Εργασία Εξαμήνου Εξόρυξης Δεδομένων

Βλάχος Ευγένιος

Πρόβλεψη εκπομπών SO_2 από μονάδες παραγωγής ηλεκτρικής ενέργειας

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας
Μάθημα Εξόρυξης Δεδομένων - 2020/2021

1 Εισαγωγή και παρουσίαση θέματος

Σκοπός της εργασίας αυτής είναι να δοκιμάσουμε διάφορα μοντέλα ώστε να μπορέσουμε να προβλέψουμε σε ικανοποιητικό βαθμό την παραγόμενη ποσότητα διοξειδίου του θείου για έναν δεδομένο σταθμό παραγωγής ηλεκτρικής ενέργειας.Με αυτόν τον τρόπο θα μπορέσουμε μελλοντικά να κάνουμε τις απαραίτητες ενέργειες για την σταδιακή μείωση του και επομένως την βελτίωση της ανθρώπινης υγείας , αλλά και του κλίματος.

Τα επίπεδα του διοξειδίου του θείου, αποτελούν έναν καλό δείκτη για την ποιότητα του ατμοσφαιρικού αέρα και είναι ένας από τους πιο χαρακτηριστικούς ρύπους της ατμόσφαιρας κυρίως για τις αστικές περιοχές.

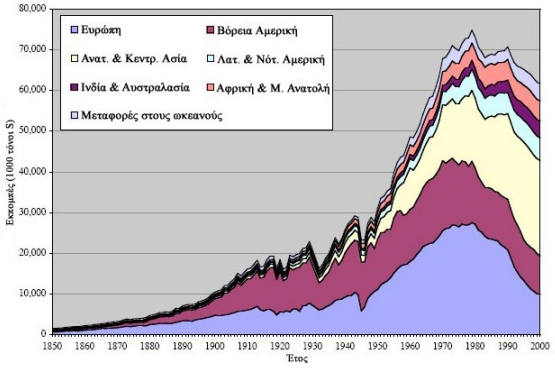
Δημιουργείται είτε φυσικά , είτε μέσω διεργασιών που προκαλούν οι άνθρωποι. Οι σημαντικότερες ανθρωπογενείς πηγές είναι η καύση ορυκτών καυσίμων που συναντούμε στους σταθμούς παραγωγής ηλεκτρικής ενέργειας, η δυίλυση του πετρελαίου, βιομηχανικές δραστηριότητες κ.ά.

Επιπλέον, εξαιτίας της εκτεταμένης συγκέντρωσης του διοξειδίου του θείου,επηρεάζεται σε μεγάλο βαθμό το κλίμα του πλανήτη , ενώ παράλληλα προσβάλλεται η ανθρώπινη υγεία λόγω της χρόνιας έκθεσης στα θειϊκά αιωρούμενα σωματίδια .

Αν και τα τελευταία χρόνια έχουν γίνει προσπάθειες για τον περιορισμό τόσο των εκπομπών διοξειδίου του θείου όσο και άλλω αέριων ρύπων κατά την διάρκεια της παραγωγής της ηλεκτρικής ενέργειας, εξαιτίας της μη τήρησης όλων τον προδιαγραφών ασφαλείας πάνω από 2 εκατομύρια άνθρωποι χάνουν τη ζωή τους λόγω της ατμοσφαιρικής ρύπανσης.

Βασικοί πλέον υπεύθυνοι για τις εκπομπές διοξειδίου του θείου στην ατμόσφαιρα είναι η Κεντρική και Ανατολική Ασία και, κατ' επέκταση η Ευρώπη και η Βόρεια Αμερική.

Παρακάτω βλέπουμε ένα διάγραμμα των εκπομπών του ανά έτος.



Σχήμα 1: Εκπομπές SO_2 ανά έτος

2 Ανάλυση δεδομένων και μεθοδολογία επεξεργασία τους

Αρχικά, τα δεδομενα που αποφασίσαμε να χρησιμοποιήσουμε είναι τα αρχεία $xlsx$ μορφής, που παρέχονται από την US Energy Information και δίνουν διάφορα στοιχεία για τις σταθμούς παραγωγής ηλεκτρικής ενέργειας. Τα δεδομένα είναι κατηγοριοποιημένα σε στήλες με βάση διάφορα χαρακτηριστικά τους. Κάποιες αξιοσημείωτες στήλες ειναι οι:

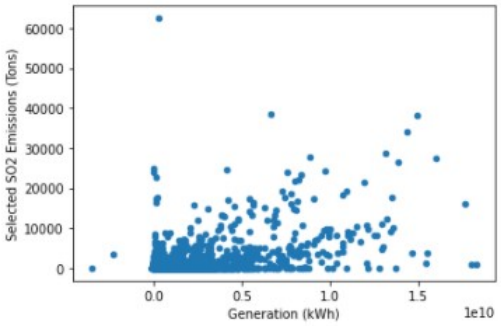
- "Aggregated Fuel Group", η οποία διαχωρίζει τους σταθμούς με βάση το είδος του καυσίμου που καταναλώνουν.
- "Generation (kWh)", δείχνει πόση ενέργεια καταναλώνει κάθε σταθμός ετησίως, σε κιλοβατώρες.
- "Total Fuel Consumption (MMBtu)", παρουσιάζει το συνολικό ποσό καυσίμου που χρησιμοποιήθηκε για παραγωγή ενέργειας, μετρημένο σε MMBtu(Metric Million British Thermal Unit).
- "Selected SO2 Emissions (Tons)",δείχνει τις μετρήσεις για τις ετήσιες εκπομπές του SO_2 σε τόνους.

Αποφασίσαμε για τον σκοπό της εργασίας να επικεντρωθούμε στην στήλη των εκπομπών του SO_2, και για αυτο επιλέχθηκαν αλγόριθμοι που θα μας βοηθήσουν να προβλέψουμε τις τιμές του.

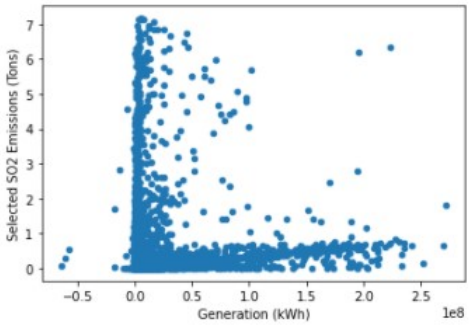
Η επεξεργασία των δεδομένων μας ξεκίνησε με την μετατροπή των categorical στηλων σε numerical με την χρήση των κατάλληλων εργαλείων στο περιβάλλον της python. Διαγράφθηκε η τελευταία στήλη των δεδομένων μας καθώς αντιληφθήκαμε μέσω του correlation matrix ότι οι τιμές της συνέπιπταν με την στήλη την οποία θέλαμε να προβλέψουμε, οπότε οι υπολογισμοί μας θα ήταν λανθασμένοι. Στη συνέχεια, αφαιρέσαμε τις στήλες των οποίων οι τιμές συνέβαλαν πολύ λίγο στη διαμόρφωση του τελικού αποτελέσματος λόγω χαμηλού correlation, άρα δημιουργούσαν outliers.

Αυτό μας οδήγησε στην αναζήτηση κάποιας μεθόδου η οποία θα μας βοηθούσε να αφαιρέσουμε αυτές τις ακραίες τιμές. Για αυτό το λόγο αποφασίσαμε να χρησιμοποιήσαμε τη μέθοδο IQR για τον εντοπισμό ακραίων τιμών για να δημιουργήσουμε ένα όριο έξω από τα $Q1$ και $Q3$, όπου $Q1$ είναι το κάτω όριο τιμών που αντιπροσωπεύει το 25% των τιμών και $Q3$ το άνω, δηλαδή το 75% των τιμών. Για την κατασκευή αυτού του ορίου παίρνουμε 1,5 φορές το $IQR(Q3 - Q1)$ και στη συνέχεια αφαιρούμε αυτήν την τιμή από το $Q1$ και την προσθέτουμε αντίστοιχα στο $Q3$. Αυτό μας δίνει τις ελάχιστες και μέγιστες θέσεις ορίου στις οποίες συγκρίνουμε κάθε παρατήρηση. Τυχόν παρατηρήσεις που είναι μικρότερες από $Q1 - 1,5 * IQR$ ή μεγαλύτερες από $Q3 + 1,5 * IQR$ θεωρούνται ακραίες τιμές. Με αυτόν τον τρόπο, βρήκαμε τις γραμμές που μας δημιουργούσαν πρόβλημα και σχηματίσαμε μία λίστα που τις περιλάμβανε, διαγράφοντάς τες από το ολικό dataset. Παρακάτω φαίνεται η διαφοροποίηση ενός εκ των στηλών αφού εφαρμόσαμε την

τεχνική IQR.



