ΔΗΜΙΟΥΡΓΙΑ ΜΟΝΤΕΛΩΝ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΤΟΥ ΑΡΙΘΜΟΥ ΚΡΟΥΣΜΑΤΩΝ, ΘΑΝΑΤΩΝ ΚΑΙ ΔΙΑΣΩΛΗΝΩΜΕΝΩΝ ΑΣΘΕΝΩΝ ΛΟΓΩ COVID-19 ΜΕ ΤΗ ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗΣ

Ματθαίος Γκόγκος

Επιβλέπων Καθηγητής: Νικόλαος Σαμαράς

Σκοπός εργασίας

Σχηματισμός αξιόπιστης πρόβλεψης του αριθμού κρουσμάτων, θανάτων και διασωληνωμένων ασθενών λόγω Covid-19.

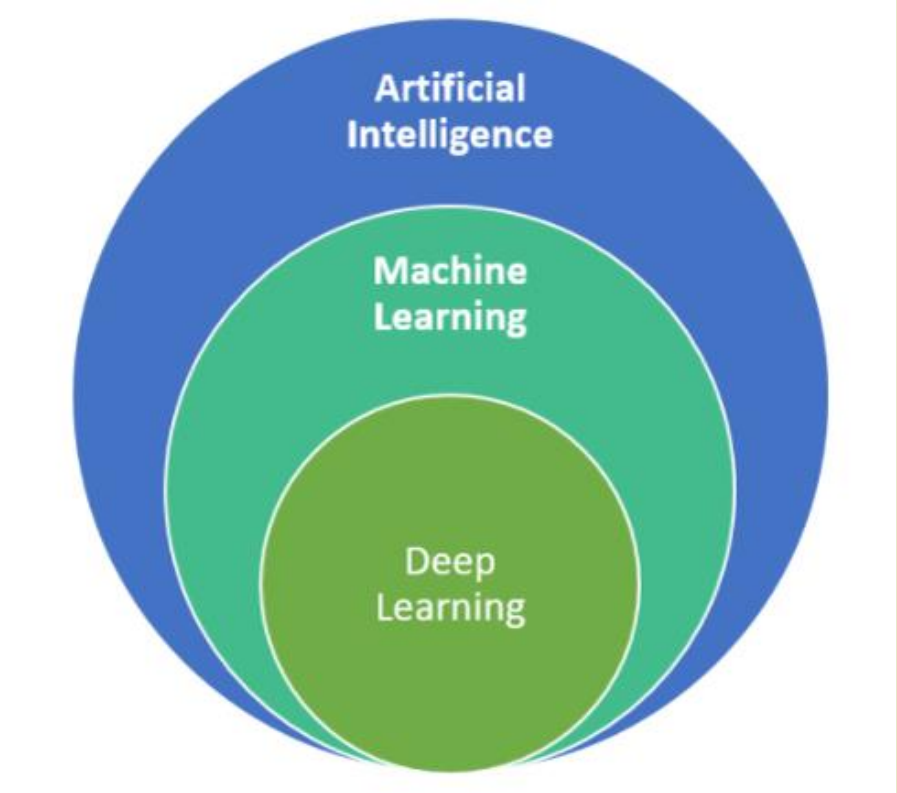
- Πολλαπλή Γραμμική Παλινδρόμηση
- Πολυωνυμική Παλινδρόμηση
- Παλινδρόμηση Διανυσμάτων Υποστήριξης
- Δέντρα Απόφασης (C4.5)
- Δέντρα Απόφασης (C5.0)

Μηχανική Μάθηση

Αντικείμενο της μηχανικής μάθησης είναι η μελέτη και κατασκευή αλγορίθμων που εκπαιδεύονται με τη χρήση δεδομένων, αποκτούν γνώση και κάνουν προβλέψεις σχετικά με αυτά.

