Πρόβλεψη μετεωρολογικών συνθηκών με τη χρήση αλγορίθμων μηχανικής μάθησης

Χρήστος Θεοδωρόπουλος
Μεγάλα Δεδομένα & Αναλυτική
Τμήμα Ψηφιακών Συστημάτων
Πανεπιστήμιο Πειραιώς
C.Theodoropoulos92@gmail.com

Περίληψη:
Τα τελευταία χρόνια χαρακτηρίζονται ως η εποχή των μεγάλων δεδομένων εξαιτίας την αύξησης των δεδομένων που παράγονται καθημερινά καθώς και λόγο της εφαρμογής αλγορίθμων μηχανικής μάθησης με ισχυρότερα υπολογιστικά συστήματα. Η μετεωρολογία και η πρόβλεψη των μετεωρολογικών συνθηκών πραγματοποιούνταν παραδοσιακά με τη προσομοίωση σύνθετων μοντέλων φυσικής. Εκμεταλλευόμενοι την ευρεία γκάμα εφαρμογής των αλγορίθμων μηχανικής μάθησης θα εξετάσουμε κατά πόσο είναι εφικτό να προβλέψουν επιτυχώς με υψηλή ακρίβεια μετεωρολογικά δεδομένα χωρίς να χρησιμοποιήσουμε καμία γνώση από την επιστήμη της Φυσικής. Τα συστήματα θα υλοποιηθούν στη γλώσσα προγραμματισμού Python.

Keywords: machine learning algorithm, Python,
μετεωρολογικά δεδομένα, πρόβλεψη μετεωρολογικών συνθηκών

1 Εισαγωγή

Με τον όρο πρόβλεψη μετεωρολογικών συνθηκών (weather forecasting) εννοούμε την πρόβλεψη των καιρικών δεικτών σε κάποια μελλοντική στιγμή και σε συγκεκριμένη τοποθεσία. Η πρόβλεψη των καιρικών συνθηκών αποτελεί κομμάτι του κλάδου της μετεωρολογίας της Φυσικής επιστήμης και βασίζεται στη προσομοίωση μοντέλων φυσικής για την εξαγωγή προβλέψεων. Πλέον οι μετεωρολογικοί σταθμοί έχουν πρόσβαση στο ιντερνέτ και με τη χρήση του IoT (Internet of Things) αλλά και των μεθόδων ML τείνουν να αντικαταστήσουν τις παραδοσιακές μεθόδους πρόβλεψης. Πλέον πολλοί μετεωρολογικοί οργανισμοί παρέχουν ευρεία πρόσβαση στις βάσεις δεδομένων τους. Αξιοποιώντας αυτά τα μεγάλα δεδομένα με προηγούμενες μετρήσεις αλλά και real time δεδομένα γίνεται πολύ προσιτή η χρήση της προβλεπτικής αναλυτικής.

Στη εργασία χρησιμοποιήθηκαν μερικές απλές μέθοδοι προβλεπτικής αναλυτικής όπως η πολλαπλή γραμμική παλινδρόμηση, η random forest, naïve bayes classifier και SVM.

Τα δεδομένα που χρησιμοποιήθηκαν προήλθαν από τον κόμβο meteo.gr, του Εθνικού Αστεροσκοπείου Αθηνών και περιλάμβαναν μετρήσεις ανά μήνα μέγιστης-ελάχιστης θερμοκρασίας, ύψος βροχόπτωσης, Βαθμοημέρες Θέρμανσης-ψύξης, μέση ταχύτητα ανέμου και μέγιστη ταχύτητα ανέμου στην περιοχή της Αθήνας για συγκεκριμένα χρονικά διαστήματα.

2 Μέθοδοι και Δεδομένα

2.1 Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν στην εργασία προήλθαν από προήλθαν από τον κόμβο meteo.gr, του Εθνικού Αστεροσκοπείου Αθηνών και αποτελούνταν από 2 αρχεία.