Σχήμα 2: Διάγραμμα παραγωγής ενέργειας πριν την επεξεργασία των δεδομένων

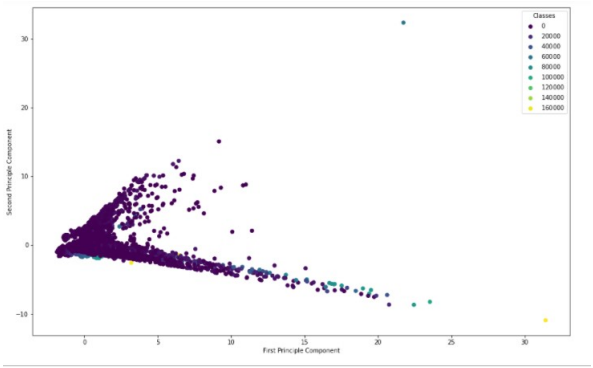


Σχήμα 3: Διάγραμμα παραγωγής ενέργειας μετά την επεξεργασία των δεδομένων

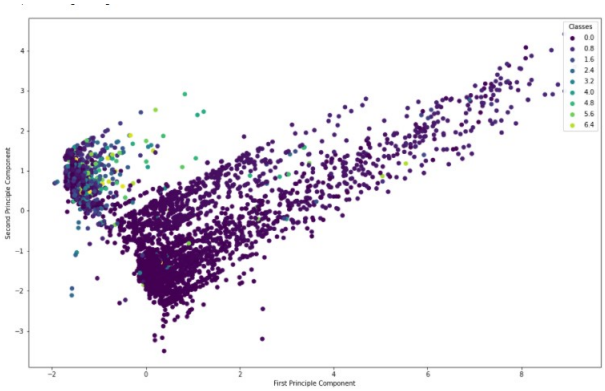
2.1 Αναπαράσταση δεδομένων με PCA

Σε αυτό το σημείο κρίναμε απαραίτητο να εφαρμόσουμε την μέθοδο PCA, για να αναπαραστήσουμε τα δεδομένα με βάση τα principle components.
ΕΦόσον οι στήλες μας ήταν πολλές δεν θα μπορούσαμε να κάνουμε μια αναπαράσταση των δεδομένων σε ενιαίο δισδυάστατο διάγραμμα. Το PCA βοήθησε στην απλούστευση των στηλών, επειδή μείωσε τις διαστάσεις σε 2 και ανέδειξε ισχυρά μοτίβα στο σύνολο των δεδομένων μας. Έτσι μετά από την εφαρμογή του πήραμε τα ακόλουθα διαγράμματα που μας δείχουν την επιτυχία των τεχνικών που εφαρμόσαμε για την επεξεργασία των δεδομένων και την αφαίρεση των ακραίων τιμών.

Τελικά, τα δεδομένα είναι έτοιμα για να ξεκινήσουμε την εφαρμογή μοντέλων και αλγορίθμων για την πρόβλεψη που θέλουμε να κάνουμε.



Σχήμα 4: Διάγραμμα δεδομένων πριν την εφαρμογή PCA



Σχήμα 5: Διάγραμμα δεδομένων μετά την εφαρμογή PCA

3 Συνοπτική παρουσίαση αλγορίθμων

Χρησιμοποιήσαμε μοντέλα παλινδρόμισης ώστε να μοντελοποιήσουμε τη σχέση μεταξύ των ανεξάρτητων χαρακτηριστικών των δεδομένων μας με την στήλη που θέλουμε να προβλέψουμε , δηλαδή τις εκκρίσεις διοξειδίου του θείου σε τόνους κατά την παραγωγή ηλεκτρικής ενέργειας από πολυάριθμες μονάδες παραγωγής σε σταθμούς στην Αμερική.

Επιλέξαμε το τετραγωνικό σφάλμα R^2 για να υπολογίσουμε την ακρίβεια της πρόβλεψης που προέκυπτε από κάθε μέθοδο. Το R^2 είναι ένα στατιστικό μέτρο που αντιπροσωπεύει το ποσοστό της διακύμανσης για μια εξαρτημένη μεταβλητή που εξηγείται από μια ανεξάρτητη μεταβλητή ή μεταβλητές σε κατάσταση παλινδρόμησης.