Υπάρχουν 3 βασικές προσεγγίσεις των αλγορίθμων μηχανικής μάθησης:

- Εποπτευόμενη Μάθηση(Supervised Learning)
- Μη εποπτευόμενη Μάθηση (Unsupervised Learning)
- Ενισχυτική Μάθηση (Reinforcement Learning)



Εποπτευόμενη Μάθηση

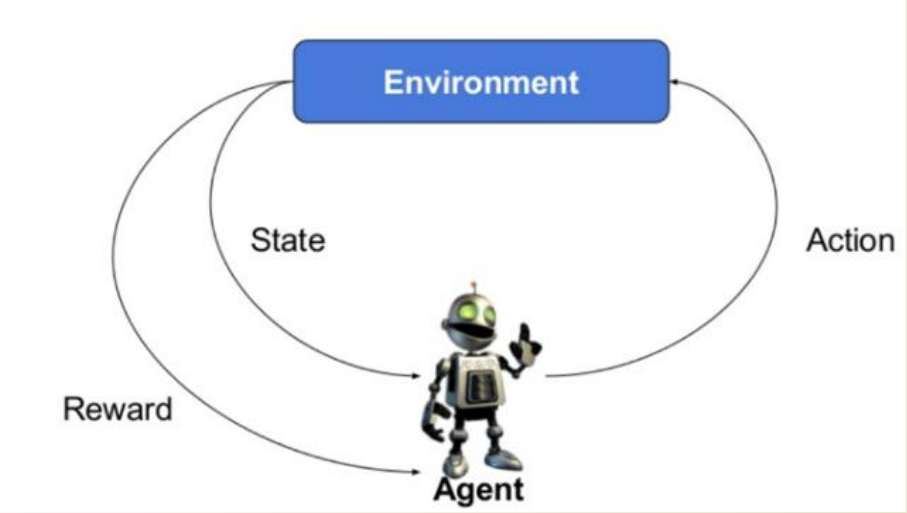
- Παροχή πληροφορίας στον αλγόριθμο
- Αντιστοίχιση μεταβλητών εισόδου με την επιθυμητή μεταβλητή εξόδου
- Εκμάθηση γενικών κανόνων και επέκταση γνώσεων
- Η εποπτευόμενη εκμάθηση θα μπορούσε να χωριστεί σε δύο κατηγορίες, την ταξινόμηση (classification) και την παλινδρόμηση (regression):
  - Οι αλγόριθμοι ταξινόμησης έχουν ως μεταβλητή εξόδου μια διακριτή μεταβλητή
  - Οι αλγόριθμοι παλινδρόμησης έχουν ως μεταβλητή εξόδου μια συνεχή μεταβλητή

Μη Εποπτευόμενη Μάθηση

- Δεν παρέχεται πληροφορία σχετικά με το τι πρέπει να αναγνωρίσει ο αλγόριθμος
- Εύρεση κρυμμένων μοτίβων
- Ομαδοποίηση δεδομένων με βάση τα κοινά τους χαρακτηριστικά

Ενισχυτική Μάθηση

- Περιβάλλον καταστάσεων που αλληλεπιδρά με την μηχανή με σκοπό την επίτευξη ενός στόχου.
- Η μηχανή δέχεται συνεχή ανατροφοδότηση και βελτιώνει την εμπειρία της.



