Εξαμηνιαία εργασία στο μάθημα «Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης»

Ακαδημαϊκό έτος 2021-2022

Των φοιτητών: Κοκκάλα Βασιλική(1067519), Φουσκαρής Γιώργος (1067404)

**Κεφάλαιο 0: Εισαγωγή**

Στα πλαίσια του μαθήματος Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης κληθήκαμε να υλοποιήσουμε μια πληθώρα συναρτήσεων για την απόκτηση γνώσης από πραγματικά σύνολα δεδομένων. Συγκεκριμένα μας δόθηκαν οι αναλυτικές καταστάσεις παραγωγής και κατανάλωσης ενέργειας για την πολιτεία της Καλιφόρνια μεταξύ 1/1/2019 και 31/12/2021 καθώς και πραγματικά δεδομένα από την βάση δεδομένων της amazon σχετικά με προϊόντα και τις αντίστοιχες λεκτικές περιγραφές τους. Στα επόμενα μέρη θα αναλύσουμε την μέθοδο με την οποία εργαστήκαμε καθώς και τα αποτελέσματα που παρήχθησαν. Ως γλώσσα υλοποίησης επιλέχθηκε η Python .

Κεφάλαιο 1: Ενεργειακά Δεδομένα Καλιφόρνιας

## Καθαρισμός των δεδομένων

Για την ενοποίηση των δοθέντων αρχείων csv και την ουσιαστική αξιοποίηση τους ήταν αναγκαία η συνένωση των αντίστοιχων αρχείων μεταξύ τους για την παραγωγή ενός συνολικού αρχείου. Ωστόσο κατά την εκτέλεση του κώδικα αντιμετωπίσαμε κάποια σφάλματα. Αυτό συνέβαινε διότι πέραν των πραγματικών ημερών εντός των αρχείων csv υπήρχαν και αρχεία για πλασματικές ημέρες χωρίς φυσικά δεδομένα. Ενδεικτικά ένα τέτοιο αρχείο ήταν το 20190231.csv . Αρχείο που αφορά την 31η Φεβρουαρίου 2019 (!). Για την διαγραφή αυτών των αρχείων τα αρχεία ταξινομήθηκαν βάσει μεγέθους και στην συνέχεια διεγράφησαν όσα εμφάνιζαν μηδενικό μέγεθος .

Στην συνέχεια παρατηρήσαμε πως αρκετές τιμές παραγωγής ενέργειας ήταν αρνητικές, πράγμα παράλογο για ορισμένες μορφές ενέργειας. Για το υπόλοιπο της εργασίας θα διατηρηθεί η σύμβαση με την οποία εργάστηκε η ομάδα μας. Οι μόνες πηγές ενέργειας οι οποίες δύνανται να έχουν αρνητικές τιμές είναι οι πηγές Large Hydro & Batteries. Αυτό συμφωνήθηκε λόγω της πιθανότητας υποδομών pumped storage στα μεγάλα υδροηλεκτρικά φράγματα ενώ για τις μπαταρίες οι δυνατότητα φόρτισης τους από το δίκτυο λειτουργεί αντιστοίχως. Και οι δύο αυτές λειτουργίες χρησιμοποιούνται ευρέως στον πραγματικό κόσμο για την ισοστάθμιση παραγωγής και ζήτησης σε στιγμές απότομης μείωσης της ζήτησης. Για τις υπόλοιπες πηγές οι αρνητικές τιμές αρχικά μεταβλήθηκαν σε 0 και στην συνέχεια με χρήση συναρτήσεων παλινδρόμησης δόθηκαν οι πιθανότερες τιμές βάσει των προηγούμενων και επόμενων τιμών.

Επιπροσθέτως αντιμετωπίσαμε το πρόβλημα της λανθασμένης μορφής της ώρας και της καταχώρισης δεδομένων. Επί παραδείγματι, σε ορισμένα αρχεία η ώρα γραφόταν 06:00 ενώ σε λίγα συγκεκριμένα γραφόταν ως 6:00. Αυτό είχε ως αποτέλεσμα την μη αποδοτική παραγωγή δεδομένων και γραφικών συναρτήσεων. Για την αντιμετώπιση του φαινομένου χρησιμοποιήθηκε συνάρτηση η οποία μετέτρεπε την κάθε τιμή στο αναμενόμενο format με δύο ψηφία για την ώρα. Ωστόσο λανθασμένη μορφή δεδομένων είχαμε και στις στήλες Large Hydro, Large hydro και Natural Gas, Natural gas για τις οποίες σε συγκεκριμένα στοιχεία τα δεδομένα καταχωρούνταν στις στήλες με τα μικρά γράμματα.  
Για την ενοποίηση των στηλών οι τιμές προστέθηκαν μεταξύ τους και καταχωρήθηκαν στις στήλες με τα κεφαλαία γράμματα στην αρχή της δεύτερης λέξης. Είχε προηγηθεί μελέτη των γραφικών παραστάσεων ώστε να βεβαιωθούμε ότι δεν υπήρχε αλληλοεπικάλυψη ή μεγάλη απόκλιση των τιμών στο σημείο συνένωσης τους.