Η ανάλυση παλινδρόμισης εκτιμά την επίδραση που έχει η αλλαγή μιας ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή ενω διατηρεί σταθερές όλες τις άλλες ανεξάρτητες μεταβλητές.

Ετσι επιγραμματικά μπορούμε να αναφέρουμε τις τεχνικές που επιλέξαμε οι οποίες παρείχαν άλλοτε ικανοποιητικά αποτελέσματα και άλλες φορές χαμηλής ακρίβειας.

- Random forest regression
- KNN regression
- Linear regression
- Decision trees regression
- SVM regression
- Gradient boosting regression
- Neural network regression

4 Αναλυτική επεξήγηση λειτουργίας του κάθε αλγορίθμου

Στη συνέχεια θα παραθέσουμε λεπτομέρειες για τον τρόπο λειτουργίας του κάθε αλγορίθμου και εάν η χρήση του ήταν αποτελεσματική για την πρόβλεψη που θέλουμε να επιτύχουμε.

Αρχικά, η παλινδρόμιση random forest αποτελεί μία supervised machine learning μέθοδο, η οποία σχηματίζει διαδοχικά decision trees κατά τη διάρκεια της εκπαίδευσης και το αποτέλεσμα στο οποίο καταλήγει είναι η μέση τιμή όλων των προβλέψεων που προκύπτουν από κάθε δέντρο.

Στη δική μας υλοποίηση παρατηρήσαμε ότι καλύτερη πρόβλεψη επιτυγχάνεται με τη χρήση 24 regressors, και παράλληλα αποτελεί τη μέθοδο κατά την οποία λαμβάνουμε την καλύτερη πρόβλεψη σε σχέση με τις υπόλοιπες, παρουσιάζοντας ακρίβεια περίπου 66%.

Ένας άλλος αλγόριθμος που χρησιμοποιήσαμε είναι ο KNN για regression. Ο αλγόριθμος KNN χρησιμοποιεί «ομοιότητα χαρακτηριστικών» για να προβλέψει τις τιμές οποιωνδήποτε νέων σημείων δεδομένων. Αυτό σημαίνει ότι στο νέο σημείο εκχωρείται μια τιμή με βάση το πόσο πολύ μοιάζει με τα σημεία στο training set.

Η παλινδρόμηση KNN είναι μια μη παραμετρική μέθοδος που, με διαισθητικό τρόπο, προσεγγίζει τη σχέση μεταξύ ανεξάρτητων μεταβλητών και του συνεχούς αποτελέσματος με μέσο όρο των παρατηρήσεων στην ίδια γειτονιά.

Στη δική μας περίπτωση με χρήση επαναληπτικής μεθόδου και αποτύπωσης του mean squared error καταλήξαμε στην τελική τιμή του συντελεστή K η οποία είναι ίση με 8.

Αν και πήραμε την βέλτιστη τιμη του συντελεστή K η μέθοδος καταλήγει σε μια μέτρια πρόβλεψη εξαιτίας της ανομοιότητας των δεδομένων μας και πετυχαίνει ακρίβεια κοντά στο 55%, εάν συνυπολογίσουμε και την τυχαία επιλογή του training set μας το οποίο αντιστοιχούσε στο 80% του συνολικού dataset, ενω το test set αντιστοιχούσε στο υπόλοιπο 20%.

Επιπλέον, χρησιμοποιήσαμε linear regression κυρίως δοκιμαστικά μιας και γνωρίζαμε ότι λόγω της φύσης των δεδομένων μας θα ήταν αδύνατον να λάβουμε μια ικανοποιητική πρόβλεψη.

Η γραμμική παλινδρόμηση επιχειρεί να μοντελοποιήσει τη σχέση μεταξύ δύο μεταβλητών προσαρμόζοντας μια γραμμική εξίσωση σε παρατηρούμενα δεδομένα. Η μία μεταβλητή θεωρείται ως επεξηγηματική μεταβλητή(explanatory variable) και η άλλη θεωρείται εξαρτημένη μεταβλητή(dependent variable).

Η ακρίβεια που παρατηρήσαμε ήταν αρκετά μικρή, με τιμή μεταξύ 15-20%. Ήταν από τις πρώτες μέθοδους που χρησιμοποιήσαμε και διαπιστώσαμε από την αρχή την δυσκολία που παρουσίαζαν τα δεδομένα μας .