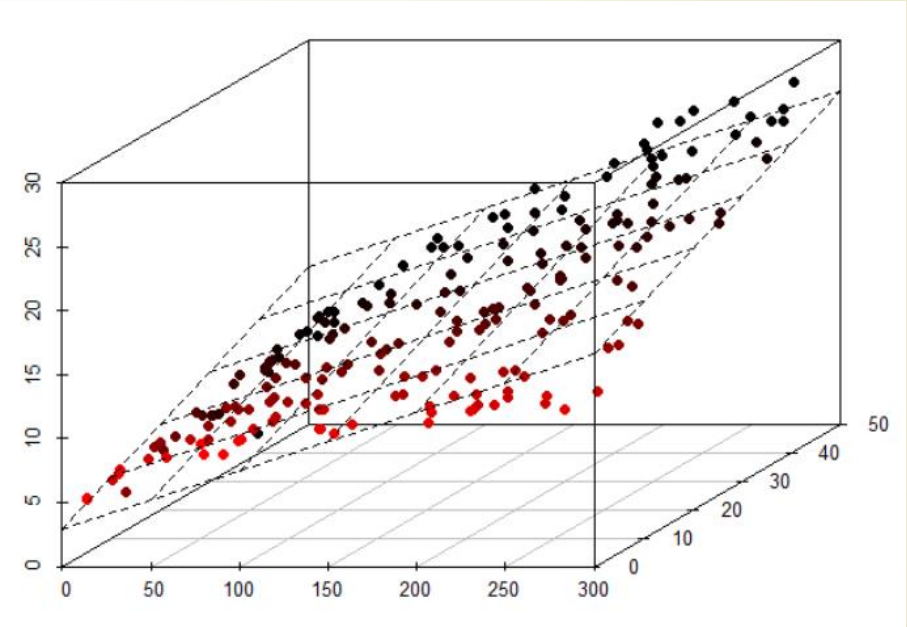
Γραμμική Παλινδρόμηση

- Στην στατιστική, η γραμμική παλινδρόμηση αποτελεί μία προσέγγιση για τη μοντελοποίηση της σχέσης μεταξύ μίας μεταβλητής απόκρισης και μίας ή περισσότερων επεξηγηματικών μεταβλητών, όπου η μεταβλητή απόκρισης αποτελεί την εξαρτημένη μεταβλητή του συστήματος, ενώ η επεξηγηματική μεταβλητή αποτελεί την ανεξάρτητη.
- Ο στόχος ενός μοντέλου γραμμικής παλινδρόμησης είναι η πρόβλεψη των μελλοντικών τιμών με το μικρότερο δυνατό σφάλμα.

Πολλαπλή Γραμμική Παλινδρόμηση

Το μοντέλο πολλαπλής γραμμικής παλινδρόμησης εκφράζεται ως εξής:

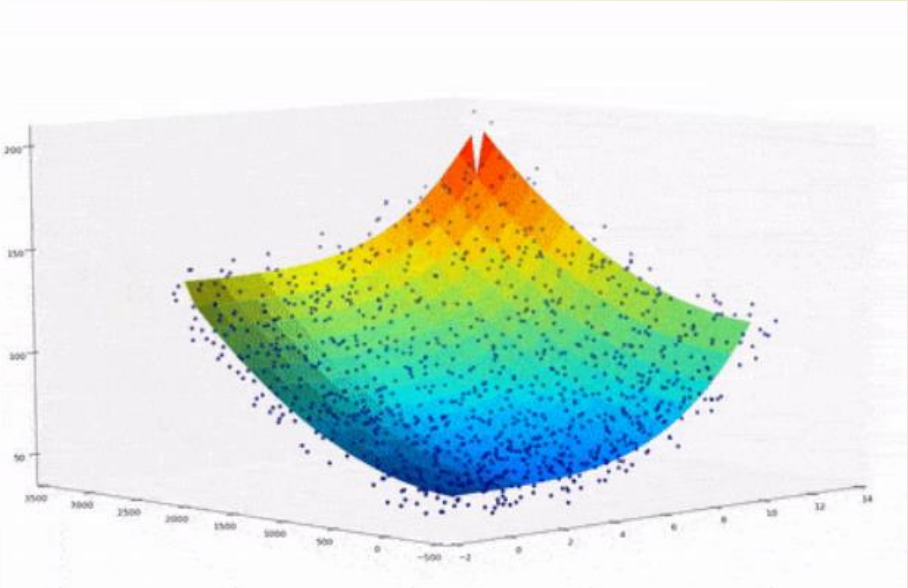
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_k$$



Πολυωνυμική Παλινδρόμηση

Το μοντέλο πολυωνυμικής γραμμικής παλινδρόμησης βαθμού k δίνεται από την εξίσωση

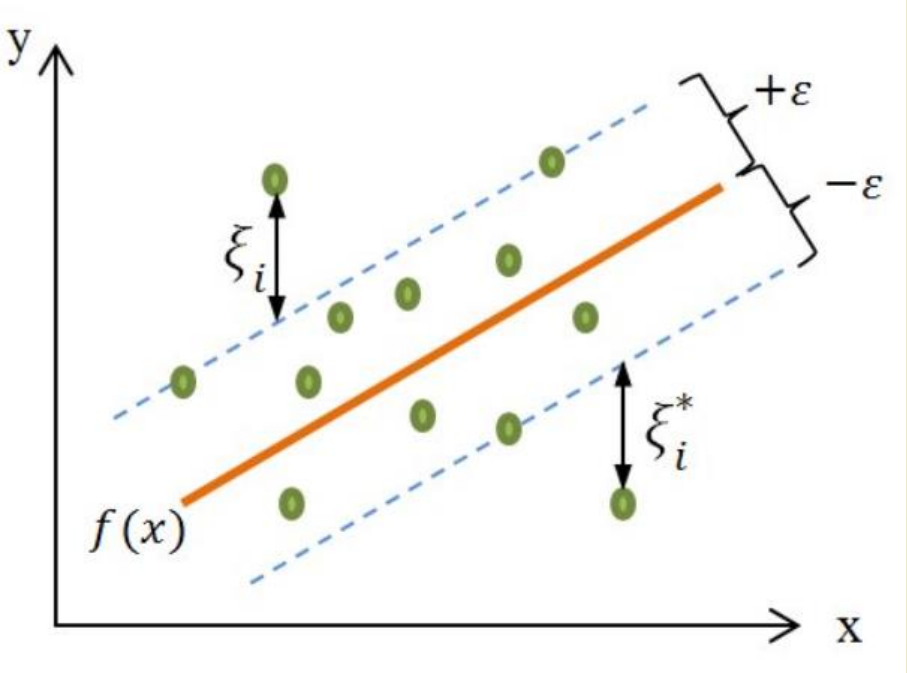
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon_k$$