2.1.1 athens_09-16.dat. Το αρχείο αυτό περιελάμβανε μετρήσεις για το κέντρο της Αθήνας για τα έτη 2009-2016. Αποτελούνταν από 8 κατηγορίες μετρήσεων (Μήνας έτους, Μέγιστη Θερμοκρασία, Ελάχιστη Θερμοκρασία, Ύψος βροχόπτωσης, Βαθμοημέρες Θέρμανσης, Βαθμοημέρες ψύξης, Μέση ταχύτητα ανέμου, Μέγιστη ταχύτητα ανέμου)
2.1.2 athens_2017.dat. Αντίστοιχα το αρχείο αυτό περιελάμβανε τις ίδιες μετρήσεις για το έτος 2017 στο κέντρο της Αθήνας.

1. Μήνας του έτους $(1, 2, \ldots, 12)$.
2. Μέγιστη Τιμή Θερμοκρασίας.
3. Ελάχιστη Τιμή Θερμοκρασίας.
4. Βαθμοημέρες Θέρμανσης.
5. Βαθμοημέρες Ψύξης.
6. Ύψος Βροχόπτωσης.
7. Μέση Ταχύτητα Ανέμου.
8. Μέγιστη Ταχύτητα Ανέμου.

2.2 Μέθοδοι

Ο πρώτος αλγόριθμος που χρησιμοποιήθηκε είναι η πολλαπλή γραμμική παλινδρόμηση. Στόχος ήταν να βρεθούν οι τιμές της μέγιστης Θερμοκρασίας και του Ύψους της βροχόπτωσης ως γραμμικός συνδυασμός των υπολοίπων μεταβλητών εκτός του Μήνα του Έτους και της Ελάχιστης Θερμοκρασίας. Η συγκεκριμένη μέθοδος επιλέχθηκε διότι μέσω των scatter plot βρέθηκε ισχυρή γραμμική συσχέτιση (θετική - αρνητική) μεταξύ των μεταβλητών και λόγο της απλότητας της.

The Formula for Multiple Linear Regression Is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

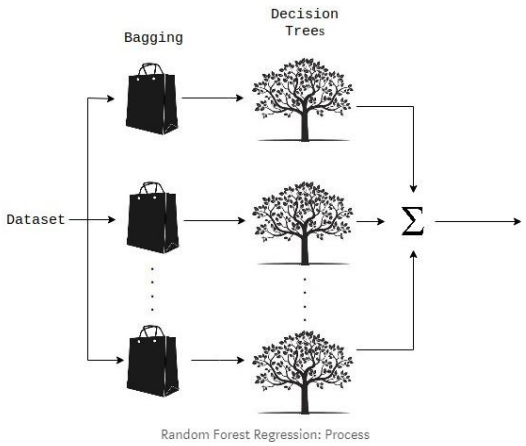
x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

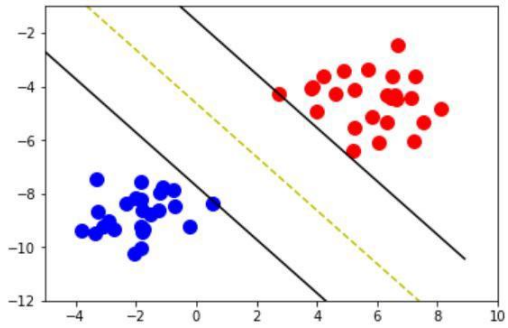
ϵ = the model's error term (also known as the residuals

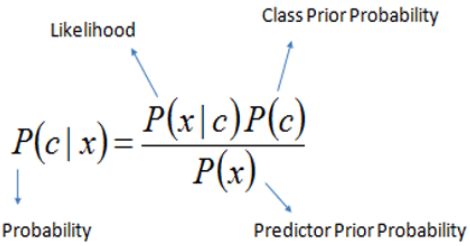
Η δεύτερη μέθοδος ήταν η RandomForestRegressor η οποία χρησιμοποιείται εξίσου και για regression αλλά και classification χρησιμοποιώντας πολλαπλά δεντρα απόφασης . Η μη ύπαρξη γραμμικής συσχέτισης μεταξύ κάποιων δεδομένων κατέστησε τη χρήση της παραπάνω μεθόδου ως ιδανική επιλογή.