Μέσω της παλινδρόμισης με δέντρα αποφάσεων αναλύεται το σύνολο δεδομένων σε μικρότερα υποσύνολα. Ένα φύλλο απόφασης χωρίζεται σε δύο ή περισσότερους κλάδους που αντιπροσωπεύουν την αξία του υπό εξέταση χαρακτηριστικού. Ο κορυφαίος κόμβος στο δέντρο αποφάσεων είναι ο καλύτερος προγνωστικός παράγοντας που ονομάζεται ρίζα. Χρησιμοποιεί προσέγγιση από πάνω προς τα κάτω και οι διαχωρισμοί γίνονται με βάση την τυπική απόκλιση. Η τελική τιμή θα είναι ο μέσος όρος των κόμβων των φύλλων.

Με το δικό μας dataset η ακρίβεια του αλγορίθμου κυμαίνεται μεταξύ 40% και 50% και έτσι σε σχέση με τους υπόλοιπους αλγορίθμους δίνει μια ενθαρρυντική πρόβλεψη.

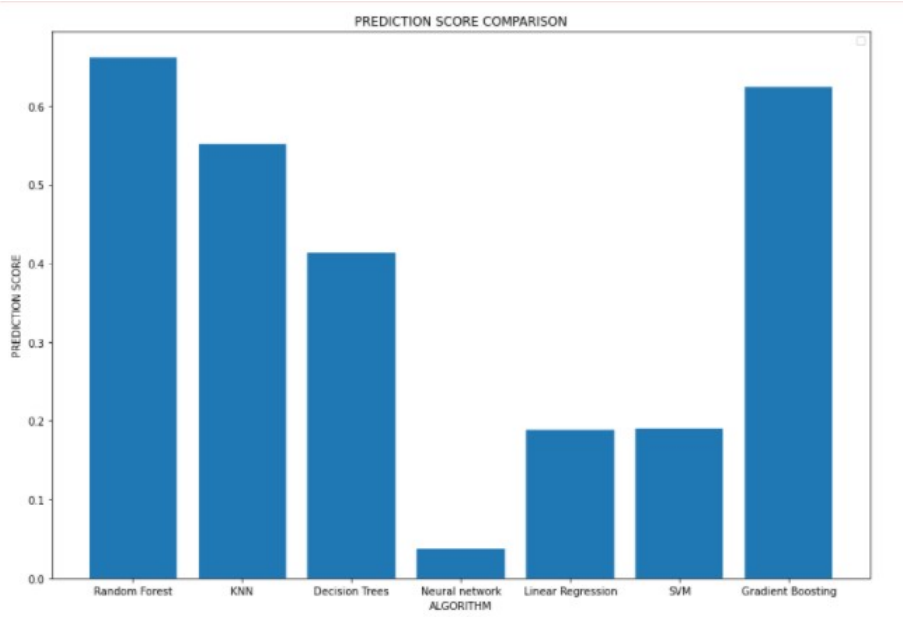
Το Support Vector Regression είναι ένας supervised machine learning αλγόριθμος που χρησιμοποιείται για την πρόβλεψη διακριτών τιμών. Το SVR χρησιμοποιεί την ίδια αρχή με τα Support Vector Machines($SVMs$). Η βασική ιδέα πίσω από το SVR είναι η εύρεση της καλύτερης γραμμής επάνω στα δεδομένα. Σε αντίθεση με άλλα μοντέλα παλινδρόμησης που προσπαθούν να ελαχιστοποιήσουν το σφάλμα μεταξύ της πραγματικής και της προβλεπόμενης τιμής. Η πολυπλοκότητα του χρόνου προσαρμογής του SVR είναι περισσότερο από τετραγωνική με τον αριθμό των δειγμάτων που καθιστά δύσκολη την κλιμάκωση σε σύνολα δεδομένων, με περισσότερα από μερικά 10.000 δείγματα.

Για αυτό το λόγο , αλλά και εξαιτίας της ανομοιομορφίας των δεδομένων δεν επιτυγχάνεται καλή πρόβλεψη με την μέγιστη που έχουμε παρατηρήσει να είναι περίπου 18%.