Παλινδρόμηση Διανυσμάτων Υποστήριξης

- Σκοπός είναι η ελαχιστοποίηση του σφάλματος πρόβλεψης της μεταβλητής στόχου
- Καθορισμός μέγιστης επιτρεπόμενης απόκλισης

$$\min \quad \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{l}(\xi_i - \xi_i *)$$



Δέντρο Απόφασης (C4.5)

- Εξέλιξη του αλγορίθμου ID3
- Διαχειρίζεται διακριτές και συνεχείς μεταβλητές
- Διαχειρίζεται ελλιπείς τιμές
- Εφαρμόζει κλάδεμα
- Κριτήριο διαχωρισμού είναι το κανονικοποιημένο κέρδος πληροφορίας

Δέντρο Απόφασης (C5.0)

- Συμβατός με νέους τύπους δεδομένων
- Κόστος λανθασμένης ταξινόμησης μιας μεταβλητής
- Δυνατότητα εφαρμογής της μεθόδου winnow
- Δυνατότητα boosting
- Κριτήριο διαχωρισμού είναι το κανονικοποιημένο κέρδος πληροφορίας

Σύνολο Δεδομένων

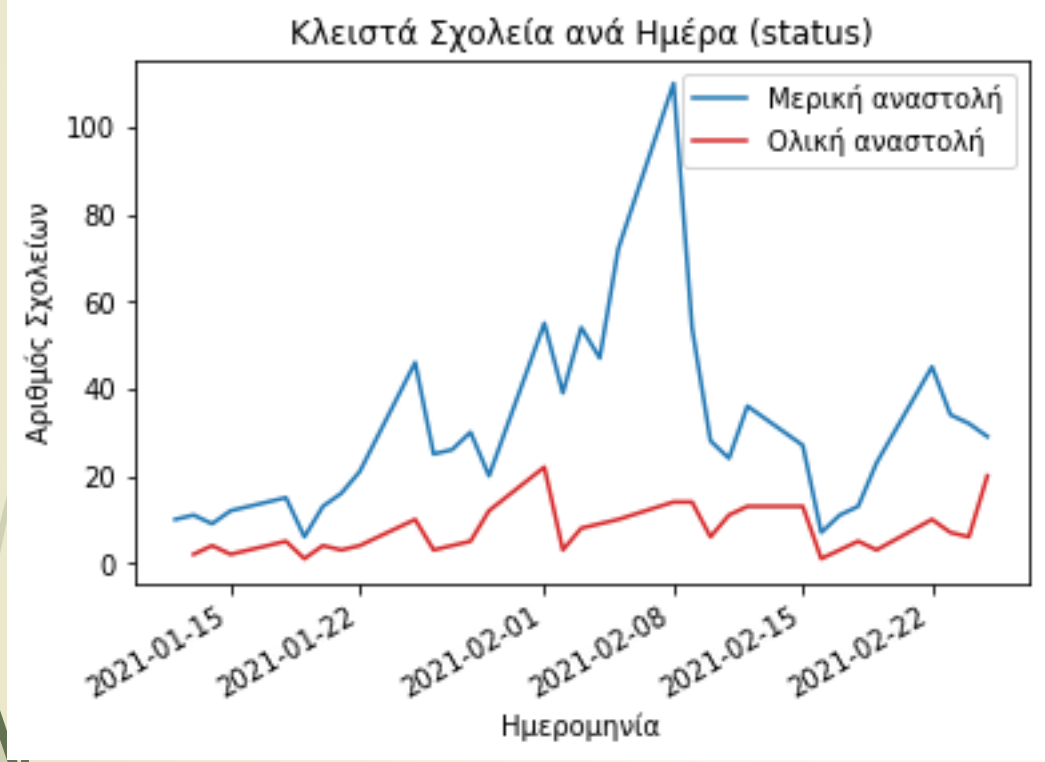
- Συλλογή δεδομένων σχετικά με το κλείσιμο των σχολείων
- Συνένωση με άλλα σύνολα δεδομένων που σχετίζονται με ημερήσια κρούσματα, θανάτους και διασωληνωμένους ασθενείς
- Διαμόρφωση των δεδομένων ώστε να είναι κατάλληλα για την τροφοδότηση των μοντέλων

