Υπολογιστική σύγκριση διάφορων αλγορίθμων κατηγοριοποίησης σε πολυμεταβλητά σύνολα δεδομένων με τη χρήση της Python

Φοιτητής: Τρίχας Θεόφιλος (mai19076)
Επιβλέπων καθηγητής: Σαμαράς Νικόλαος

Περιεχόμενα εργασίας

Εισαγωγή

Η **μηχανική μάθηση (Machine Learning)** είναι μία μελέτη αλγορίθμων υπολογιστών η οποία βελτιώνεται μέσα από την χρήση δεδομένων.

Οι **οικογένειες αλγορίθμων** είναι:

- Κατηγοριοποίηση (Classification)
- Συσταδοποίησης (Clustering)

Σκοπός της εργασίας είναι να συγκρίνει τους βασικότερους αλγόριθμους κατηγοριοποίησης σε σύνολα δεδομένων διαφόρων πολυπλοκότητας χρησιμοποιώντας ακρίβειες μέσα από τα διάφορα μοντέλα.

Κατηγοριοποίηση (Classification)

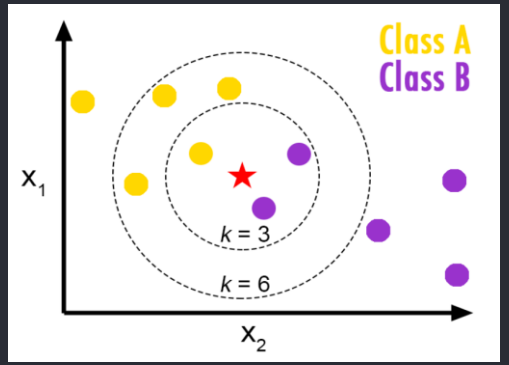
Οι παρακάτω αλγόριθμοι χρησιμοποιήθηκαν για την κατηγοριοποίηση των δεδομένων:

1. K-nearest Neighbors
2. Naïve Bayes
3. Logistic regression
4. Decision Tree (CART)

K-Nearest Neighbors

Ο αλγόριθμος K-NN είναι:

- Μία μη παραμετρική (non-parametric) μέθοδος που χρησιμοποιείται για την ταξινόμηση
- Από τους πιο δημοφιλείς αλγορίθμους ταξινόμησης
- Ικανός να αντιμετωπίσει προβλήματα που έχουν άγνωστες και μη κανονικές κατανομές.

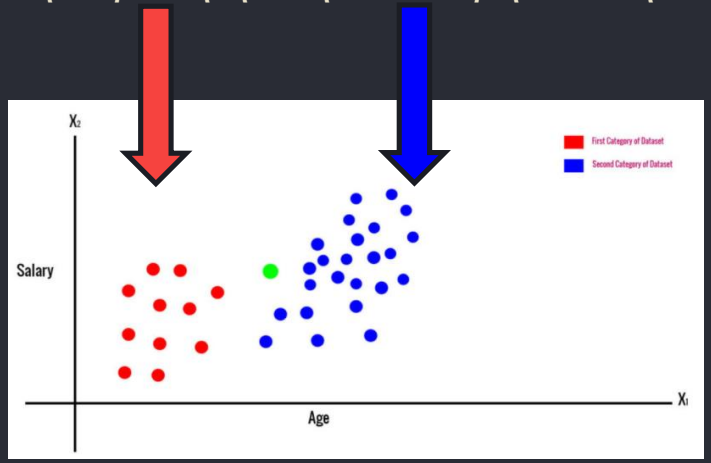


Ωστόσο, ο συγκεκριμένος αλγόριθμος χρειάζεται πάντα ένα μεγάλο σύνολο δεδομένων για να αποδώσει το μέγιστο δυνατό.

Naïve Bayes

Το Naïve Bayes είναι ένα μοντέλο πιθανότητας υπό όρους που χρησιμοποιεί το Bayes Θεώρημα για να υπολογίσει την πιθανότητα για κάθε ετικέτα κατηγορίας C_k.

Στην ουσία θέλουμε να υπολογίσουμε την πιθανότητα της τυχαίας πράσινης κουκίδας να ανήκει στην πρώτη ή στην δεύτερη κλάση κατηγορίας.



Logistic Regression

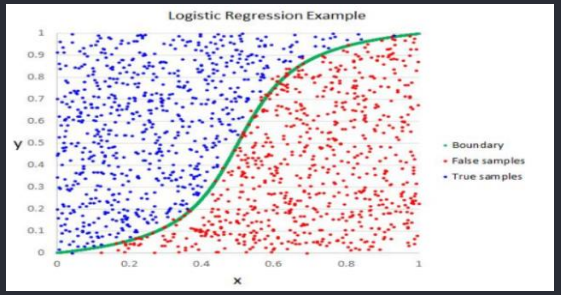
Στην στατιστική, το logistic μοντέλο χρησιμοποιείται για την μοντελοποίηση μίας συγκεκριμένης κλάσης ή ενός γεγονότος (πχ True ή False).

