3η ΘΕ – 2η ΥΕ – Δεύτερη Εφαρμογή

Εισαγωγή

Στα πλαίσια της 2ης ολοκληρωμένης εφαρμογής, θα αξιοποιήσουμε ανοικτά δεδομένα που μπορούμε να βρούμε στον Παγκόσμιο Ιστό, για να φτιάξουμε ένα απλό προγνωστικό μοντέλο. Η εφαρμογή μας έχει ως αντικείμενο τον καιρό και συνεπώς το προγνωστικό μας μοντέλο θα επιχειρεί να προβλέψει τις μετεωρολογικές συνθήκες που θα επικρατήσουν σε μεταγενέστερη στιγμή.

Για να πετύχουμε το σκοπό μας, θα μελετήσουμε:

- Πως μπορούμε να συνδεθούμε σε μια ανοικτή διαδικτυακή προγραμματιστική διεπαφή, η οποία μας επιτρέπει να ανακτήσουμε δεδομένα με χρήση κάποιων παραμέτρων (π.χ. ημερομηνία, τοποθεσία κ.α.).
- Πως αξιοποιούμε ανοικτά δεδομένα που προσφέρονται με τη δομή JSON (JavaScript Object Notation) σε προγράμματα της Python.
- Πως μετασχηματίζουμε τα δεδομένα σε πινακοειδή μορφή που διευκολύνει την επεξεργασία και την αποθήκευσή τους στον υπολογιστή μας με χρήση του εργαλείου pandas.
- Πως οπτικοποιούμε δεδομένα με τη βιβλιοθήκη matplotlib.
- Πως δημιουργούμε, αξιολογούμε και αξιοποιούμε ένα απλό προγνωστικό μοντέλο γραμμικής παλινδρόμησης με τη βιβλιοθήκη sklearn.

Οι ανωτέρω διαδικασίες (ανάκτηση δεδομένων, μορφοποίηση δεδομένων, οπτικοποίηση δεδομένων και εκπαίδευση/αξιοποίηση προγνωστικών μοντέλων) είναι τα βασικά στάδια που εφαρμόζονται σε όλα τα project της επιστημονικής επεξεργασίας της πληροφορίας (data science). Φυσικά η πολυπλοκότητα των σταδίων ποικίλλει ανάλογα και με το είδος του project και τη φύση του προβλήματος που αντιμετωπίζουμε. Ωστόσο η βασική φιλοσοφία παραμένει η ίδια και στην εφαρμογή αυτή, θα πάρετε μια πρώτη και απλοποιημένη γεύση από τον πολύ ενδιαφέροντα τομέα του data science.

Αρχή Λειτουργίας

Το πρώτο βήμα για να ξεκινήσει κανείς στο data science είναι να αποκτήσει ένα σύνολο δεδομένων (dataset) για το πρόβλημα που τον ενδιαφέρει. Συχνά τέτοια datasets υπάρχουν έτοιμα προς μεταφόρτωση, ωστόσο αυτά μπορεί να μην καλύπτουν τις ανάγκες μας. Σε αρκετές περιπτώσεις αναγκαζόμαστε να χτίσουμε αυτά τα σύνολα δεδομένων, αποκτώντας δεδομένα από μία ή περισσότερες πηγές.

Οι πηγές αυτές μπορεί να περιέχουν έτοιμα σύνολα προς μεταφόρτωση (από τα οποία επιλέγουμε) ή, πιο συχνά, προσφέρουν μια προγραμματιστική διεπαφή (Application Programming Interface, API), που μας επιτρέπει μέσω ορισμάτων να μεταφορτώσουμε μόνο εκείνο το υποσύνολο δεδομένων που μπορεί να μας ενδιαφέρει. Για τα δεδομένα καιρού, θα χρησιμοποιήσουμε την υπηρεσία visualcrossing.com, η οποία παρέχει τρέχοντα και ιστορικά δεδομένα καιρού. Η υπηρεσία προσφέρει μέχρι 1000 δωρεάν εγγραφές δεδομένων την ημέρα, οι οποίες είναι αρκετές για το σκοπό μας. Θα πρέπει να εγγραφείτε στην υπηρεσία δημιουργώντας ένα λογαριασμό, και να αποκτήσετε ένα «κλειδί» (API key) το οποίο χρησιμοποιείται ως μέσο ταυτοποίησης σε όλες τις κλήσεις που θα κάνουμε για να ζητήσουμε δεδομένα.