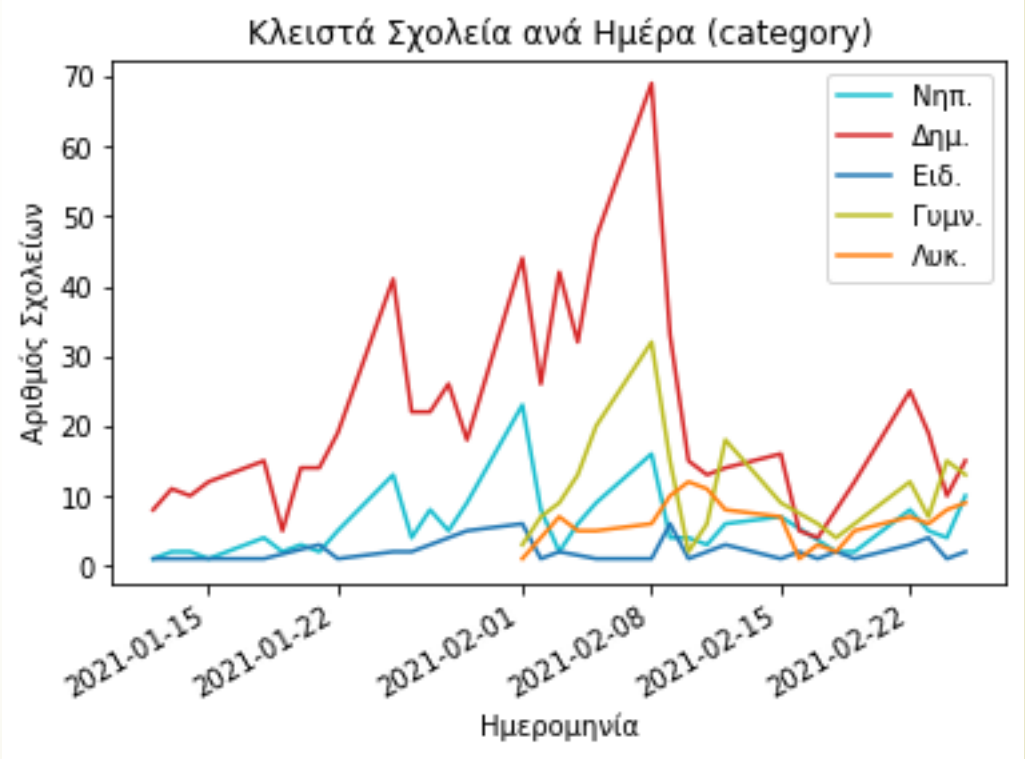
Σύνολο Δεδομένων

- Στο σύνολο των 1237 σχολείων, 1000 σχολεία ανέστειλαν μερικώς τη λειτουργία τους και 237 ολικώς.

