Αξιολόγηση Άρθρων Ειδήσεων

Δημήτρης Σταθόπουλος Ε18151

Γρηγόριος Μάριος Φραγκάκης Ε18173
Ψηφιακών Συστημάτων
Πανεπιστήμιο Πειραιώς
Αθήνα Ελλάδα
grigorisfragkakis@gmail.com
demetresstathopoulos8@gmail.com

Περίληψη
Για την την υλοποίηση του μοντέλου εντοπισμού αναξιόπιστων άρθρων ειδήσεων χρησιμοποιήθηκαν δυο συγκεκριμένοι αλγόριθμοι κατηγοριοποίησης.Συγκεκριμένα : Λογιστική Παλινδρόμηση και Μηχανές διανυσμάτων υποστήριξης .Χρειάστηκε μια προεπεξεργασία των δεδομένων που περιέχονται στα train και test σύνολα δεδομένων .Έγιναν βήματα προς αποφυγή της υπερπροσαρμογής και έγινε αξιολόγηση αποτελεσμάτων .

Εισαγωγή
Η κατηγοριοποίηση έγινε με σκοπό, να βρεθεί η ακρίβεια με την οποία μπορούμε να προβλέψουμε σε ποία κατηγορία θα ανήκει η κάθε νέα είδηση που εισάγεται στο σύστημα.

Περιγραφή του συνόλου δεδομένων
Το σύνολο δεδομένων αφορά μια συλογή απο άρθρα που υπάρχει στο www.kaggle.com .Εμπεριέχει άρθρα απο πραγματικές ειδήσεις και σε κάθε άρθρο αντιστοιχεί ένας συγγραφέας ένα μοναδικό αναγνωριστικό (id), ο τίτλος του άρθρου και το περιεχόμενο (text) του άρθρου αυτού .Εκείνα τα άρθρα τα οποία χαρακτηρίζονται « αληθές » έχουν την τιμή στην στήλη label και αυτά που χαρακτηρίζονται ως « ψευδές » έχουν την τιμή 0 στην στήλη label.

Προ-επεξεργασία δεδομένων

Στην προ-επεξεργασία των δεδομένων μας μέσω της χρήσης βιβλιοθηκών της python αρχικά συγχωνεύσαμε τις στήλες author και title .Σε επόμενο στάδιο η στήλη label αποδώθηκε σε ξεχωριστή μεταβλητή απο τις υπόλοιπες στήλες .Ακόμη κάθε κείμενο υπήλθε απο πολλαπλές διαδικάσιες μετατροπής κειμένου δηλαδή διαδικασίες όπως αφαίρεσης σημείων στίξης ,μετατροπής των κεφαλαίων γραμμάτων σε πεζά γράμματα, γέμισαμε τα κενά με αντικατάστατο ενα κενό string και τέλος αφαιρέθηκε κάθε κενό που υπήρχε μεταξύ των λέξεων.

Αλγόριθμοι Κατηγοριοποίησης

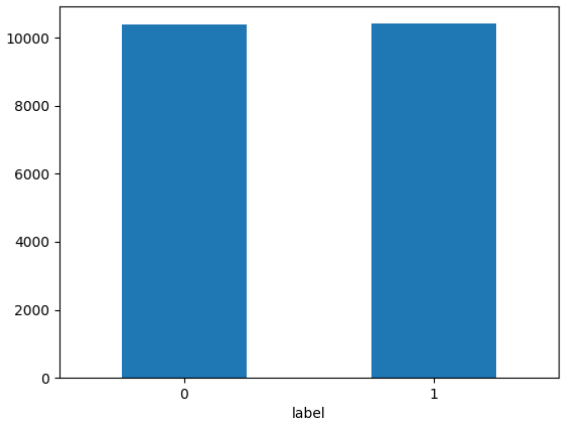
Λογιστική Παλινδρόμηση:Επειδή η κατηγοριοποιήση που θα κάνουμε είναι δυαδική(αληθής ή ψευδής είδηση),ταιριάζει απόλυτα να χρησιμοποιηθεί η συγκεκριμένη μέθοδος κατηγοριοποίησης. Αυτό διότι η λογιστική παλινδρόμηση παράγει αποτελέσματα σε κλίμακα απο το μηδέν έως το ένα.Κοντά στο μηδέν θεωρείται
«ψευδές» ενώ κοντά στο ένα «αληθές».

Μηχανές Διανυσμάτων Υποστήριξης:Παρόμοια με την Λογιστική Παλινδρόμηση, αυτός ο κατηγοριοποιητής,ορίζει το αποτέλεσμα ως «αληθές» αν ειναι κοντά στο 1 και «ψευδές» αν είναι κοντα στο -1.Επίσης το σύνολο δεδομένων είναι μικρό οπότε δεν θα υπάρχει πρόβλημα με την κλιμάκωση του μοντέλου.