Η εφαρμογή λοιπόν πρέπει να:

1. Συνδέεται με το API του visualcrossing και να κατεβάζει ιστορικά δεδομένα των τρέχουσων μετεωρολογικών συνθηκών για τις τελευταίες 2 εβδομάδες. Τα δεδομένα αφορούν όλα τα διαθέσιμα από την υπηρεσία μετεωρολογικά στοιχεία (θερμοκρασία, υγρασία, ατμοσφαιρική πίεση κ.α.) ανά μία ώρα. Αν λοιπόν φανταστούμε πως κάθε υποσύνολο τιμών για τις παραμέτρους αυτές αφορά μια ώρα της ημέρας, συνολικά θα έχουμε 24 ώρες * 7 ημέρες * 2 εβδομάδες = 336 υποσύνολα δεδομένων. Οι συντεταγμένες γεωγραφικού μήκους και πλάτους που απαιτούνται για την ανάκτηση δεδομένων ορίζονται ως μεταβλητές που αρχικοποιούνται εντός του κώδικα από εσάς (μπορείτε να χρησιμοποιήσετε όποιο ζεύγος επιθυμείτε).


```
"hours": [
 {
  "datetime": "12:00:00",
  "datetimeEpoch": 1654938000,
  "temp": 25,
  "feelslike": 25,
  "humidity": 62.13,
  "dew": 17.2,
  "precip": 0,
  "precipprob": 0,
  "snow": 0,
  "snowdepth": 0,
  "preciptype": null,
  "windgust": 16.6,
  "windspeed": 7.8,
  "winddir": 320,
  "pressure": 1007,
  "visibility": 10,
  "cloudcover": 50,
  "solarradiation": 1053,
  "solarenergy": 3.8,
  "uvindex": 10,
  "conditions": "Partially cloudy",
  "icon": "partly-cloudy-day",
  "stations": [
   "LGRX",
   "D6214",
   "LGAD"
  ],
  "source": "obs",
  "datetimeInstance": "2022-06-11T09:00:00.000Z"
 },
```


Παράδειγμα μέρους της δομής επιστρεφόμενων δεδομένων από την κλήση στο visualcrossing για συγκεκριμένη ιστορική χρονική στιγμή και γεωγραφική τοποθεσία. Παρατηρήστε ότι η δομή μοιάζει πάρα πολύ με τα λεξικά της Python!

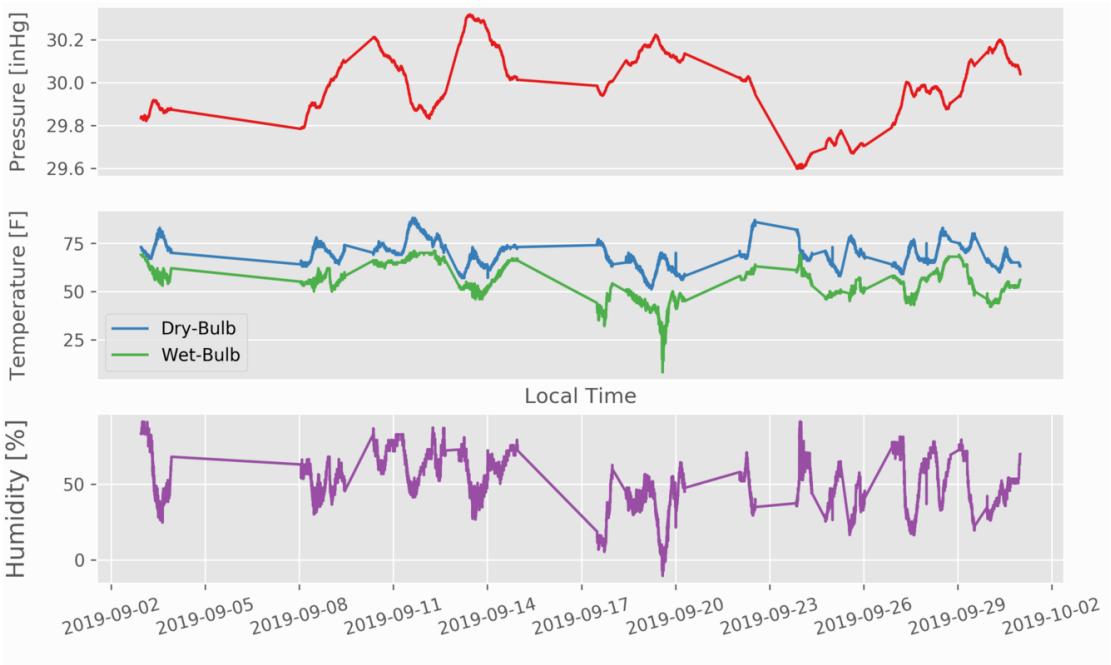
2. Από τα δεδομένα αυτά η εφαρμογή δημιουργεί ένα dataframe (pandas) που διατάσσει τα δεδομένα σε πινακοειδή μορφή. Από τα δεδομένα αυτά διατηρούμε μόνο τα στοιχεία θερμοκρασία, υγρασία και ατμοσφαιρική πίεση και ώρα της ημέρας.

Index	Hour	Temperature	Humidity	Pressure
0	14	25.3	64	1029
1	15	26.4	62	1018
….	….	….	….	….
335	23	17.6	82	800

Παράδειγμα διάταξης των επιστρεφόμενων αποτελεσμάτων σε πίνακα pandas.