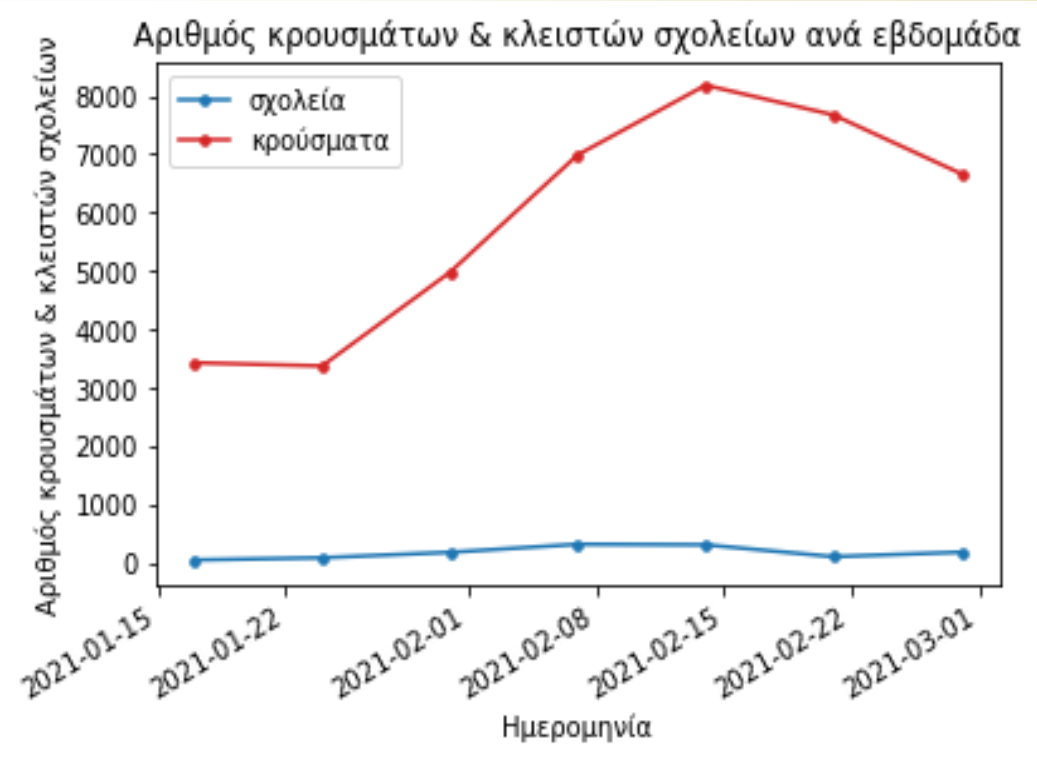
category	
Δημοτικά	686
Γυμνάσια	197
Νηπιαγωγεία	180
Λύκεια	113
Ειδικά σχολεία	61