Μεθοδολογία

Έγινε χρήση της βιβλιοθήκης Scikit-Learn και συγκεκριμένα της μεθόδου train_test_split,για να δημιουργήσουμε τύχαια σύνολα δοκιμής και εκπαίδευσης .Δηλαδή δημιουργήσαμε δύο σύνολα δεδομένων .Ύστερα τα δεδομένα των συνόλων αυτών παίρνουν την μορφή διανύσματος με την χρήση της μεθόδου TdifVectorizer που παρέχει η βιβλιοθήκη που αναφέρεται παραπάνω .Ακόμη δημιουργήθηκαν δυο βασικές μεθόδοι που αφορούν την λογιστική παλινδρόμηση και τον αλγόριθμο svm εντός των οποίων γίνεται η δήλωση του εκάστοτε μοντέλου που θα υλοποιηθεί .Επίσης εντός των συναρτήσεων γίνονται δοκιμές και στο σύνολο δοκιμής καθώς και εκπαίδευσης με σκοπό την αναγνώριση τυχών προβλήματος υπερπροσαρμογής .Δηλαδή εκτιμήθηκε τελικά εαν υπάρχει πρόβλημα υπερπροσαρμογής και στους δύο αλγορίθμους .Το ίδιο ακριβώς σκεπτικό υλοποίησης πραγματοποιήθηκε όσον αφορά την τυχαία επιλογή ενός σετ (π.χ. 1000) δειγμάτων .Όσον αφορά την χρονομέτρηση εκπαίδευσης του μοντέλου και την χρονομέτρηση υπολογισμού εκτίμησης ενός νέου κειμένου υπολογίστηκαν με την βοήθεια της βιβλιοθήκης time .Τέλος όσον αφορά την αξιολόγηση ενός κειμένου εισαγώμενο απο τον χρήστη έγινε η χρήση του πακέτου/βιβλιοθήκη flask δημιουργόντας μία σχετικά απλή εφαρμογή στο επίπεδο μιας ιστοσελίδας η οποία δέχεται απο το πληρκτρολόγιο ένα κείμενο σε μορφή text .Ύστερα το κείμενο αυτό υπόκειται σε προ-επεξαργασία δεδομένων όπως αυτή αναφέρθηκε προηγουμένος και μετά την εισαγωγή του στο μοντέλο αξιολογείται η αληθότητα αυτού.

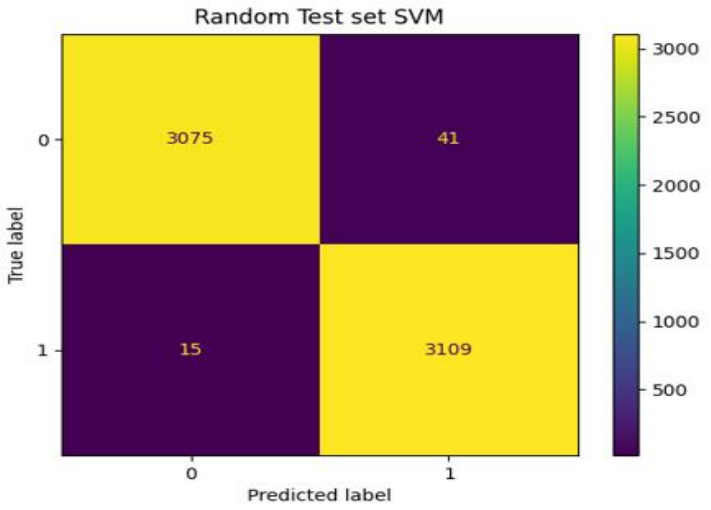
ΠειραματΙκή ΑξΙΟλόγηση



Λογιστική παλινδρόμηση (test-set)

0.9910256410256411

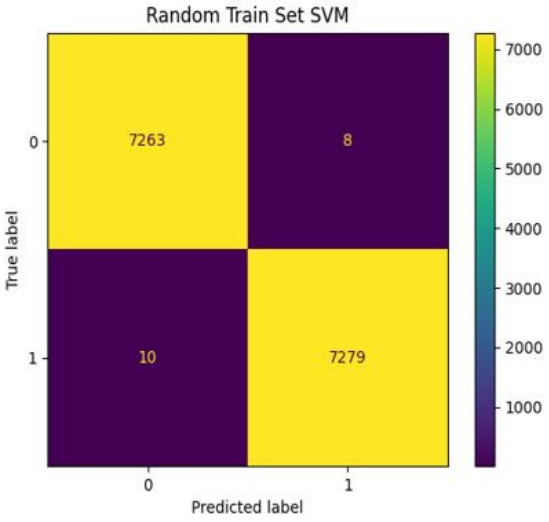
	precision	recall	f1-score	support
0	1.00	0.99	0.99	3116
1	0.99	1.00	0.99	3124
accuracy			0.99	6240
macro avg	0.99	0.99	0.99	6240
weighted avg	0.99	0.99	0.99	6240



