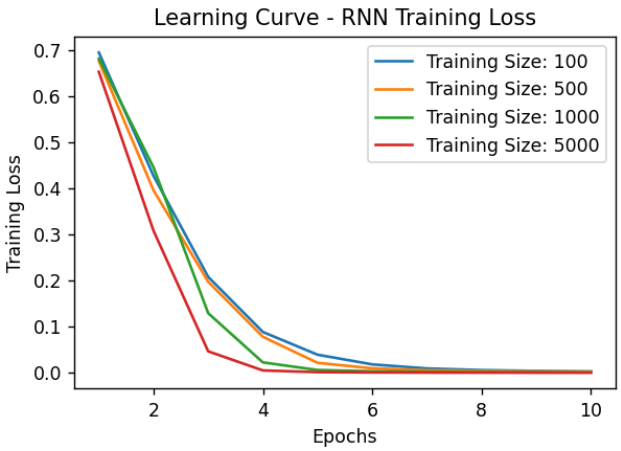
Γ) Αποτελέσματα για τις ίδιες τιμές παραμέτρων (m=vocabulary size, training sizes) χρησιμοποιώντας RNN με word embeddings, καμπύλες μάθησης και πίνακες για δεδομένα εκπαίδευσης και ελέγχου και καμπύλες μεταβολής σφάλματος δεδομένων εκπαίδευσης και ανάπτυξης για κάθε μέγεθος δεδομένων εκπαίδευσης

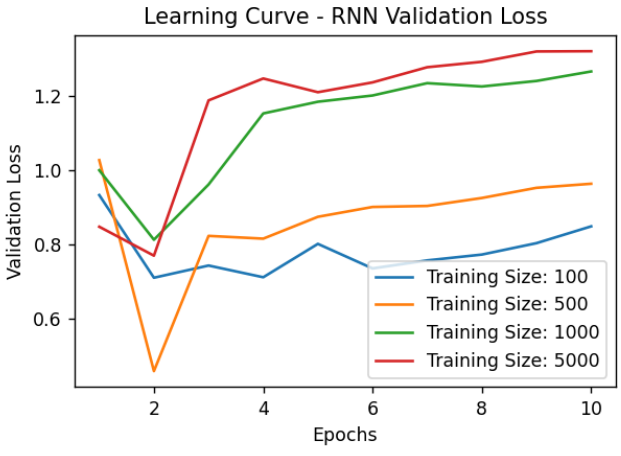












```
RESULTS FOR TRAINING DATA SIZE = 100

Training Data:
Accuracy: 91.00%
Precision for positive reviews: 84.75%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 91.74%

Test Data:
Accuracy: 51.55%
Precision for positive reviews: 51.38%
Recall for positive reviews: 57.80%
F1 Score for positive reviews: 54.40%

RESULTS FOR TRAINING DATA SIZE = 500

Training Data:
Accuracy: 90.00%
Precision for positive reviews: 83.33%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 90.91%

Test Data:
Accuracy: 52.00%
Precision for positive reviews: 51.79%
Recall for positive reviews: 58.00%
F1 Score for positive reviews: 54.72%
```

```
RESULTS FOR TRAINING DATA SIZE = 1000

Training Data:
Accuracy: 88.50%
Precision for positive reviews: 81.30%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 89.69%

Test Data:
Accuracy: 50.30%
Precision for positive reviews: 50.26%
Recall for positive reviews: 58.00%
F1 Score for positive reviews: 53.85%

RESULTS FOR TRAINING DATA SIZE = 5000

Training Data:
Accuracy: 92.18%
Precision for positive reviews: 86.48%
Recall for positive reviews: 100.00%
F1 Score for positive reviews: 92.75%

Test Data:
Accuracy: 72.55%
Precision for positive reviews: 69.36%
Recall for positive reviews: 80.80%
F1 Score for positive reviews: 74.64%
```

Και στα 3 μέρη της εργασίας Α,Β,Γ χρησιμοποιήθηκαν ως δεδομένα ελέγχου 2000 κριτικές από την IMDB, 1000 θετικές και 1000 αρνητικές.