Η παλινδρόμιση Gradient Boosting βασίζεται στη διαίσθηση ότι το καλύτερο δυνατό επόμενο μοντέλο, όταν συνδυάζεται με προηγούμενα μοντέλα, ελαχιστοποιεί το συνολικό σφάλμα πρόβλεψης. Εάν μια μικρή αλλαγή στην πρόβλεψη για μια υπόθεση προκαλεί μεγάλη πτώση του σφάλματος, τότε το επόμενο αποτέλεσμα στόχου της υπόθεσης είναι υψηλό. Οι προβλέψεις από το νέο μοντέλο που πλησιάζουν τους στόχους του θα μειώσουν το σφάλμα. Εάν μια μικρή αλλαγή στην πρόβλεψη για μια υπόθεση δεν προκαλεί καμία αλλαγή στο σφάλμα, τότε το επόμενο αποτέλεσμα στόχου της υπόθεσης είναι μηδέν. Η αλλαγή αυτής της πρόβλεψης δεν μειώνει το σφάλμα.

Πρέπει να αναφέρουμε ότι με τη συγκεκριμένη μέθοδο παρατηρούμε τιμές άνω του 60% καθιστώντας την μία από τις καλύτερες μεθόδους για να εφαρμόσουμε στα δεδομένα μας.

5 Συγκεντρωτικό διάγραμμα προβλέψεων



Σχήμα 6: Συγκριτική αποτύπωση της αποτελεσματικότητας των αλγορίθμων

6 Σχολιασμός και ερμηνεία αποτελεσμάτων

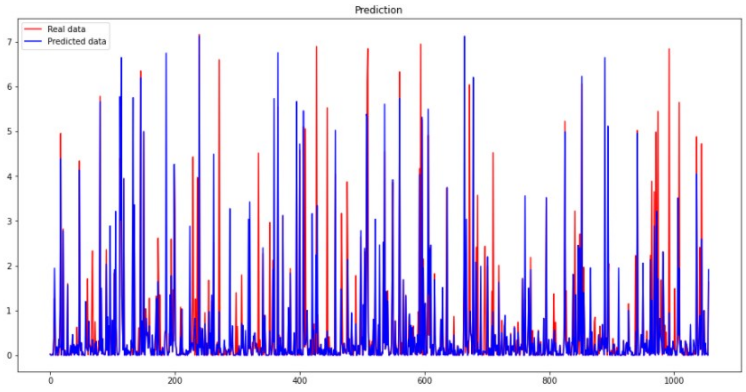
Εξαιτίας της φύσης των δεδομένων μας, δηλαδή την έλλειψη γραμμικότητας παρατηρήσαμε ότι η επίδοση των περισσότερων αλγορίθμων δεν είναι ικανή ώστε να μας δώσει μια επιθυμητή πρόβλεψη. Έτσι, είδαμε τιμές πρόβλεψης μεταξύ 15% και 20% εκτός τριών μεθόδων παλινδρόμισης. Αυτό συνέβη κατα κύριο λόγο επειδή η μορφή των δεδομένων μας δεν έδωσε στους υπόλοιπους αλγορίθμους την ευκαιρία να λειτουργήσουν σωστά και να βγάλουν ένα καλύτερο αποτέλεσμα στην πρόβλεψή τους. Οι μέθοδοι που αντεπεξήλθαν σε ικανοποιητικό βαθμό είναι οι εξής:

- Random Forest Regression
- Decision Trees Regression
- Gradient Boosting Regression
- KNN Regression

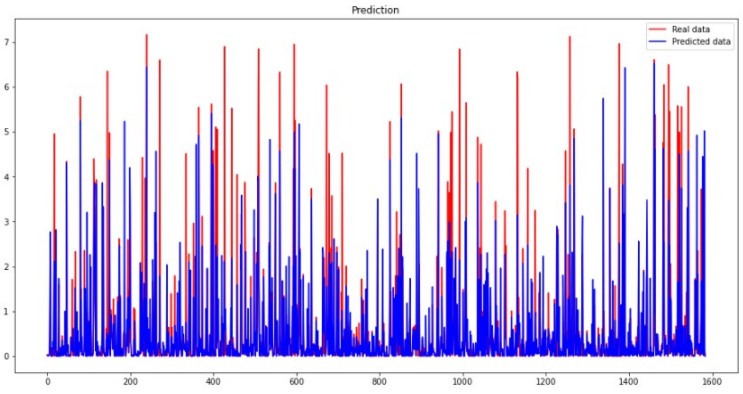
Με τις μεθόδους decision trees & KNN παρατηρήσαμε τιμές πρόβλεψης μεταξύ 40% και 55% ανάλογα με την επιλογή του training set.Από την άλλη πλευρά,