Random Forest Regression: Process

Η τρίτη και η τέταρτη μέθοδοι αντίστοιχα ήταν οι SVM και naïve bayes. Η ανάγκη για επίλυση ενός προβλήματος classification παρουσιάστηκε στο τρίτο ερώτημα της εργασίας με στόχο την πρόβλεψη της εποχής (winter-autumn-summer-spring). Η δυνατότητα τους να χειριστούν multiple classification καθώς και η απλότητα της κατασκευής των μοντέλων μας οδήγησε στην επιλογή τους. Αξιοποιώντας την στήλη με τους Μήνες δημιουργήσαμε μια ακόμα στήλη με το label της εποχής για κάθε γραμμή και χρησιμοποιήθηκε στους κώδικες .

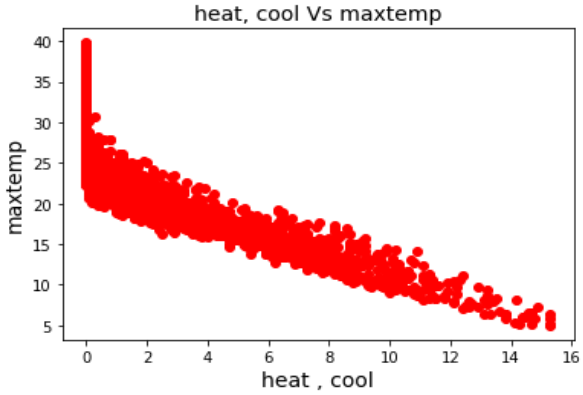




$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

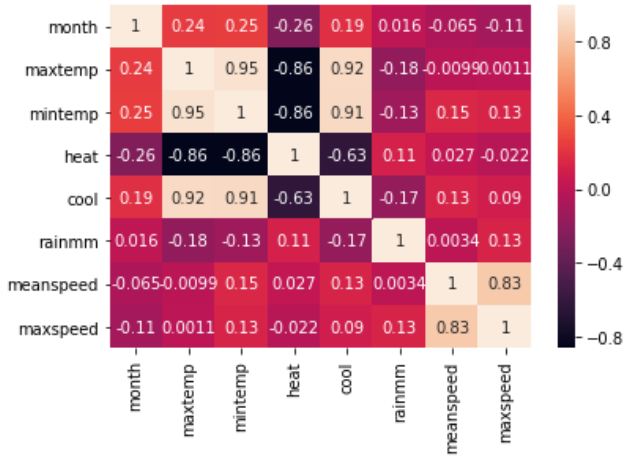
3 Αποτελέσματα

Στο πρώτο ερώτημα εξετάσαμε την πρόβλεψη της μέγιστης θερμοκρασίας. Αρχικά τα δεδομένα ελέχθησαν για τυχόν κενές τιμές που θα αλλοίωναν το αποτέλεσμα. Προκειμένου να εξετάσουμε εάν εφαρμόζεται η μέθοδος της γραμμικής παλινδρόμησης κατασκευάστηκαν τα διαγράμματα διασποράς μεταξύ της εξεταζόμενης μεταβλητής και των υπολοιπων και διαπιστώθηκε γραμμική συσχέτιση. Ενδεικτικά



Εικόνα 1: scatter plot maxtemp-heat

Επίσης βρέθηκαν οι συντελεστές συσχέτισης με τιμές διαφορετικές του 0 που δείχνουν τη γραμμική συσχέτιση των μεταβλητών.(κοντά στο -1, 1 ισχυρή συσχέτιση)