Λογιστική Παλινδρόμηση(Train set)

0.9987637362637363

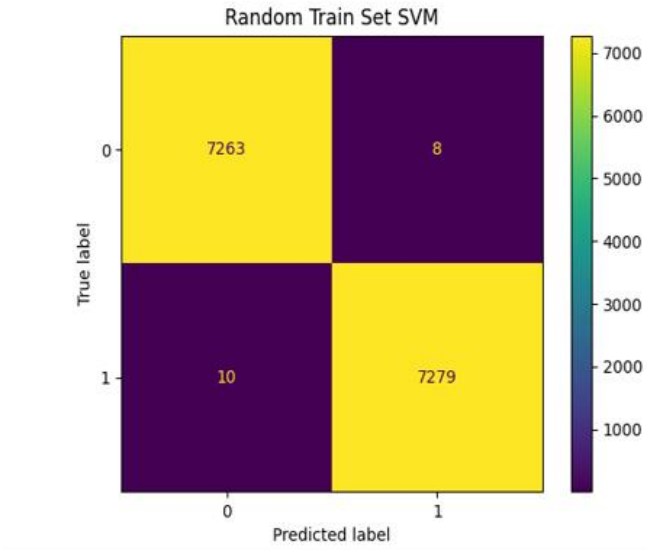
	precision	recall	f1-score	support
0	1.00	1.00	1.00	7271
1	1.00	1.00	1.00	7289
accuracy			1.00	14560
macro avg	1.00	1.00	1.00	14560
weighted avg	1.00	1.00	1.00	14560



Support Vector Machine (train-set)

0.9987637362637363

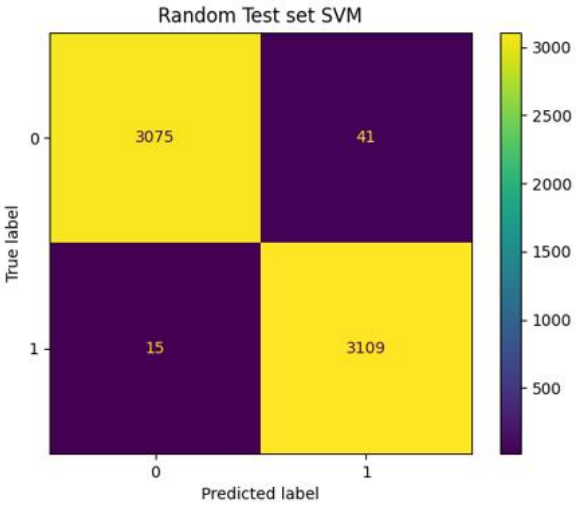
	precision	recall	f1-score	support
0	1.00	1.00	1.00	7271
1	1.00	1.00	1.00	7289
accuracy			1.00	14560
macro avg	1.00	1.00	1.00	14560
weighted avg	1.00	1.00	1.00	14560



Support Vector Machine (test-set)

0.9910256410256411

	precision	recall	f1-score	support
0	1.00	0.99	0.99	3116
1	0.99	1.00	0.99	3124
accuracy			0.99	6240
macro avg	0.99	0.99	0.99	6240
weighted avg	0.99	0.99	0.99	6240



Logistic Regression (1000)

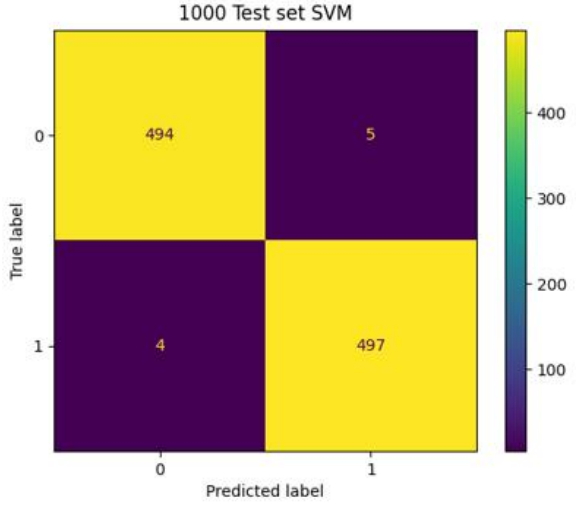
0.98

	precision	recall	f1-score	support
0	0.99	0.97	0.98	499
1	0.97	0.99	0.98	501
accuracy			0.98	1000
macro avg	0.98	0.98	0.98	1000
weighted avg	0.98	0.98	0.98	1000

Support Vector Machine (1000)

0.991

	precision	recall	f1-score	support
0	0.99	0.99	0.99	499
1	0.99	0.99	0.99	501
accuracy			0.99	1000
macro avg	0.99	0.99	0.99	1000
weighted avg	0.99	0.99	0.99	1000



REFERENCES

- https://flask.palletsprojects.com/en/2.0.x/api/
- https://scikit-learn.org/stable/