Δίνει πιθανότητα μεταξύ του 0 και 1.

Το logistic regression είναι ένα στατιστικό μοντέλο που μας δίνει μία τέτοια πιθανότητα παίρνοντας ένα μοντέλο και χρησιμοποιώντας μία λογιστική συνάρτηση, μοντελοποιεί μία δυαδική μεταβλητή.

$$\mathrm{l} = \log_b \frac{p}{p-1} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$$

- l είναι το log-odds
- b είναι η βάση του λογάριθμου
- β_i είναι οι παράμετροι του μοντέλου



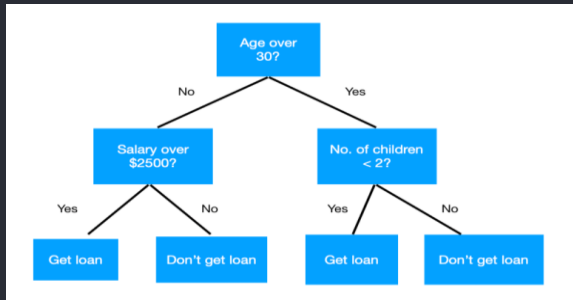
Decision Tree (CART)

Η εκμάθηση δέντρων αποφάσεων χρησιμοποιείται τόσο στην στατιστική όσο και στην εξόρυξη δεδομένων καθώς και στην μηχανική μάθηση.

Βοηθάει να ταξινομήσουμε τα δεδομένα μας

Τα μέτρα ποιότητας που χρησιμοποιούνται είναι συνήθως τα μέτρα entropy ή ο δείκτης Gini

Διαχωρίζει τα δεδομένα στη δυνατότητα που ελαχιστοποιεί το $G\ (N, \{j, t_N\})$ και συνεχίζει την ίδια διαδικασία αντίστοιχα για Q_{left} και Q_{right} έως ότου δεν υπάρχουν άλλα κριτήρια διαχωρισμού των δεδομένων



Συσταδοποίηση (Clustering)

Η συσταδοποίηση βοηθάει στο να χωριστούν τα δεδομένα μεταξύ τους σε κλάσεις, έχοντας κάποιες ομοιότητες και κάνοντάς τα να ξεχωρίζουν από τα υπόλοιπα σύνολα.

Μία μέθοδος συσταδοποίησης είναι καλή αν οι συστάδες που παράγει είναι καλής ποιότητας.

Ομοιότητα μέσα στην συστάδα

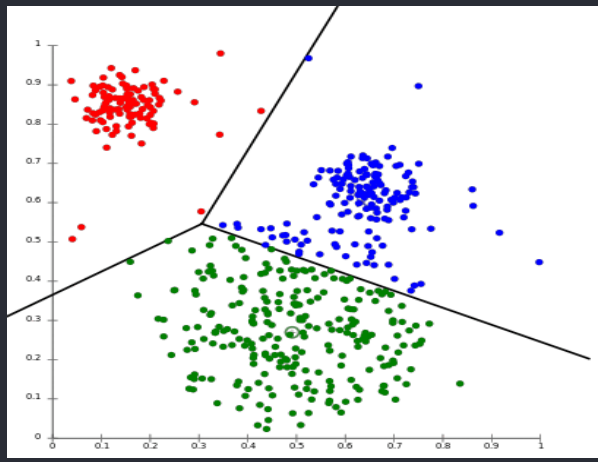
Σημαντικά μικρή ομοιότητα ανάμεσα στις συστάδες

K-Means Model

Ο K-Means αλγόριθμος προσπαθεί και στοχεύει να χωρίσει ένα σύνολο n παρατηρήσεων $(x_1, x_2, \ldots x_n)$ σε k σύνολα S = $\{S_1, S_2, \ldots, S_k\}$ όπου το k <= n έτσι ώστε να ελαχιστοποιηθεί το άθροισμα των τετραγώνων εντός τις συστάδας.

Τα αρχικά κέντρα επιλέγονται τυχαία.

Η εγγύτητα των σημείων υπολογίζεται με βάση κάποια απόσταση που εξαρτάται από το είδος των σημείων.



Data Sets

1. INTERNET FIREWALL DATA SET
2. TEACHING ASSISTANT EVALUATION DATA SET
3. CAR EVALUATION DATA SET
4. ELECTRICAL GRID STABILITY SIMULATED DATA SET

Internet Firewall Data Set

Αποτελείται από 65.532 εγγραφές οι οποίες περιγράφουν την κίνηση ανάμεσα σε ένα δίκτυο επικοινωνίας ενός Πανεπιστημίου.