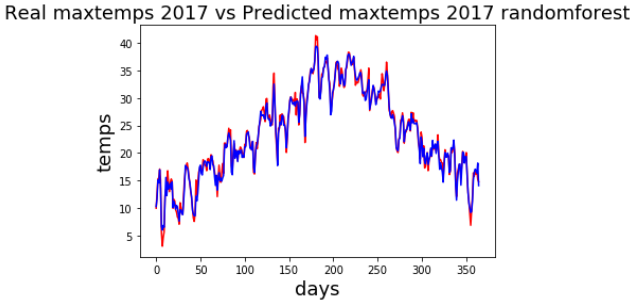


Εικόνα 2:συντελεστής συσχέτισης

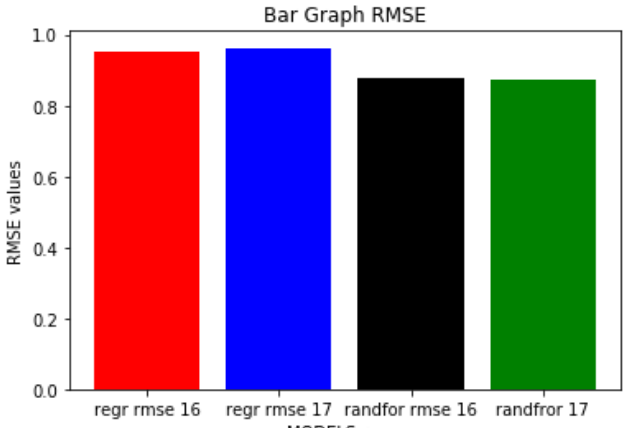
Υλοποιήθηκαν οι αλγόριθμοι της πολλαπλής γραμμ. Παλινδρόμησης και του randomforest. Τα δεδομένα του 2009-2016 χωρίστηκαν σε 70% training set – 30% test set και αφού εκπαιδεύτηκαν τα μοντέλα, τους τροφοδοτήσαμε τα δεδομένα του 2017 και είχαμε τα παρακάτω αποτελέσματα.

```
Intercept:
 22.939879219668402
Coefficients:
 [-1.04269400e+00  1.08873692e+00 -3.17925462e-02 -1.89447032e-01
 -9.99029380e-04]
      Real 2016  Predicted 2016
209        10.6       12.854499
2354       15.1       14.780218
235        15.3       14.170063
2137       16.4       14.914283
1504       21.3       20.785753
479        36.1       35.484787
1827       31.7       31.334567
1774       17.6       17.249261
1833       35.7       35.909990
877        32.8       32.998139
2205       29.4       29.757266
773        24.6       24.663270
1754       25.1       25.261949
1459       14.3       14.477373
857        11.3       13.361495
137        20.2       21.769101
2026       20.0       20.457726
59         21.3       20.079504
```

Εικόνα 3:Παράμετροι γραμμικής παλινδρόμησης και τιμές



Εικόνα 4:real temp vs predicted temp

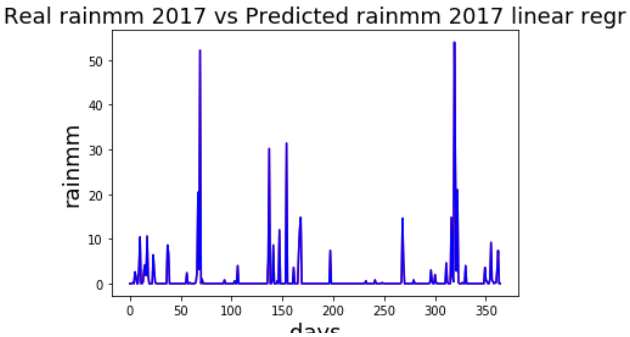


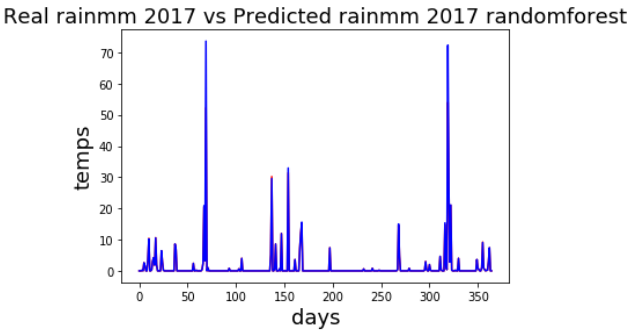
Εικόνα 5:RMSE

Παρατηρούμε ότι επιτυγχάνουμε χαμηλό μέσο τετραγωνικό σφάλμα και με τις 2 μεθόδους αλλά με επικρατέστερη μέθοδο εκείνη του random forest. Επίσης επιβεβαιώνεται από το ότι οι προβλεπόμενες τιμές τείνουν στις πραγματικές και υπάρχει μία πολύ καλή ταύτιση και στη χρονοσειρά