Κατόπιν όμως αυτής της επεξεργασίας, παρέμεναν ορισμένες μη λογικές τιμές.  
Είτε μεμονωμένες είτε συγκεκριμένες μέρες εμφάνιζαν παράλογες τιμές όπως για παράδειγμα, έντονη παραγωγή ηλιακής ενέργειας στις 1 π.μ. ή στιγμιαίος δεκαπλασιασμός ακολουθούμενος από στιγμιαίο αποδεκατισμό της παραγωγής υδροηλεκτρικής ενέργειας. Για την αντιμετώπιση αυτών χρησιμοποιήθηκαν συναρτήσεις παλινδρόμησης για την τοποθέτηση αληθοφανών τιμών.

## Γραφική Αναπαράσταση των δεδομένων και βασικά στατιστικά στοιχεία

Για την οπτικοποίηση των διαθέσιμων δεδομένων χρησιμοποιήθηκαν δύο τύποι γραφημάτων, ένα διάγραμμα Scatter που συγκέντρωνε τις τιμές ως προς την ώρα της ημέρας, καθώς και ένα διάγραμμα της εκάστοτε μεταβλητής ως προς την χρονική στιγμή.

Παρατίθενται οι δύο γραφικές αναπαραστάσεις με σχόλια καθώς και τα ανάλογα στατιστικά δεδομένα

## Ομαδοποίηση δεδομένων (Clustering)

## Εκπαίδευση νευρωνικού δικτύου LSTM και πρόβλεψη μελλοντικών τιμών

Όσον αφορά την πρόβλεψη μελλοντικών τιμών ζήτησης για τις μη ανανεώσιμες πηγές ενέργειας, χρησιμοποιήσαμε τα προηγούμενα στοιχεία ως σημείο αναφοράς. Πιο συγκεκριμένα μη ανανεώσιμες πηγές ενέργειας στα δεδομένα μας είναι ο άνθρακας (Coal) καθώς και το φυσικό αέριο το οποίο όπως προαναφέρθηκε παραπάνω έχει τις τιμές του στην στήλη Natural Gas merged. Η πρόσθεση των δύο ανωτέρω στηλών μας δίνει τα ιστορικά δεδομένα για τις ανάγκες ενέργειας της Καλιφόρνια από μη ανανεώσιμες πηγές. Χρησιμοποιήθηκε η βιβλιοθήκη tensorflow με το keras ως διεπαφή για την δημιουργία και εκπαίδευση νευρωνικού δικτύου. Το νευρωνικό δίκτυο μελετά το 70% του πλήθους των τιμών σε 6-άδες όπου χρησιμοποιώντας τις 5 πρώτες επιχειρεί να εκτιμήσει την έκτη τιμή. Η διαδικασία αυτή επαναλαμβάνεται για το σύνολο των δεδομένων σε 20 epochs. Στην συνέχεια για τον έλεγχο ακρίβειας εισάγεται τυχαία 5-άδα δεδομένων από το υπολειπόμενο 30% των δεδομένων και το δίκτυο εκτυπώνει την πρόβλεψη του καθώς και τα εισαγόμενα δεδομένα. Από τις παρατηρήσεις μας θεωρούμε ότι η ακρίβεια του δικτύου αν και όχι απόλυτη είναι ικανοποιητική συναρτήσει του χρόνου για την εκτέλεση του προγράμματος. Γνωρίζουμε πως αν αυξάναμε τον αριθμό τον epochs, το πλήθος των δεδομένων εκπαίδευσης αλλά και ενδεχομένως το πλήθος Ν στις Ν-αδες εισαγωγής δεδομένων θα είχαμε ενδεχομένως καλύτερες επιδόσεις. Για τον εκπαιδευτικό σκοπό του μαθήματος και με τα μέσα που διαθέτουμε θεωρούμε το αποτέλεσμα ικανοποιητικό.  
Ακολουθούν εκτυπώσεις του δικτύου.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Αποκλίσεις 0.1-0.2%

Παρατήρηση

Λόγω προγραμματιστικού λάθους στις πρώτες εκτελέσεις του προγράμματος το σύστημα εκπαιδευόταν στις πρώτες 80 τιμές δεδομένων και εξεταζόταν στις υπόλοιπες. Αυτό είχε ως αποτέλεσμα σαφώς υποδεέστερη ακρίβεια αλλά τάχιστη εκπαίδευση και παραγωγή αποτελεσμάτων. Ενδεικτικά, το σύστημα εκπαιδεύτηκε σε 10.000 epochs εντός χρόνου μικρότερου των 3 λεπτών και απέδωσε αποτέλεσμα με απόκλιση της τάξης του 2-5%. Αν και δεν είμαστε σε θέση να γνωρίζουμε τα αποδεκτά επίπεδα ακρίβειας θεωρούμε πως ένα υπεργρήγορο σύστημα που εκπαιδεύεται σε μικρό χρόνο και παράγει έγκυρα αποτελέσματα είναι άξιο αναφοράς. Ενδεχομένως να οφείλεται στην μορφή των δεδομένων με τα επαναλαμβανόμενα μοτίβα ή στην πυκνότητα των δεδομένων στο αρχικό dataset. Θα συμπεριληφθεί και ο ελαττωματικός κώδικας στην τελική παράδοση.