Σύνολο Δεδομένων

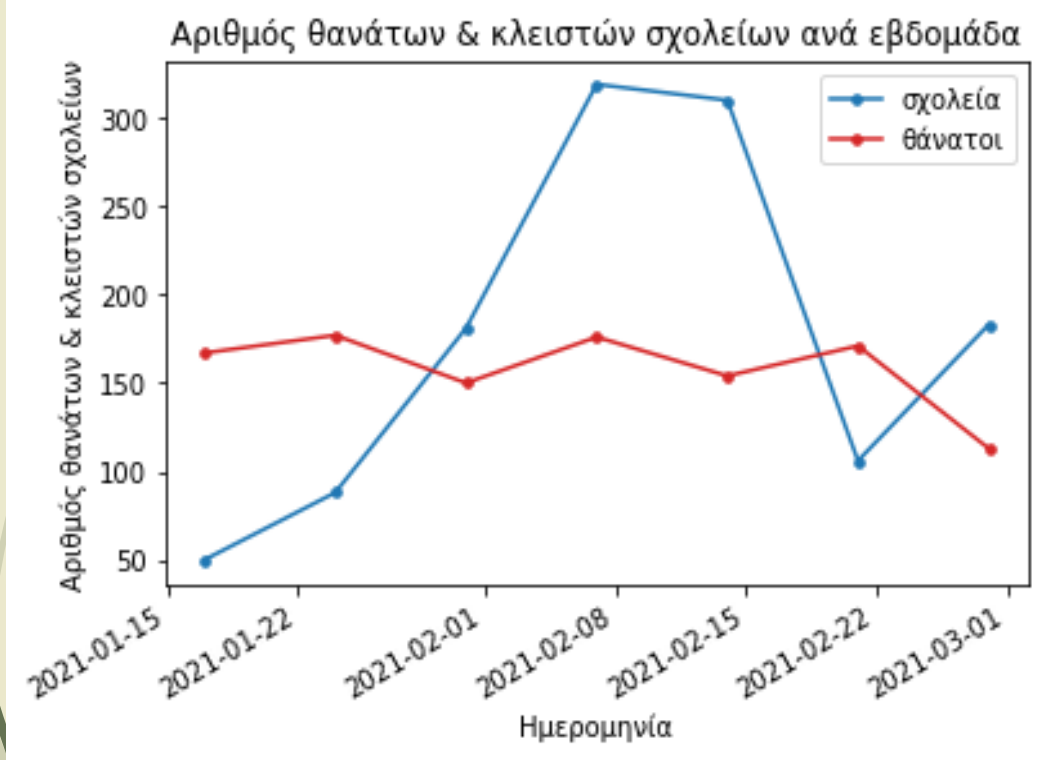


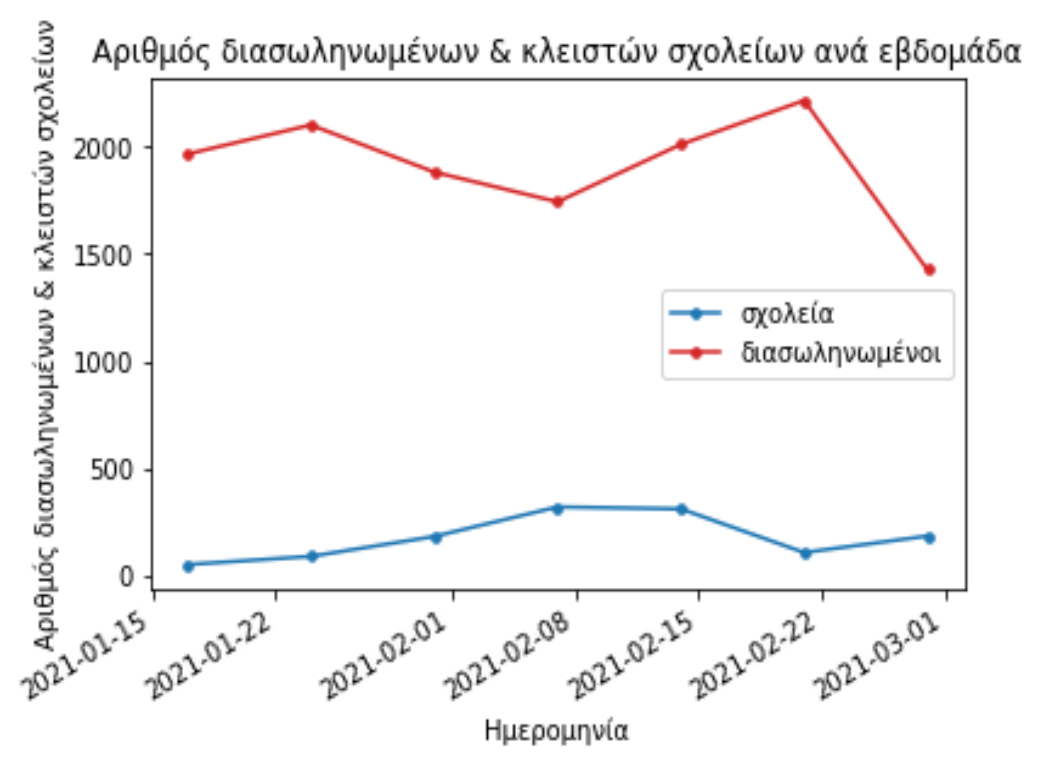


Σύνολο Δεδομένων



Σύνολο Δεδομένων





Αποτελέσματα

- Μοντελοποίηση αριθμού κρουσμάτων

	Accuracy (%)
Multiple Linear Regression	74.9
Polynomial Regression	98.7
Support Vector Regression	98.7
C4.5	88.7
C5.0	95.1

Αποτελέσματα

- Μοντελοποίηση αριθμού θανάτων

	Accuracy (%)
Multiple Linear Regression	52.2
Polynomial Regression	95.2
Support Vector Regression	89.8
C4.5	89.1
C5.0	95.1

Αποτελέσματα

- Μοντελοποίηση αριθμού διασωληνωμένων ασθενών

	Accuracy (%)
Multiple Linear Regression	83.5
Polynomial Regression	98.9
Support Vector Regression	98.9
C4.5	81
C5.0	95.9

Συμπεράσματα

- Ο Multiple Linear Regression είναι ο λιγότερο αποδοτικός αλγόριθμος
- Polynomial Regression, Support Vector Regression και C5.0 εξίσου αποδοτικοί σε όλα τα προβλήματα
- Ο C5.0 αποδίδει καλύτερα από τον C4.5
- Δεν κατέστη εφικτή η εξαγωγή συμπερασμάτων για την συμβολή της κάθε μεταβλητής στο εκάστοτε πρόβλημα

Ευχαριστώ πολύ!