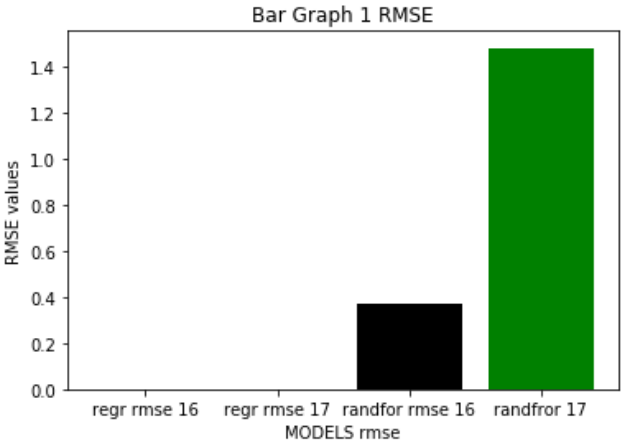
Στο δεύτερο ερώτημα αντίστοιχα εξετάσαμε τη πρόβλεψη του Ύψους της Βροχόπτωσης. Έχοντας παρατηρήσει τους πολύ χαμηλούς συντελεστές συσχέτισης αλλά και τα διαγράμματα διασποράς αναμέναμε λιγότερη ακρίβεια στα αποτελέσματα μας με υψηλό μέσο τετραγωνικό σφάλμα στη πολλαπλή γραμμική μέθοδο.

```
Intercept:
 4.6629367034256575e-14
Coefficients:
 [-1.24319300e-15 -1.44328993e-15 -2.75474088e-15  2.53269627e-15
  1.00000000e+00 -9.03682511e-17  1.86865633e-16]
      Real 2016  Predicted 2016
209          7.6     7.600000e+00
2354         0.2     2.000000e-01
235          0.0     2.561211e-15
2137         0.2     2.000000e-01
1504         0.0     2.969902e-15
479          0.0     2.532986e-15
1827         0.0     1.046063e-15
1774         0.0     3.601905e-15
1833         0.0     1.878732e-15
877          0.0    -6.085522e-16
2205         0.0     8.713911e-16
773          0.0     2.775621e-15
1754         0.0    -1.051078e-15
```





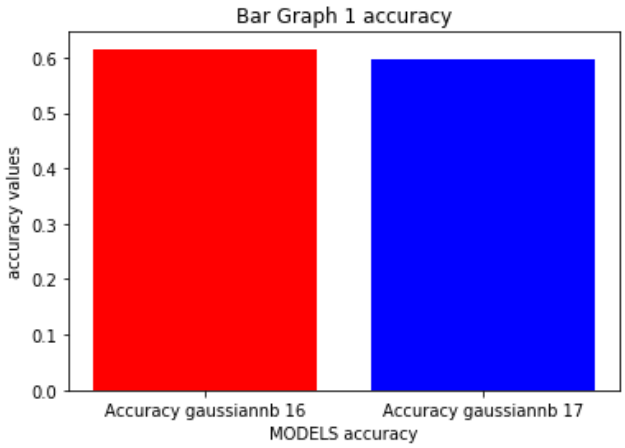
Εικόνα 5:real temp vs predicted temp



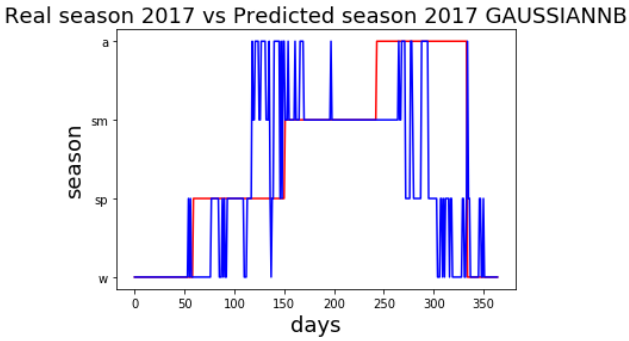
Εικόνα 6:RMSE

Όπως επιβεβαιώθηκε και από τα δεδομένα η μέθοδος της πολλαπλής γραμμικής παλινδρόμησης δεν είχε καλά αποτελέσματα. Αντίθετα η randomforest κατάφερε να επιτύχει ένα χαμηλό μέσο τετραγωνικό σφάλμα. Επιπλέον παρατηρήθηκε ότι στο 2017 dataset παρουσιάστηκε πολύ υψηλό (σχεδόν τριπλάσιο) RMSE που πιθανόν να οφείλεται σε κάποιο overfitting στο dataset 2016.