Χωρίζονται σε 12 χαρακτηριστικά.

Η εξαρτημένη μεταβλητή επιτρέπει ή απαγορεύει την κίνηση δεδομένων στο δίκτυο.

Χρησιμοποιήθηκε ο αλγόριθμος Support Vector και προέκυψαν τα παρακάτω αποτελέσματα .

Αλγόριθμος	F1 Score	Precision	Recall
SVM Linear	75,4	67,5	85,3
SVM Linear	53,6	61,8	47,4
SVM RBF	76,4	63,0	97,1
SVM Sigmoid	74,8	60,3	98,5

Teaching Assistant Evaluation Data Set

Αποτελείται από 151 εγγραφές οι οποίες αξιολογούν διδακτικές αποδόσεις σε 5 εξάμηνα ενός Πανεπιστημίου.

Χωρίζονται σε 6 χαρακτηριστικά.

Η εξαρτημένη μεταβλητή δείχνει το χαρακτηριστικό κάθε τάξης.

Χρησιμοποιήθηκαν 33 αλγόριθμοι και η σύγκριση έγινε με ένα μέσο όρο του Error rate.

	Error rates
Ελάχιστο όριο	0,33
Μέγιστο όριο	0,66

Car Evaluation Data Set

Αποτελείται από 1.728 εγγραφές οι οποίες με βάση 6 χαρακτηριστικά αξιολογούν ένα αυτοκίνητο.

Η εξαρτημένη μεταβλητή δείχνει την κλάση του αυτοκινήτου.

Χρησιμοποιήθηκε ο αλγόριθμος Naive-Bayes και προέκυψαν τα παρακάτω αποτελέσματα .

	Accuracy			
Dataset/Algorithm	Naïve-Bayes	BN Augmented	TREE Augmented	General BNs
Car Evaluation	86,58	94,04	94,10	86,11

Electrical Grid Stability Simulated Data Set

Αποτελείται από 10.000 εγγραφές οι οποίες περιγράφουν την σταθερότητα της ηλεκτρικής ενέργειας.

Χωρίζονται σε 12 χαρακτηριστικά.

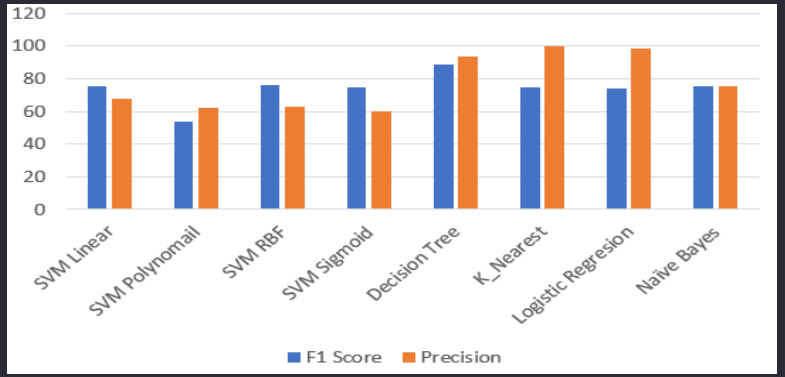
Η εξαρτημένη μεταβλητή δείχνει την σταθερότητα του συστήματος.

Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο CART και η ακρίβεια που υπήρξε στα πειράματα ήταν 80%.

Αποτελέσματα έρευνας

Στα προϋπάρχοντα αποτελέσματα του Internet Firewall Dataset προστέθηκαν τα εξής:

Αλγόριθμος	F1 Score	Precision	Accuracy
K-Nearest	75,0	99,7	99,67
Naïve Bayes	75,2	75,1	99,1
Logistic Regresion	73,8	98,6	98,5
Decision Tree	88,4	93,5	99,7



Ο πιο αποδοτικός ήταν ο decision tree αλγόριθμος όπου το F1 Score συγκριτικά με τους υπόλοιπους είχε αρκετή διαφορά της τάξης του 13%.

Σημαντικός ρόλος → Μεγάλο δείγμα δεδομένων.

Αποτελέσματα έρευνας

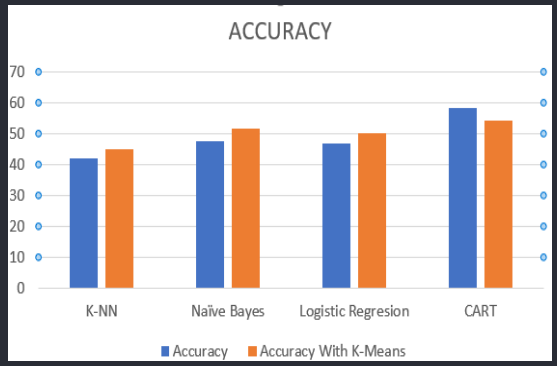
Στα προϋπάρχοντα αποτελέσματα του Teaching Assistant Evaluation Data Set προστέθηκαν τα εξής:

Συνολικά αποτελέσματα

Αλγόριθμος	Min Error	Max Error
K-NN	0,33	0,75
Naïve Bayes	0,25	0,75
Logistic Regression	0,33	0,66
CART	0,25	0,66

Συνολικά αποτελέσματα με K-Means

Αλγόριθμος	Min Error	Max Error
K-NN	0,33	0,83
Naïve Bayes	0,16	0,66
Logistic Regression	0,25	0,75
CART	0,25	0,66

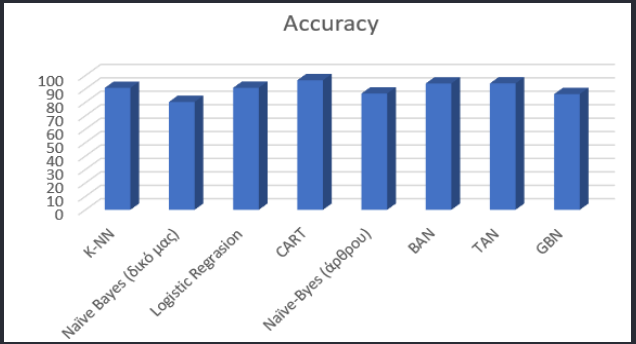


Το μικρότερο ποσοστό error rate στην δικιά μας έρευνα το πετυχαίνει ο naïve bayes αλγόριθμος.

Αποτελέσματα έρευνας

Στα προϋπάρχοντα αποτελέσματα του Car Evaluation Data Set προστέθηκαν τα εξής:

Accuracy				
Αλγόριθμος	K-NN	Naïve-Bayes	Logistic Regression	CART
Data Set	90,73	80,30	90,80	96,45

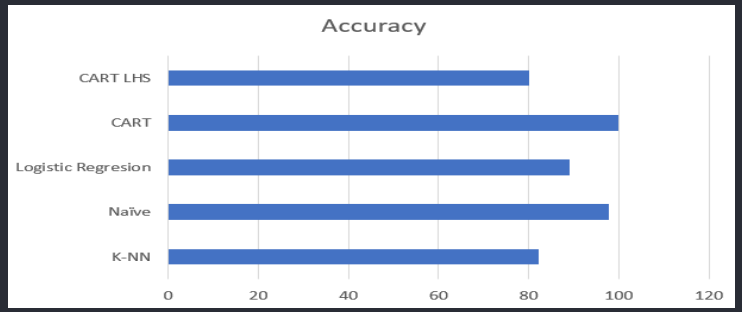


Ο CART αλγόριθμος είχε την καλύτερη ακρίβεια.

Αποτελέσματα έρευνας

Στα προϋπάρχοντα αποτελέσματα του Electrical Grid Stability Data Set προστέθηκαν τα εξής:

Accuracy				
Dataset/Algorithm	K-NN	Naïve Bayes	Logistic Regression	CART
Electrical Grid Stability	82,15	97,92	89,14	99,98

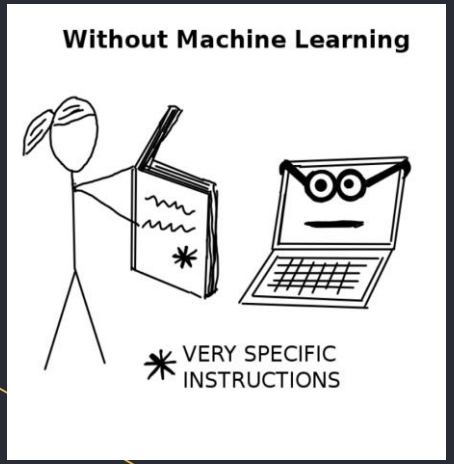


Υψηλές ακρίβειες με όλους τους αλγορίθμους.

Καλύτερος ξανά ο CART αλγόριθμος.

Συμπεράσματα

- Βελτίωση αποτελεσμάτων συγκριτικά με τις αντίστοιχες προηγούμενες έρευνες.
- Εν αντιθέσει με τους υπόλοιπους αλγορίθμους στον Teaching Assistant Evaluation Data Set η βελτίωση ήταν ελάχιστη → Μικρό δείγμα δεδομένων.
- CART αλγόριθμος καλύτερη εκπαίδευση δεδομένων.
- Χώρος βελτίωσης των αποτελεσμάτων.
- Δεν υπάρχει οδηγία για το ποιος αλγόριθμος αποδίδει καλύτερα στα δεδομένα που έχει κάποιος → Δοκιμές πάνω στα δεδομένα και σύγκριση αυτών.



"Computers are able to see, hear and learn. Welcome to the future."

~Dave Waters