2η Εργασία Τεχνητή Νοημοσύνη

Μιχαήλ Πρωτονοτάριος

Μάριος Γεωργοπετρέας

**Naïve Bayes algorithm:**

**a)**

Χρησιμοποιήσαμε 27 features (λέξεις του λεξιλογίου), φαίνονται κάποιες από αυτές στο παρακάτω pandas dataframe. Καταλήξαμε σε αυτές τις 27 λέξεις φιλτράροντας τα κείμενα από σημεία στίξης, κενά και άλλες άχρηστες λέξεις, έπειτα πήραμε τις 700 πιο συχνές λέξεις από όλα τα κείμενα εκπαίδευσης (αρνητικά και θετικά) και κρατήσαμε αυτές που είχαν information gain μεγαλύτερο ή ίσο του 0.008.

A screenshot of a computer

Description automatically generated

**Stats of our implementation**

* Training data report

A screenshot of a computer

Description automatically generated

* Test data report

A screenshot of a computer

Description automatically generated

Training Curve (Accuracy)

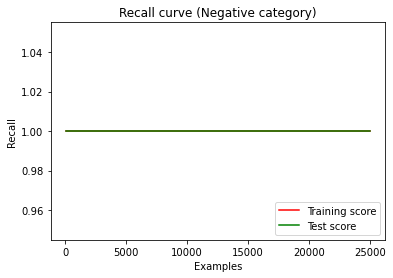
Αυτή η καμπύλη εκπαίδευσης δείχνει το συνολικό ποσοστό ορθότητας (accuracy) και στις δύο κατηγορίες μαζί (δηλαδή τη συνολική ορθότητα και σε θετικά και σε αρνητικά reviews)

A graph on a white background

Description automatically generated

Recall Curve (Negative category)

Τα recall των test data και training data ταυτίζονται και είναι σταθερά στο 1



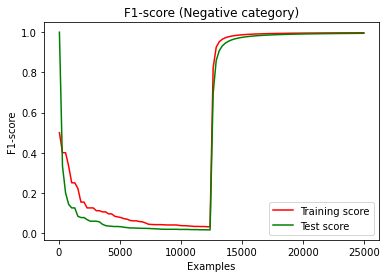
Precision curve (Negative category)

H καμπύλη έχει αυτήν την μορφή γιατί αρχικά τα πρώτα παραδείγματα που διαβάζουμε είναι positive και στα 12500 όπου τα positive τελειώνουν και ξεκινούν τα negative αρχίζει και ανεβαίνει το precision

A graph of a training curve

Description automatically generated with medium confidence

F1- Curve (Negative Category)



**b)**

Σύγκριση των στατιστικών μας με τα στατιστικά της υλοποίησης του naïve bayes algorithm από την βιβλιοθήκη sklearn

Παρατηρούμε ότι τα στατιστικά της υλοποίησης μας σε σχέση με την υλοποίηση της sklearn είναι σχεδόν ίδια και στις δύο κατηγορίες (negative positive) σε όλα τα metrics (precision, recall, f1-score και accuracy)

A screenshot of a computer program

Description automatically generated

Training curve (Accuracy):

A graph of a graph

Description automatically generated with medium confidence

Recall curve:

A white rectangular object with green lines

Description automatically generated

Precision curve:

A graph of a training

Description automatically generated

F1 curve:

A green line graph on a white background

Description automatically generated

**Random Forest algorithm:**

**a)**

Χρησιμοποιήσαμε τα ίδια 27 features για τα δεδομένα μας. Ως υπερπαραμέτρους χρησιμοποιήσαμε 150 δέντρα με 10 τυχαία επιλεγμένα features για το κάθε ένα.

**Stats of our implementation:**

* **Training data report:**

**A screenshot of a computer

Description automatically generated**

* **Test data report:**

A screenshot of a computer

Description automatically generated

Training curve:

A graph of a graph

Description automatically generated

Recall Curve:

A screen shot of a computer

Description automatically generated

Precision Curve:

A graph of a training curve

Description automatically generated

F1-score Curve:

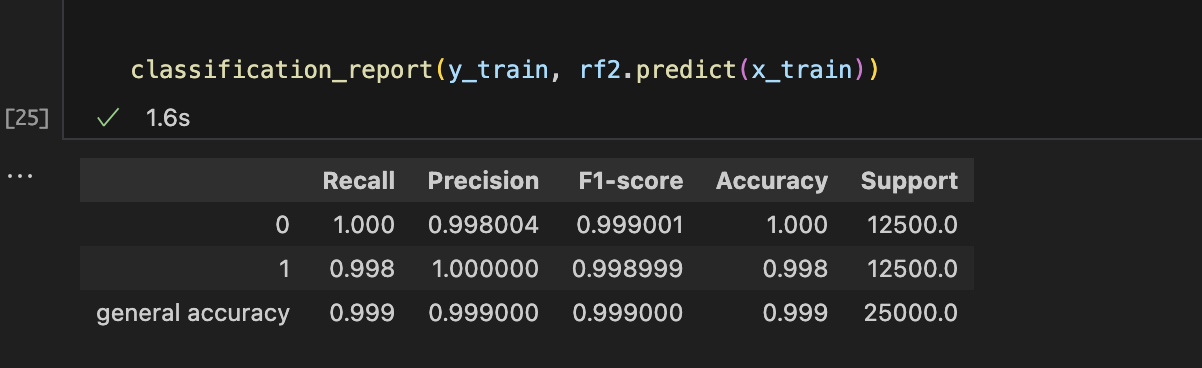
A graph on a white background

Description automatically generated

**b)**

Σύγκριση των στατιστικών μας με τα στατιστικά της υλοποίησης του random forest algorithm από την βιβλιοθήκη Sklearn

Παρατηρούμε ότι οι διαφορές στα στατιστικά της βιβλιοθήκης Sklearn σε σχέση με την δική μας υλοποίηση είναι πολύ μικρές, με την υλοποίηση της Sklearn να είναι περίπου κατά 2% πιο ακριβής, και στα test data και στα training data



A screen shot of a computer

Description automatically generated

* Training Curve (Accuracy):

A graph of a graph

Description automatically generated with medium confidence

* Recall Curve:

A screen shot of a computer

Description automatically generated

* Precision Curve:

A graph of a training

Description automatically generated

* F1-score Curve:

A graph of a training

Description automatically generated