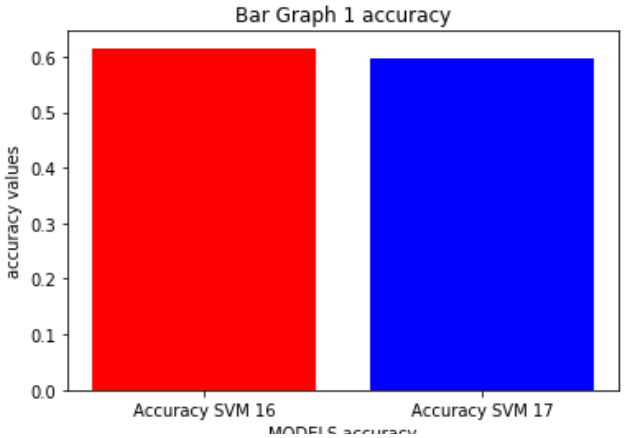
Στο τρίτο ερώτημα ήταν ένα πρόβλημα classification με σκοπό να βρεθεί η εποχή με βάση τις υπόλοιπες μεταβλητές. Αρχικά και στα 2 dataset δημιουργήθηκε μια νέα μεταβλητή season όπου αναφερόταν η εποχή ανα γραμμή. Στη συνέχεια αφαιρέθηκε η μεταβλητή Month και το dataset 2016 χωρίστηκε πάλι σε 70% train set και 30% test set. Υλοποιήθηκαν οι αλγόριθμοι SVM και NAÏVE BAYES.



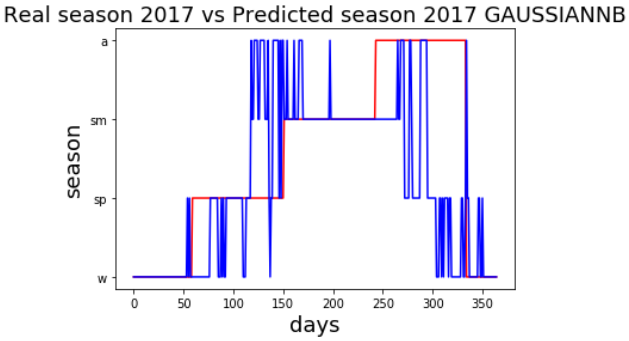
Εικόνα 7:Accuracy



Εικόνα 8:real season vs predicted season



Εικόνα 9: Accuracy



Εικόνα 8:real season vs predicted season

Παρατηρούμε ότι και οι 2 μέθοδοι επιτυγχάνουν χαμηλό accuracy, πιθανόν λόγο λάθος διαχωρισμού του dataset 2016 (Μεγάλη συχνότητα μια εποχής στο train set).

Καταλήγοντας κατανοούμε ότι η χρήση machine learning αλγορίθμων θα μπορούσε να αντικαταστήσει την κλασσική μέθοδο πρόβλεψης. Τα παραπάνω αποτελέσματα παραμετροποιούνται ώστε να επιτευχτεί μεγαλύτερη ακρίβεια. Μια επιπλέον ανάλυση των δεδομένων θα ήταν η πρόβλεψη κάποιας μεταβλητής σε μελλοντικό χρονικό διάστημα με τη χρήση χρονοσειρών. (πχ μέθοδος ARIMA)

Παραπομπές

[1] http://www.ep.liu.se/ecp/153/024/ecp18153024.pdf
[2] http://www.ep.liu.se/ecp/153/024/ecp18153024.pdf.
[3] https://medium.com/@shivamtrivedi25/what-is-the-weather-prediction-algorithm-how-it-works-what-is-the-future-a159040dd269
[4] https://www.linkedin.com/pulse/weather-forecasting-using-machine-learning-models-model-kinjal-ami/.