με τις υπόλοιπες 2 μεθόδους είδαμε αρκετά καλύτερες τιμές πρόβλεψης που κυμαίνονταν μεταξύ 60% και 65%. Συγκριτικά όμως πρέπει να αναφέρουμε ότι με την gradient boosting παλινδρόμιση θα είχαμε πιο βέλτιστη πρόβλεψη εάν δεν υπήρχε τόσος θόρυβος στα δεδομένα μας. Επίσης, για τη παλινδρόμιση random forest παρατηρήσαμε ότι παίρνει περισσότερο χρόνο, σε σχέση με τη παλινδρόμιση gradient boosting η εκπαίδευση του training set. Άρα συμπεραίνουμε ότι η παλινδρόμιση gradient boosting καταλήγει σε ποιοτικότερα αποτελέσματα και χωρίς την μεγάλη κατανάλωση πόρων, ενώ εάν τα δεδομένα μας είχαν λιγότερο θόρυβο, η πρόβλεψη θα ήταν αρκετά πιο ικανοποιητική.

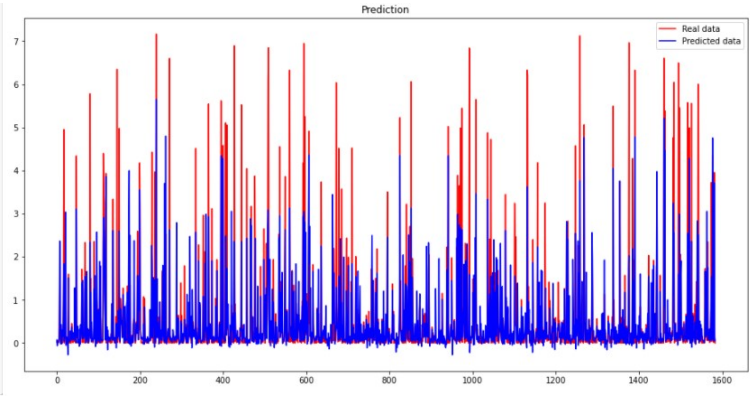
Παρακάτω παραθέτουμε τα διαγράμματα πρόβλεψης των τριών πιο αποδοτικών αλγορίθμων για να δείξουμε την σχέση μεταξύ πραγματικής και προβλεπόμενης τιμής.



Σχήμα 7: Πρόβλεψη *Decision Trees Regression*



Σχήμα 8: Πρόβλεψη *Random Forest Regression*



Σχήμα 9: Πρόβλεψη *Gradient Boosting Regression*

7 Μελλοντική προεκτάσεις και ευαισθητοποίηση

Ολοκληρώνουμε υπενθυμίζοντας πως δυστυχώς οι ανάγκες των ανθρώπων για ενέργεια είναι ατελείωτες και για αυτό τον λόγο κρίνεται επιτακτική η ανάγκη για εύρεση μεθόδων που περιορίζουν τους αέριους ρύπους και εξασφαλίζουν καλύτερη ποιότητα ζωής παράλληλα με την κάλυψη βασικών αναγκών για ενέργεια. Επισης, οι αυξημένες ποσότητες του διοξειδίου του θείου στην ατμόσφαιρα , έχουν πολλές επιζήμιες συνέπειες, όπως το φαινόμενο της όξινης βροχής και βλάπτει ιδιαίτερα τόσο τα φυτικά όσο και τα ζωικά οικοσυστήματα καθως και τους υδρόβιους οργανισμούς. Γίνεται λοιπόν αντιληπτό πως η επιτυχημένες προβλέψεις των βλαβερών εκπομπών απο τις μεγάλες εταιρείες παραγωγής ενέργειας, μπορούν να σύμβαλουν στην λήψη κρίσιμων αποφάσεων, οι οποίες θα είναι σωτήριες για το περιβάλλον και τον άνθρωπο.

8 Σύνδεσμοι

Σύνδεσμος για τα δεδομένα και επεξηγηματικά βίντεο:
https://www.eia.gov/electricity/data/emissions/?fbclid=IwAR1gIet_d_9uCnpzL_8Zv04ZxiXvoPd0N5YUPm6PehU9n1rlqNSn-YXxtw4
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
https://www.statisticshowto.com/probability-and-statistics/interquartile-range/
https://www.youtube.com/watch?v=Vc4cXIAa69Y&lc=z22oxtwqkonselayl04t1aokg3cfzaqlrzx2jsef0j1ub
https://statisticsbyjim.com/regression/interpret-r-squared-regression/