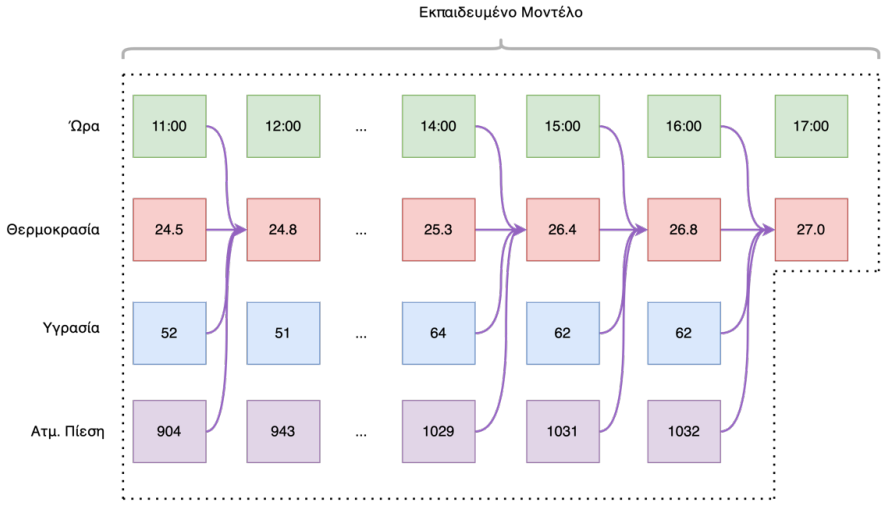
3. Η εφαρμογή αποθηκεύει τα δεδομένα που έχουν μετασχηματιστεί σε πινακοειδή μορφή σε αρχείο .csv στον υπολογιστή μας και ξαναχτίζει τον πίνακα διαβάζοντας από το ίδιο csv αρχείο. Αυτό είναι σημαντικό ώστε να μη δαπανούμε χρόνο αλλά και τις δωρεάν κλήσεις που έχουμε, ώστε να ανακτούμε συνέχεια τα δεδομένα από το openweathermap.

4. Η εφαρμογή παράγει γραφήματα που απεικονίζουν τη διακύμανση των χρονοσειρών για τα μετεωρολογικά στοιχεία που μας ενδιαφέρουν.



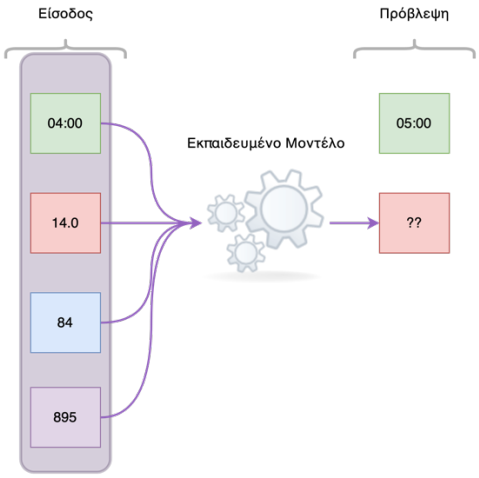
Παράδειγμα απεικόνισης χρονοσειρών σε γραφικές παραστάσεις

5. Η εφαρμογή δημιουργεί ένα μοντέλο πρόγνωσης που στηρίζεται στα δεδομένα θερμοκρασίας, υγρασίας και πίεσης της προηγούμενης ώρας, ώστε να μαντεύει την τρέχουσα θερμοκρασία. Η προτεινόμενη μεθοδολογία δημιουργίας του μοντέλου παρουσιάζεται στο εισαγωγικό βίντεο και το συνοδευτικό υλικό.



Το μοντέλο εκπαιδεύεται μαθαίνοντας την θερμοκρασία για κάποια συγκεκριμένη ώρα, με βάση τα δεδομένα της αμέσως προηγούμενης ώρας

6. Η εφαρμογή αξιοποιεί το προγνωστικό μοντέλο ώστε να μπορεί να δέχεται σαν είσοδο α) την ώρα της ημέρας και β) τις μετεωρολογικές συνθήκες που αφορούν την ώρα αυτή, ώστε να παρέχει ως έξοδο την πρόγνωση για την αμέσως επόμενη ώρα.



Το μοντέλο μπορεί να δεχθεί ως είσοδο μια πλειάδα δεδομένων ώρας, θερμοκρασίας, υγρασίας και πίεσης, και στη συνέχεια παράγει ως έξοδο την προβλεπόμενη θερμοκρασία, η οποία θα αφορά την επόμενη ώρα. Η έξοδος δε συμπεριλαμβάνει την ώρα, αλλά αυτή υπονοείται από τον τρόπο με τον οποίον έχουμε εκπαιδεύσει το μοντέλο – πάντα η παραγόμενη έξοδος αφορά την αμέσως επόμενη ώρα από αυτή που δώσαμε ως είσοδο!